

## INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.
2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.
3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.
4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.
5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

**University  
Microfilms  
International**

300 N. Zeeb Road  
Ann Arbor, MI 48106

8315610

McPherson, Donald John

USING THE RASCH MODEL TO EVALUATE TEST ITEMS FOR GRADE 4 AND  
GRADE 7 MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM CRITERION-  
REFERENCED READING TESTS ADMINISTERED 1973 THROUGH 1979

Wayne State University

PH.D. 1983

**University  
Microfilms  
International** 300 N. Zeeb Road, Ann Arbor, MI 48106

**Copyright** 1983

by

McPherson, Donald John

**All Rights Reserved**

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark .

1. Glossy photographs or pages \_\_\_\_\_
2. Colored illustrations, paper or print \_\_\_\_\_
3. Photographs with dark background \_\_\_\_\_
4. Illustrations are poor copy \_\_\_\_\_
5. Pages with black marks, not original copy \_\_\_\_\_
6. Print shows through as there is text on both sides of page \_\_\_\_\_
7. Indistinct, broken or small print on several pages
8. Print exceeds margin requirements \_\_\_\_\_
9. Tightly bound copy with print lost in spine \_\_\_\_\_
10. Computer printout pages with indistinct print
11. Page(s) \_\_\_\_\_ lacking when material received, and not available from school or author.
12. Page(s) \_\_\_\_\_ seem to be missing in numbering only as text follows.
13. Two pages numbered \_\_\_\_\_. Text follows.
14. Curling and wrinkled pages \_\_\_\_\_
15. Other \_\_\_\_\_

University  
Microfilms  
International

USING THE RASCH MODEL TO EVALUATE TEST ITEMS FOR  
GRADE 4 AND GRADE 7 MICHIGAN EDUCATIONAL  
ASSESSMENT PROGRAM CRITERION-REFERENCED READING  
TESTS ADMINISTERED 1973 THROUGH 1979

by

DONALD JOHN MCPHERSON

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

1983

MAJOR: EDUCATIONAL  
EVALUATION AND RESEARCH

Approved by:

Donald M. Smith Dec. 6, 1987  
Advisor date

Joseph A. Leluta  
Joseph Z. Kostelnik  
Eugene P. Smith

## ACKNOWLEDGMENTS

The author wishes to express his deep appreciation and thanks to:

Dr. Donald Marcotte for his guidance, patience, and support.

Dr. Joseph Labuta, Dr. Joseph Posch Jr., and Dr. Eugene Smith for their direct assistance during the preparation of this research.

Dr. Danial Schooley for encouraging the selection of the Michigan Educational Assessment Program tests as the research objects of this study and Dr. Edward Roeber for assisting in obtaining the data.

Dr. Benjamin Wright, Dr. Ronald Mead, and Mr. Richard Smith for the information they provided toward gaining a better understanding of the conception of the item fit statistic.

Dr. Ernie Bauer and Dr. Robert Veitch for the information that made it possible to successfully run BICAL.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . .	ii
LIST OF TABLES . . . . .	v
LIST OF APPENDICES . . . . .	vii
CHAPTER I.	
STATEMENT OF THE PROBLEM . . . . .	1
INTRODUCTION	
THE PROBLEM	
RESEARCH QUESTIONS	
DISCUSSION	
SIGNIFICANCE	
DEFINITIONS OF TERMS	
CHAPTER II.	
BACKGROUND AND REVIEW OF RELATED LITERATURE . . . . .	12
INTRODUCTION	
BACKGROUND LITERATURE	
RELATED LITERATURE	
SUMMARY	
CHAPTER III.	
RESEARCH DESIGN - METHODS, PROCEDURES AND LIMITATIONS . . . . .	37
PREPARATION OF DATA	
LIMITATIONS OF STUDY	
STATISTICAL ANALYSIS	
SUMMARY	
CHAPTER IV.	
RESULTS - DISCUSSION OF RESEARCH QUESTIONS AND PRESENTATION OF ANALYSIS RESULTS . . . . .	83

INTRODUCTION

DESCRIPTIVE STATISTICAL ANALYSIS

SUMMARY

CHAPTER V.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH . . . . . 150

SUMMARY OF CONCLUSIONS

RECOMMENDATIONS FOR FUTURE RESEARCH

BIBLIOGRAPHY . . . . . 272

ABSTRACT . . . . . 277

AUTOBIOGRAPHICAL STATEMENT . . . . . 279

LIST OF TABLES

TABLE 1. DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL . . . . .	86
TABLE 2. NUMBER OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM . . . . .	95
TABLE 2A. SCORES OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH QUARTILE LEVEL . . . . .	100
TABLE 3. PROPORTION OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH QUARTILE LEVEL WHO CORRECTLY ANSWER EACH ITEM . . . . .	102
TABLE 4. ESTIMATES OF DIFFICULTY FOR EACH TEST ITEM MEASURED IN TERMS OF LOG-ODDS . . . . .	108
TABLE 5. t-STATISTIC AND ONE-TAIL PROBABILITY LEVEL (I. E., ALPHA LEVEL) DERIVED ON COMPARISON OF AVERAGE SCORE OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST . . . . .	115
TABLE 6. THE AVERAGE SCORE OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST . . .	117
TABLE 7. THE SCORE VARIANCE OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST . . .	120
TABLE 8. THE STANDARD ERROR OF ESTIMATE ON SCORES OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST . . . . .	122
TABLE 9. t-STATISTIC AND ONE-TAIL PROBABILITY LEVEL (I. E., ALPHA LEVEL) DERIVED ON COMPARISON OF THE LARGEST PROPORTION OF DIFFICULT ITEMS TO THE 1979 PROPORTION OF DIFFICULT ITEMS IN THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHICH DO NOT FIT THE RASCH MODEL . . . . .	125
TABLE 10. DIFFICULT ITEMS IN THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHICH DO NOT FIT THE RASCH MODEL SHOWN BY QUESTION NUMBER 1 TO 5 WITHIN TEST OBJECTIVE . . . . .	130
TABLE 11. $\chi^2$ -STATISTIC AND ONE-TAIL PROBABILITY LEVEL (I. E., ALPHA LEVEL) DERIVED ON COMPARISON OF THE SMALLEST PROPORTION OF PASSABLE OBJECTIVES TO THE PROPORTION OF PASSABLE OBJECTIVES IN THE 1979 FOURTH AND SEVENTH GRADE MEAP READING TESTS AFTER DELETING HARD ITEMS WHICH DO NOT FIT THE RASCH MODEL . . . . .	135



TABLE 12. PROBABILITY OF PASSING TEST OBJECTIVES IN THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHERE IT IS ASSUMED THAT STUDENTS WILL ALWAYS GET DIFFICULT ITEMS WRONG WHEN THOSE ITEMS ARE SO DIFFICULT THAT THEY DO NOT FIT THE RASCH MODEL . . . . . 139

TABLE 13.  $\chi^2$ -STATISTIC AND ONE-TAIL PROBABILITY LEVEL (I. E., ALPHA LEVEL) DERIVED ON COMPARISON OF THE PROPORTION OF STUDENTS PASSING LESS THAN 40% OF THE 1979 TEST OBJECTIVES BEFORE AND AFTER DELETING HARD ITEMS WHICH DO NOT FIT THE RASCH MODEL . . . . . 143

LIST OF APPENDICES

APPENDIX A	
DATA REQUEST AND ASSURANCES AGREEMENT . . . . .	162
APPENDIX B	
FOURTH GRADE AND SEVENTH GRADE INDIVIDUAL STUDENT REPORT FORM FOR 1973-74 MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM AND CORRESPONDING LISTS OF ITEMS MEASURING EACH OBJECTIVE . . . . .	165
APPENDIX C	
THE ITEM FIT STATISTIC - EVOLUTIONARY CHANGES OF INTERPRETATION	170
APPENDIX D	
SPSS CONTROL SET EXAMPLE FOR DESCRIPTIVE STATISTICS ON SAMPLE OF 5,000 DATA . . . . .	186
APPENDIX E	
SPSS CONTROL SET EXAMPLE FOR CREATING STANDARD FORMAT SOURCE FILES ON SAMPLE OF 5,000 DATA . . . . .	188
APPENDIX F	
BICAL - A COMPUTER PROGRAM FOR USE IN CONDUCTING RASCH ANALYSIS: CODING BICAL CONTROL CARDS . . . . .	191
APPENDIX G	
SPSS CONTROL SET EXAMPLE FOR TRANSFORMING SCORES ON MEAP OBJECTIVES IN SAMPLE OF 5,000 DATA WHEN ITEMS THAT DO NOT FIT THE RASCH MODEL ARE RESCORED . . . . .	245
APPENDIX H	
LOG OF COMPUTER RUNS USING BICAL TO PROCESS FOURTEEN RANDOM SAMPLES TAKEN FROM SAMPLE OF 5,000 DATA . . . . .	257

## CHAPTER I

### STATEMENT OF THE PROBLEM

#### INTRODUCTION

The Michigan Educational Assessment Program (MEAP) has been in operation since 1969 when the state legislature enacted enabling legislation. Under the program, students at the fourth and seventh grade levels have been tested for mastery of basic skills in mathematics and reading every year since 1970. The first three years tests were of the norm-referenced type. That is, individual achievement was determined by referencing a score on the test to the scores previously earned on the test by some "standard" group which is intended to be representative of the type of persons expected to take the test. However since 1973, criterion-referenced tests have been employed in the MEAP program. Unfortunately, many of the classical evaluation methods applied to norm-referenced tests of long standing do not work in connection with criterion-referenced tests.

#### THE PROBLEM

MEAP Tests and remedial education in Michigan have been tied together by the decision of the State Board of Education in 1974 to use MEAP scores to identify those students who need remedial instruction. The Board adopted the criterion that students passing fewer than 40% of the MEAP Test objectives will qualify for participation in programs eligible for federal funding.

The decision to provide remedial education is determined by whether or not a student passes 40% of the MEAP Test objectives. How accurate is the determination that a student has passed that proportion of objectives, while another has not? This question focuses on the

measurement problem that constitutes the basis of this study. Rasch measurement seems to offer improved evaluation methods. Should this new technique be used to supplement, or perhaps even replace, those techniques which are presently used in connection with the Michigan Educational Assessment Program? A better concept of measurement theory would be extremely useful in developing these tests and in evaluating the scores that result from their use. If the application of Rasch measurement theory to MEAP tests should result in better measurement tools, this result would certainly improve the level of confidence in the judgements made on the basis of MEAP test results.

#### RESEARCH QUESTIONS

Five research questions are dealt with in this investigation with seven samples drawn from MEAP tests given from 1973 through 1979:

Question 1: How many items in the MEAP Reading Test for the years 1973 through 1979 fit the Rasch model?

Question 2: Is there a statistically significant increase in the measurement efficiency of MEAP reading tests after items which do not fit the Rasch model, because they are too difficult, have been credited to students who get those items wrong?

Question 3: Is there any change in pattern respecting item fit to the Rasch model which would suggest either an increase or decrease in difficulty as items are used over time?

Question 4: Can a negative effect of items identified as being too difficult to fit the Rasch model be demonstrated on the probability that a student will pass the MEAP Reading Test learning objectives by re-scoring these items in favor of the student and treating the items as if they had been originally calibrated to fit the Rasch model?

Question 5: Do changes occur in the proportion of students who are "qualified" for remedial instruction between scores reported on MEAP Reading Test objectives compared to the proportion of students who would be qualified if scores were based solely on items which have been re-scored to compensate for the adverse effect perceived in this study by the method described? That is, does a change in proportion of qualified

students occur when students are credited for too-difficult items they have missed?

#### DISCUSSION

It is unfortunate that so much of the statistical literature that is available appears to have so little application to criterion-referenced tests like those used in the Michigan Educational Assessment Program. The situation exists where the statistical tools that are available may not be up to the demands being placed on them in practice today. Facts in any area of human endeavor are extremely hard to discover. In this case an accepted and relatively reliable measure for human achievement is needed which may be used in connection with objective tests, but one is not available now.

Better measures of analysis are needed to use with MEAP Test data than are presently available for making educational decisions which are intended to help low achievers. The problem is a serious one. The comments of Wright and Stone (1979) pointedly put the basic measurement problem into perspective this way:

It is an old problem in educational testing. Alfred Binet worried about it 60 years ago. Louis Thurstone worried about it 40 years ago. The problem is still unsolved. To some it may seem a small point. But when you consider it carefully, I think you will find that this small point is a matter of life and death to the science of mental measurement. The truth is that so-called measurements we now make in educational testing are no damn good! . . . .

The scales on which ability is measured are uncomfortably slippery. They have no regular unit. Their meaning and estimated quality depend upon the specific set of items actually standardized and the particular ability distribution of the children who happen to appear in the standardized sample. . . .

Change the children and you have a new yardstick. Change the items and you have a new yardstick again. Each collection of items measures an ability of its own. Each measure depends for its meaning on its own family of test takers. How can we make objective mental measurements and build a science of mental development when we work with rubber yardsticks? (p. xi)

## SIGNIFICANCE

The Rasch model appears to offer a better standard of measurement than is available now. Rasch measurement theory appears to offer escape from the limitations of nominal and ordinal measures in testing. Sample dependence is foremost among these limitations. The Rasch model offers psychometricians a vastly improved standard of measurement. It offers a way to apply an interval measurement scale in testing that is independent of both the sample of persons tested or the sample of test items used. An improved standard of measurement, needless to say, affords the opportunity to improve the decisions based upon that measurement. This research is designed to add to knowledge of that standard. Perhaps enough may be learned in this investigation to demonstrate the value of the Rasch model in connection with MEAP tests. If Rasch theory can be employed to improve criterion-referenced measurement, every objective of this program will be easier to accomplish, including the difficult task of determining which student should receive remedial education. Rasch measurement concepts are generally developed in the testing literature in discussions revolving around norm referenced testing. There does not appear to be much discussion about applicability to criterion referenced tests, the type of test which is considered in this study. The literature suggests that Rasch measurement can be applied to all forms of mental measurement, but criterion referenced examples were not found. This investigation explores whether it is appropriate for use in the context of criterion referenced testing, and, if so, how? While the results of this investigation may add something to the general understanding of Rasch analysis, the primary intent is to explore its use with criterion

referenced tests. The tests used in connection with the Michigan Educational Testing Program provide an exceptional opportunity. This is a major program affecting tens of thousands of young students each year. The program is, therefore, large and expensive, and it directly affects the quality of education offered to great numbers of children in the State of Michigan. This program was chosen as the research object of this study in the belief that it effectively illustrates the importance of finding the best standard in test measurement possible.

Creation and implementation of a program like the Michigan Educational Assessment Program is difficult and expensive. There is always hope that enough resources in personnel and financing will be available to do the important jobs in education that need to be done. Few are more important than the MEAP objectives geared to improving the quality of education offered in Michigan. But experience often shows that competing demands exceed scarce resources in the best of times, and these are not the best of times. Eroding tax bases, inflation, and unemployment have created a stressful political and economic climate in Michigan which threatens to diminish already severely limited educational resources in this state. The citizens of Michigan, ever concerned that taxes are well spent, can be expected to be increasingly watchful that programs as visible and costly as MEAP tests give a good accounting of themselves. Unfortunately, at a time when they are likely to be most needed, adequate means for assessing the program may not be available.

The creation and implementation of remedial education programs, and a host of other educational efforts related to the MEAP program may cost

far more than the benefits warrant. It is not easy to determine whether a program is cost effective.

Efficiency will become an increasingly important component in decisions to plan and fund new programs. Need it be said that educators should be equally concerned for the quality of educational programs? Probably not, but there is greater urgency now to find new means to satisfy the profession that existing educational programs are quality programs.

Most of the funding for remedial education in Michigan is derived from the federal Elementary and Secondary Education Act of 1965; Public Law 95-561. Specifically:

1. Title I, Financial Assistance to Meet Special Educational Needs of Children, and
2. Title III, Supplementary Educational Centers and Services; Guidance, Counseling, and Testing.

The Act emphasizes the needs of "educationally deprived" children from families whose income is below the "current poverty level." The criterion for participation is set at a maximum of 40% of the total number of students in a district between the ages of five and seventeen years. Determination of the students who will participate is to be made by "an assessment of educational needs each year" (Public Law 95-561, 1978), provided that said assessment identifies:

1. those children in greatest need,
2. the instructional areas concerned with that need, and
3. the extent of need for remedial education across the district.



The Michigan Educational Assessment Program attempts to identify all three. This investigation explores the Rasch model within the context of the Michigan Educational Assessment Program. Rasch measurement is an approach to test item analysis. The latent trait model Rasch devised makes it possible to analyze the performance of test takers without any need to consider the items which make up the test. Conversely, analysis of item-difficulty and fit to the model are concurrently possible without need to consider the persons who have taken the test or any other group which may have done so.

Traditional methods for evaluating test results must relate to some group to give these results meaning. This is the method used to give norm referenced test scores, for example, meaning. Scores by themselves have no meaning in traditional psychometric theory. One can not know if a score is good or bad without some standard of acceptable performance having been first established by a representative individual or, more often, group of test takers. All subsequent scores are related to the performance of this criterion group. Norm referencing procedures, however, measure performance in terms of variance from the expected score rather than by reference to the performance of a representative individual or group. In any case, traditional methods of test analysis must relate in some way to group performance to have meaning.

In Rasch measurement, test scores have meaning in themselves. They do not take on varying shades of meaning, depending on the group taking the test or by comparison to the performance of some other group. Nor is it necessary to evaluate test items in terms of different groups. In Rasch analysis, test items measure the underlying variable to a known degree of difficulty. Therefore it is unnecessary to know the ability

level of the group taking a test to evaluate the appropriateness of an item in that test. Rasch developed a mathematical model which makes it possible to evaluate person-ability independently of test item difficulty, and, conversely, to evaluate item-difficulty independently of person-ability. The model has only two parameters: person-ability and item-difficulty. Scores from tests based upon this model contain all the information necessary to determine performance for either persons or items.

#### DEFINITIONS OF TERMS

Certain terms should be defined at this point to provide the meanings intended in the subsequent discussion:

Between fit mean square: that measure of variance based upon the sum of the squared differences between actual and expected performance on an item by every person in a score subgroup defined by the BICAL program, divided by the standard deviation of that subgroup performance.

BICAL: the version of a computer program developed at the University of Chicago School of Education used in this study to perform Rasch analysis.

BICAL.3: a more recent version of a computer program developed at the University of Chicago School of Education to perform Rasch analysis.

Calibration process: the procedure for determining item difficulty, within prescribed limits, to determine test item "fit" to the Rasch model and the probability that it will be answered correctly at each score level.

CALFIT: a computer program developed at the University of Chicago School of Education, which precedes BICAL and BICAL.3, for estimating the item-difficulty parameter employed in Rasch analysis.

Column marginal: the sum of the entries in one of the columns of a double-entry table; column marginal totals are customarily at the foot of the respective column.

Criterion-referenced test: a test comprised of items designed to elicit specific behaviors which correspond directly to intended learning outcomes; behaviors which are the conscious result of the instruction process. The criterion referenced tests employed in the Michigan Educational Assessment Program, based on minimal performance objectives, are an example.

Efficient statistic: a statistic that, as the size of the sample is increased, approaches the true population value as a limit and has a normal distribution of error and a smaller standard error than any other measure that could be used to estimate the true value of a particular statistical constant (Good, 1973).

Independence of items: lack of correlation between test items.

Independence of subjects: lack of influence of individual ability between test takers.

Latent trait: the component of a variable that is inherent in, and therefore measured by, a test item.

LOG: a section of a computer program (i.e., BICAL, BICAL.3, or CALFIT) used to estimate the item-difficulty parameter in Rasch analysis. LOG determines the initial estimate of item-

difficulty which is used as input to an iterative process in a section of the program, called MAX, which computes the final estimate.

Log-odds: a mathematical unit which is based upon the natural logarithm of a ratio of the number of times an event happens to the number of times it does not happen. It is used to measure, or calibrate, both item-difficulty and person-ability. For item-difficulty, the log value constitutes the log odds for succeeding on the type of items used to measure the latent trait. For person-ability, the log value constitutes the log odds for failing persons which have an ability level at the midpoint of the ability scale.

Logit: a synonym for log-odds.

MAX: a section of a computer program (i.e., BICAL, BICAL.3, and CALFIT) used to compute the final estimate for the item-difficulty parameter in Rasch analysis.

Non-parallel items: the test items which measure different underlying test variables.

Parallel items: test items which measure the same underlying test variables.

Sample-of-5,000: annual sample of fourth grade and seventh grade Michigan students; each grade totaling 5,000 students.

Score level: the total score attained on a test by one or more students taking the test.

Statistical equivalence: a characteristic of test item subsets which means that observed ability differences between subset scores differ only to the extent that measurement error is

present; thus there should be no statistically significant difference between any subset statistic and the comparable statistic for the complete item pool.

Sufficient statistic: a statistic, derived from a set of observations, which contains all the information in that set of observations relevant to the estimate being made. The arithmetic mean is an example of a sufficient statistic of observations from a population with a normal distribution (James, 1968).

Test characteristic curve: a probability curve which plots the level of expectancy that students at a specific score level are likely to possess a corresponding level of ability.

Total fit mean square: that measure of variance based upon the sum of the squared differences between actual and expected performance on a test item by every person in the sample, divided by the standard deviation of that total performance.

## CHAPTER II

### BACKGROUND AND REVIEW OF RELATED LITERATURE

#### INTRODUCTION

Test item analysis based upon the Rasch model is a technique which employs computer processing capabilities to excellent advantage. Though somewhat less precise manual methods are available which work well in pilot tests, or with relatively small data sets, it would be difficult to conceive of an extensive application outside a computer environment. The model has been criticized for being computer dependent by Whitely and Dawis (1974). However Wright (1977), who is probably the foremost advocate of the Rasch model in American education, sharply disagrees with Whitely-Dawis on this and a number of other points.

While the manual procedures may be employed extensively in practice, there are few examples in the literature. It is difficult to know how extensively the Rasch model is used. The literature is not particularly extensive. Fewer than 100 articles have appeared in the professional journals since Georg Rasch introduced the model (1960). Fully a third of these have been written by researchers in Europe. Fewer than two dozen doctoral dissertations dealing with various aspects of the Rasch model have been completed by graduates of American universities. Obviously, in terms of historical development, the model itself and the very useful (if not absolutely essential) computer software supporting it, are relatively new when compared to the more traditional approaches to the analysis of paper and pencil tests used in mental measurement.

Researchers who have investigated the Rasch model and written about their conclusions predominantly favor its application. There appear to

be distinct advantages in objectivity and evaluation not found in classical techniques. There are a few detractors, but for the most part they confine their criticism to specific applications of the model while supporting its overall usefulness as a measurement tool. With the possible exception of Whitely and Dawis (1974) article, the literature does not clearly advocate classical methods as being superior to the Rasch model. For the most part peer evaluation has shown the Rasch model to be mathematically sound, statistically robust, an efficient tool of measurement, and useful in very practical terms. It appears, therefore, to be technically sound and, especially to those statistically oriented educators who are comfortable in a computer environment, relatively easy to use. The incentive to use Rasch analysis lies in the fact that the Rasch model fills a definite void in test-theory: objective measurement. The reasons that it has not been used more extensively are obscure. Of course there are a number of major applications including "The National Reference Scale for Reading" (Rents, 1974), "Equating Reading Tests with the Rasch model" (Rents & Bashaw, 1976), and "The Validation and Refinement of Measures of Literal Comprehension in Reading for use in Policy Research and Classroom Management" (O'Reilly, Schuder, Kidder, Salter, and Hayford, 1976). In light of its great promise, interest in the Rasch model appears to be growing.

As might be expected of any major contribution to basic knowledge in any discipline, the Rasch model has drawn the attention of leading scholars in psychology and education. Each year since 1966, when Rasch's definitive articles on the model were first published, a significant number of mathematicians, statisticians, and psychometricians in Europe and the United States have contributed to the

literature relating to the Rasch model. As noted previously, all of the writers have not completely supported Rasch's theories, but detractors are few and, for the most part, unsupported in the literature. The underlying theory has so far withstood very able criticism. The model has been applied in a number of practical situations and with apparently satisfactory results. And the great appeal of this model is in the seemingly endless variety of circumstances where it may be used when traditional approaches to evaluation can not be considered. Lack of familiarity with the Rasch model seems to be the greatest hindrance to its future application.

It is a probability model intended for use in test item analysis. Sophisticated mathematical and statistical considerations are involved which warrant appropriate attention from mathematicians and statisticians to be sure. Once past these technical considerations, the model seems to have its greatest appeal to psychometricians because it is so easy to use. It appears that the technical issues have been resolved. The point seems to have been reached where emphasis could be directed toward building confidence in the use of Rasch measurement in connection with increasingly varied and challenging attempts to evaluate human behavior and achievement. Psychometricians have been so long preoccupied with traditional measurement theory that it may be some time yet before the Rasch model may, in practice, live up to the potential its theory would suggest.

Depending on the frame of reference, several individuals should be mentioned in connection with the development of the Rasch model. In terms of the objectives of this research, the most notable contributor to the development of means to apply the model in practical testing



situations is the American professor of education, Benjamin Wright of the Department of Education, the University of Chicago. Wright has done more in this regard, both in the United States and world-wide, than any other person. Wright and his associates have interpreted the Rasch model and provided a computer program to use it. The value of those contributions can be measured by the extensive participation Wright has had in the application of the model in major test evaluation programs in the United States, England, and on the continent. The research in this investigation pertains to the application of basic Rasch theory to a practical test situation. This has, therefore, been the orientation of the literature review to follow. Material has been selected for inclusion on the basis of its contribution toward understanding the major tenets of the model as they apply to objective achievement testing. Elaboration of mathematical and statistical theory has been confined to the essentials necessary to understanding use of the model in practice.

#### BACKGROUND LITERATURE

At this writing, only twenty years have passed since the Danish mathematician Georg Rasch (1960) first explored the basic concepts of the latent trait model. The concept, that has since come to be known as the Rasch model, has evolved as part of a more generalized discussion of various types of psychological tests and alternative item analysis models which might be applied to them. The next year, in 1961, Rasch developed some of the additional theoretical framework for his ideas, but it was not until 1966 that he focused his attention, in two separate articles, on the theoretical composite attributed to Rasch ("An Individualistic Approach to Item Analysis," "An Item Analysis Which

Takes Individual Differences into Account"). The subject matter of these articles is virtually identical, the second being a somewhat abbreviated version of the first. In these articles, Rasch describes an approach to test item analysis that is genuinely innovative. For a number of years he had been concerned with the question of whether or not an approach to the analysis of psychological tests could be devised which would be independent of the tests themselves or the subjects taking them.

Traditionally the properties of a psychological test are defined in terms of variations within some specified population. In practice such populations may be selected in various reasonable ways, and accordingly the properties referred to--for example, the reliability coefficient--are not specific to the test itself but may vary according to how the population is defined. Similarly, the evaluation of a subject is usually linked up with a population by a standardization of some kind and is therefore not specific to the subject per se. Our aim is to develop probabilistic models in the application of which the population can be ignored. It was a discovery of some mathematical significance that such models could be constructed, and it seemed remarkable that data collected in routine psychological testing could be fairly well represented by such models. (Rasch, 1966a, p.89)

By 1966 Rasch had come to believe that one simple model, having only two parameters, would serve this purpose well. He left no doubt that this was his position: "But only recently it has become quite clear that this model is in fact the complete answer to the requirements about the parameters and the adequacy of a discrete probabilistic model be objective in a sense to be fully specified" (p. 89). Rasch introduces his model in terms of three assumptions:

1. every encounter of a given subject with a specific test item has a corresponding probability of a correct answer;
2. the description of this person/item encounter is the product of two factors (i.e., variables): the subject parameter (i.e., ability) and the item parameter (i.e., difficulty);

3. the probability of a correct answer to an item is independent of the probability of a correct answer to any other item. (p. 90)

Rasch demonstrates that all of the information necessary to determine either of the two factors pertinent to a test situation (i.e., person-ability or item-difficulty) is inherent in the marginal total of a matrix of 1's and 0's that shows whether each person tested answered every item correctly (1) or incorrectly (0), and that these factors may be determined from observable data (i.e., scores), independently of each other (pp. 96-104).

The unknown person-ability in these expressions has been replaced with observable quantities, the individual totals of persons who get each correct, and . . . with item totals for each person, in consequence of which we can estimate the person-ability factors without knowing or simultaneously estimating the individual item-difficulty factors. (p. 103)

Rasch pointed out that this "independence" carried out to any subgroup of persons or items within the tests.

In fact, the comparison of any two subjects can be carried out in such a way that no other parameters are involved than those of the two subjects . . . Similarly, any two stimuli (items) can be compared independently of all other parameters than those of the two stimuli, the parameters of all other stimuli as well as the parameters of the subjects having been replaced by observable numbers. (pp. 104-105)

The numbers referred to are the marginal totals of the scoring matrix. Thus, the mathematician Rasch advanced the basic tenets of his model. While he was primarily concerned with psychological testing, he was quick to point out that it need not be restricted to psychology; indeed--that it should not be.

Wright saw the application of the Rasch model to mental measurement done in education through achievement testing. Wright and his associates have written one book four monographs and nearly a dozen

articles on the subject since 1969. Wright and Naugis Panchapakesan, a former student of his at the University of Chicago, thoroughly introduced the subject to the professional literature of education that year in an article entitled "A Procedure for Sample-free Item Analysis" (1969). Following a thorough introduction of the model, this article introduces a statistical test for determining when an item fits the model. This capability is essential to the application of the Rasch model.

If a given set of items fits the model this is the evidence that they refer to a unidimensional ability, that they form a comfortable fit. Fit to the model also implies that item discriminations are uniform and substantial, that there are no errors in item scoring and that guessing has had a negligible effect. Thus the criterion of fit to the model enables us to identify and delete "bad" items. Item calibration is concluded by reanalyzing the retained items to obtain the final estimates of their easiness. (p. 25)

There are two major aspects to item calibration: determination of fit to the model and determination of how easy or how difficult an item is. Once the individual test items have been calibrated, they may then be used to determine individual ability on the basis of score alone.

An important consequence of this model is that the number of current responses to a given set of items is a sufficient statistic for estimating person ability. This score is the only information needed from the data to make the ability estimate.

Therefore, we need only estimate an ability for each possible score. Any person who gets a certain score will be estimated to have the ability associated with that score. All persons who get the same score will be estimated to have the same ability. (p. 24)

And this is how Wright-Panchapakesan proposed to use the model, viz., estimate person-ability. The procedure follows two stages. The first stage is item calibration which is mostly concerned with determination of the items fit to the model. The second stage is the application of calibrated items to determine individual ability. "In the second stage, person measurement, some or all of the calibrated items are used to obtain a test score. An estimate of person-ability and the standard error of this estimate

are made from the score and from the easiness of the items used". (p. 26)

In using the Rasch model, test reliability ceases to be a factor for concern. The calibration process, among other things, is intended to enhance measurement precision. The process eliminates "bad" items and retains items that measure a single trait within prescribed limits of variability.

In this procedure the reliability of a test, a concept which depends on the ability distribution of the sample, is replaced by the precision of measurement. The standard error of the ability estimate is a measure of the precision attained. This standard error depends on the number of items used. The range of item easiness with respect to the ability level being measured, also effects the standard error of the ability measurement. But in practice this effect is minor compared to the effect of test length . . . It is possible to reach any desired level of precision by varying the number of items used in the measurement, just providing that the range of item easiness is reasonably appropriate to the ability being measured. (p. 26)

The article provides an extensive explanation of two procedures for item calibration. The first, a procedure referred to as "LOG," is most appropriate to samples of 500 or more students. When used with smaller samples, an undesirable amount of estimate bias is introduced.

The LOG method for estimating item easiness and variability is described:

The log method of estimates is based on using the observed proportion of success . . . within a particular score group . . . as an estimate of the probability . . . of obtaining a correct response , for any person in (a) score group . . . to an item of easiness. (p. 27)

Because LOG has this potential for biasing estimates of item easiness in small samples, the authors favor a second procedure referred to as "MAX." MAX begins where LOG ends. That is, the easiness estimates begin with the LOG procedure. These estimates, in turn, became the

initial input to the MAX procedure, which is based upon an iterative method for maximizing parameter estimates called the "Newton-Raphson equation."

In our analysis we use the Newton-Raphson procedure to solve for the unknown parameter estimates. This procedure is an iterative one. We start with an initial estimate . . . and using the Newton-Raphson equation to obtain an improved estimate . . . now using the new value . . . as the starting estimate, we repeat the procedure until the estimates do not change appreciably. (p. 35)

The computer programs necessary for making item estimates are presented in the article. Reference is also made to the fact that they are available at cost from the University of Chicago. Once items have been calibrated using either the LOG or MAX procedure, the items may then be considered for use in making person-ability measurements. Up to this point, item calibration is a straight-forward computational process, but the result of this process calls for a certain amount of judgment which is entirely independent of the technical processes so far used. At the end of the computer run, there are very likely two item sets: One fits the Rasch model and the other does not. Of course those that fit the model stand as ready for use in making ability measures. There is some question that those items which do not fit the model should be used, but it is by no means determined at this point that they should be eliminated from use in making person measures. Some items that do not fit the model may be retained, but discussion of this aspect of the calibration procedure was not part of this article. Once the final evaluation of items that do not fit the model has been completed, the items retained following the computer calibration procedure may then be used in test calibration.

During item calibration it is necessary to decide whether all the items that have been tried are to be retained for the final pool. We need a statistical criterion for deciding whether an item

is good enough from the point of view of the model . . . To make this decision we need to investigate how the elements . . . in the data matrix [i.e., the score matrix] . . . depend on the estimates [of item difficulty and of person-ability]. If we can derive expectations . . . of these elements in terms of the obtained estimates we can form a standard deviate . . . and use this deviate as the basis for a test of item fit. If item  $i$  fits the model, and the score group . . . is large enough, then the [standard deviate] will have an approximately normal distribution. (p. 44)

An important product of the calibration process is a matrix of item-difficulty estimates, one for each item, at every score level. It is this matrix which is used to determine whether the item fits the model. The estimate of difficulty is constant at each score level. Since, under the model, each score level corresponds to an ability level, one would not expect persons at a given score level to answer a question correctly that is too difficult for them. Conversely, one would not expect persons at a given score level to miss a question with a lower difficulty estimate. That is, at a given level of ability, persons would not be expected to get "hard" questions right or "easy" questions wrong. Beyond predetermined limits of probability, such questions do not fit the Rasch model. Such questions would have to be carefully evaluated, and, if found wanting, dropped from the item pool. This evaluation is done from the item-difficulty estimates generated by the calibration process described in this article.

Examination of the matrix  $Y$ , with the standard deviates . . . as elements will show us how well the items fit, and indicate where there are signs of misfit . . . From the matrix  $Y$  we can obtain statistics which will enable us to evaluate the fit of the model to the data as a whole, and we can also form approximate statistics which will help identify items which are bad, and hence need to be reconsidered . . . The overall statistic used in the procedure is a chi-square statistic which is obtained by summing the squared unit normal deviates over the entire matrix  $Y$  . . . (p. 45)

Thus the method for evaluating item fit to the Rasch model employs Chi-square analysis of item-difficulty estimates. The authors caution

that lack of fit to the model does not automatically mean to drop the item from the pool.

Since the . . . equation for Chi-square is an approximation, we do not think it is advisable to mechanically delete all items which Chi-square is significant at some level. We prefer to examine in detail items which Chi-square is large. This may mean evaluating the possible effects of discrimination and guessing in these bad items. Then when we have decided which of the "bad" items to delete, we rerun the analysis to see how the remaining set of items look. (p. 45)

The authors devote the balance of this benchmark article to the FORTRAN computer program segments which are used to accomplish the procedures described. The entire procedure constitutes sample-free item analysis.

The full spectrum of literature on the Rasch model falls primarily into four categories:

1. Description;
2. Development of underlying mathematical theory and related proofs;
3. Development of research design and applications theory; and
4. Empirical evaluation of theory and applications.

Description of the Rasch model has been provided by Wright and his associates in a series of articles over a ten year period (Wright & Panchapakesan, 1969, Wright & Stone, 1979). Excellent descriptions of the model also appear in articles by Anderson, Kearney & Everett (1968), Whitely & Dawis (1974), and in a text by Wilmott & Fowles (1974). The Whitely & Dawis article raises important implications which detract from the model. The article challenges the Rasch measurement concept in ways that, if left unanswered, leave serious doubt affecting its usefulness in the practice of mental measurement. The challenge was met by Wright in a thorough and carefully reasoned response (1977) three years after



the Whitely & Dawis article appeared. The nature of the controversy will be dealt with presently, but it is important to emphasize at this point that there have been no effective challenges to date regarding any basic aspect of Rasch theory.

The underlying mathematical theory relevant to the Rasch model and the proofs of that theory were carefully developed by Rasch (1966, pp. 50-56) and Anderson (1972, pp. 43-50). Both articles tend to be highly technical with primary emphasis on the underlying concepts upon which the Rasch model is based. The Anderson article (1972) restates these basic concepts and goes on to present the underlying theory for "MAX", the parameter estimation procedure favored in most computer applications of the Rasch model. Both of these articles employ probability concepts extensively. They are very helpful in gaining general assurance that the Rasch model is conceptually sound. Test designers who lack this assurance will possibly feel uncomfortable in any attempt they may make to use the model. Therefore, while it is probably not necessary to place great importance upon full understanding of the mathematics behind Rasch measurement, a fundamental understanding of probability will help a great deal: "In that the Rasch model is a probability model, distributions of probabilities of item and person parameters is assumed and produced" (Brink, 1972, p. 924).

The research design and applications theory appropriate to the Rasch model have appeal because they are parsimonious. Aside from the prospect of objective measurement, which it offers, the attraction of the model largely rests upon this apparent simplicity. There are, after all, only two variables to consider, item-easiness and person-ability; but a major shift in thinking for those oriented to classical test

theory may be in order. "Of all latent trait models posed for person measurement, the Rasch model has the fewest ingredients, just one ability parameter for each person and one difficulty parameter for each item. These parameters represent the positions of persons and items on the latent variable they share" (Wright, 1977b, p. 97). As psychometricians are probably more familiar with classical measurement theory, they are likely to prefer its use to Rasch analysis. The intention of this investigation is to provide information which will clarify the distinctions which can be made between classical measurement theory and Rasch measurement. Perhaps the ability to draw these distinctions will enhance understanding of both approaches. On the other hand, lack of appreciation for these distinctions will lead to confusion, at best, or misapplication of theory out of appropriate context, at worst. For example,

The absolute value of the true score has been relatively unimportant for classical measurement. This is because the primary purpose of classical testing is to rank order examinees consistently. For this purpose, the critical problem is not determining the true score, but rather it is determining the correlation between the scores and observed scores. (Epstein, 1975, p. 627)

In Rasch item analysis, the test score is of paramount importance. "An important consequence of this model is that the number of correct responses to a given set of items is a sufficient statistic for estimating person-ability. This score is the only information needed from the data to make the ability estimate" (Wright, 1969, p. 24). Another major distinction between Rasch theory and classical theory pertains to item discrimination. "The classical perspective assumes that all items are parallel; have the same ratio of true score information; relative to error score information, and that error score variance is

randomly distributed with a mean of zero" (Ryan & Hamm, 1976, P. 1). On the other hand, Rasch analysis assumes equal item discrimination.

Lack of equal item discrimination produces a test with poor fit to the model. In that a measuring instrument should be calibrated using one unit of measure, Rasch recommends that items be refined to produce good fit to the model and thus maintain objectivity rather than adding parameters to account for other parameters as item discrimination. (Bring, 1970)

Test data for which the Rasch model is applicable must have two other properties:

First, all items must have equal discrimination. That is, the rate at which the probability of passing the item increases with total score must be equal for all items. The Rasch model does not contain a parameter for item discrimination.

Second, there must be minimal guessing so that the probability of passing an item by chance is minimized. (Whitely & Dawis, 1974, p. 166)

The substitution of equal item discriminations, rather than maximum item discriminations as a goal in item writing, may appear counter-intuitive to the test construction expert steeped in classical test theory. While it is true a highly discriminating item is capable of providing more information concerning the placement of an individual on the continuum of some latent trait, the highly discriminating item functions over a narrower range of abilities than a less discriminating item. An item with perfect discrimination would provide complete information about a single point on the ability continuum and no information about any other point. Therefore, for any given test, there will exist an optimal range of discrimination. If the test characteristic curve is to rise steeply through a narrow range of abilities, more highly discriminating items will be desirable than if the test is to function over a broad range of abilities. (Dinero, 1976, pp. 14-15)

Rasch item analysis theory requires that the test designer control discrimination and guessing, while classical theory treats them as item parameters. "This means that variation in additional item characteristics like guessing and discrimination must be dealt with during the construction and selection of items for the final sample-free

pool. The aim is to create a pool of items with similar discrimination and minimal guessing" (Wright, 1969, p. 23). Thus, by controlling item discrimination and guessing through design and selection of test items, Rasch theory suggests that they may be ignored in the determination of item-difficulty and person-ability. But they are ignored only after the best possible effort is made to minimize their effect. This is done during the item calibration procedure which determines how well items fit the model. Those who are uneasy with such an approach might consider the alternative classical theory offers. Classical theory is no better. The more general models used in classical analysis lead nowhere in their attempts to deal with guessing and discrimination.

The estimation of discrimination while frequently attempted, is clouded by uncertainties as to whether it can in fact be reliably estimated. The values actually obtained for a particular set of items are highly sensitive to the particular distribution of person abilities which happen to occur in the calibrating sample . . . The estimation of guessing is even more obscure. (Wright, 1977, p. 220)

By selecting items which fit the model, discrimination and guessing, under Rasch theory, can be effectively controlled as a source of score variability. Discrimination is dealt with in the item selection phase of test design. The objective is to choose items having a discrimination index reasonably close to 1. Ideally, all items used in the test will have identical item characteristic curves across the entire ability spectrum. If a reasonable explanation for extremes in the discrimination index or atypical ICC curve can not be found, an item should not be used. That item does not fit the Rasch model. Guessing, on the other hand, must be considered after each test administration. Guessing behavior becomes apparent when response patterns are presented in difficulty order. For example, one such pattern might be revealed when an individual shows consistently good performance as item-

difficulty increases, up to a point, and then misses a number of more difficult items before getting some items near the top of the difficulty scale correct. It is possible that the string of correct responses at the lower end of the difficulty scale better represents the person's true ability and that the few correct responses at the top end of the scale resulted from guessing. The latter items might be eliminated from the score during the analysis in an effort to gain a more correct measure of ability. The decision to do so is a matter of personal judgement on the part of the evaluator and should be made after consideration of all the pertinent facts.

Whitely and Dawis (1974) are foremost among an extremely small group of testing experts who do not give the Rasch model their full support. With the exception of Whitely and Dawis, all of the researchers so far noted do support application of the model, and it would be very easy to catalogue still more names of persons who support it fully or with very limited reservations. The articles by Whitely and Dawis (1974, 1977) are, for the most part, the only notable exceptions so far observed in the literature. Among the half-dozen criticisms leveled against the Rasch model by these authors, two are relevant to this research. Whitely and Dawis suggest that the Rasch model requires that:

1. The calibration procedure should employ very large samples.
2. Test forms developed from items in a calibrated pool should result in scores of equal variance. That is, the resulting test forms should demonstrate "reliability" in the classical sense.

Wright challenges both contentions (1977). The Whitely-Dawis position on sample size is that:

The shift in emphasis from populations to raw score groups has one important operational application: huge N's are required. Unlike classical item analysis, each score group is used to give independent estimates of item parameters. However, even when as many as 500 persons are used for item calibration, extreme scores may not be obtained frequently enough to provide very stable estimates of the [probabilities]. Even if scores on a 50-item test formed a perfect rectangular distribution, for instance, a total N of 500 would produce only 10 persons per score group. Typically, however, mid-range score groups have very high frequencies and extreme score groups may have few or no observations at all. Although the [probabilities] can be estimated from the model, the need for very large N's during test development should be obvious. (1974, pp. 168-169)

Wright's rebuttal to this point suggests that Whitely and Dawis did not understand the Rasch model (1977):

Whitely and Dawis (1974) . . . recommended a two-stage estimation procedure which is actually unnecessary and impractical. As a result, they conclude that only huge sample sizes make it possible to apply the Rasch model to real data. [Since] the Rasch model can be and has been applied productively to sets of data as small as 100 persons, and under these circumstances lead to useful results, it is important to correct this misunderstanding and the incorrect conclusion drawn from it. (p. 219)

Wright went on to develop the sampling theory appropriate to the Rasch model, including several useful formulas, and concluded:

These findings, coupled with the information in Table 1, lead to the conclusion that calibration sample sizes of 500 are more than adequate in practice and that useful information can be obtained from samples as small as 100.

But Whitely did not back down (1977). Her rejoinder did not so much refute Wright's reply as point out that the real issue was a matter of statistical power; that, while possibly technically sound in a way, Wright's recommendations on sample size would not support a sufficiently powerful test of fit to the Rasch model, an issue not raised in the first Whitely- Dawis article:

Given the importance of testing fit, and the need for a reasonably powerful statistical test, successful application of Rasch's model requires large sample sizes at some phase in the test development process. Since the power of a test of fit is dependent on N, the choice of sample size should be guided by the degree of departure

from the model that the test developer wishes to detect. At the extremes, a sample of several thousand can detect trivial departures, while a small N (less than 800) fails to detect sizeable differences. (p. 231)

So it would appear that Wright feels that samples of 500 are more than adequate while Whitely can not support use of a sample of less than 800. Statistical power is considered to be an important factor in this study. Students taking MEAP tests must pass four out of five items to receive credit for an objective and it is on the basis of the percentage of objectives passed that remedial education decisions are based in Michigan. Assuming the pass/fail criterion remains unchanged, when a single item from a set of items comprising an objective is eliminated because the item does not fit the Rasch model, the odds against a student passing an objective are significantly increased. If two items are eliminated, the student has no chance at all of passing the objective. Therefore, the statistical power and sample size appropriate to this investigation will be a major consideration.

The Whitely-Dawis position on parallel test forms favors classical theory:

The empirical results generally substantiated the theoretical interpretation of the nature of equivalent forms from the Rasch model. Only under extreme conditions did the measurement errors fail to account for the observed differences between subjects. However, none of the subsets were equivalent in the traditional sense. Alternate form correlations were only moderate, and there was some evidence that precision might have been increased by using more efficient techniques in selecting items. Although the Rasch item parameter may be invariant over populations, precision is specific to the trait distribution in a given population. If the goal of item selection is to develop fixed content tests, then the classical techniques of having item difficulties close to .50 and matching extreme item difficulties will yield the most precise equivalent forms. (Whitely, 1974, pp. 163-178)

Wright, emphatically, did not agree with Whitely-Dawis on this point either:

The Whitely-Dawis implication on page 176, that traditional equivalent forms are in some way better because they maximize precision and statistical equivalency, is incorrect. The precision of a measure depends on the relevance of the item to the target of measurement and the number of items used . . . The more essential question is, do the set of items all bear on a single common latent variable? If they do not, then the set of items contain a mixture of variables and there is no simple, efficient or unique way to know their utility for measuring. (Wright, 1977, p.223)

This point is also important to this research because, while it is not the primary intention to develop parallel forms, that will be the result. The calibration process can be expected to result in the elimination of MEAP test items. This will be done to identify those items which fit the Rasch model and those that do not. The purpose will be to devise a more precise set of test items on which to judge the reading test performance of Michigan fourth and seventh grade students. As Wright puts it, "The opportunity to determine whether or not there is a possibility of objective measurement in some data, by checking their fit to the Rasch model, represents the model's most important contribution to the scientific method" (Wright, 1977, p. 223). In effect, the Rasch model will be used in this research to improve measurement efficiency. Whitely-Dawis suggest that this is not possible: "Although the use of the Rasch model cannot improve precision of fixed-content tests, the special properties of the latent trait model permit the desired degree of precision for any person to be obtained from the fewest possible items" (Whitely-Dawis, 1974, p. 177). They possibly intended their statement to mean that efficiency of an existing test could not be improved in its original form. However, since much of the effort in this investigation relates to the possibility of improving the efficiency of existing MEAP reading tests by re-scoring items that do not fit the Rasch model, this could be a troublesome assumption. It



is contrary to the relatively common practice of improving the reliability of norm referenced tests by increasing their length. There is support in Rasch theory for the concept of improving criterion referenced test efficiency, on the other hand, by dropping items from a test which do not fit the Rasch model. An actual instance of support was found in the work of Joseph P. Ryan and Debra W. Hamm (1976):

In summary, the procedure described in this paper offers practical advice to a teacher who wishes to maximize [a test's] reliability. The teacher might be interested to know that adding additional parallel items to the test will theoretically increase its reliability the next time it is used.

For a teacher who wishes to score students on a test they have already taken, however, it is more useful to provide a procedure that can increase the reliability of the test by deleting items from the existing data set. (Ryan-Hamm, 1976, p. 8)

Those who are familiar with classical item analysis probably recognize the principle of increasing test reliability by adding items. It is not so likely that they will be equally comfortable with the idea, under Rasch theory, of increasing test efficiency (i.e., reliability) by deleting items. Ryan and Hamm (1976) provide a very illuminating discussion of this point:

In contrast [to classical theory], the Rasch model argues that test items are not necessarily parallel and consequently some items more accurately manifest the latent trait being measured than other items. When some items do not measure the same trait, error information will differ among items which necessarily implies that at least some items will add error information faster than true score information. Items which do not fit the Rasch model are those which add error information at a higher rate than the rest of the items on the test . . . (p. 8)

Eliminating items that do not fit the Rasch model should increase the reliability of a test because it will delete non-parallel items, the items with greatest error variation. Magnussen (1967) suggests the reasonableness of this assertion when he writes:

The internal consistence coefficient we obtain from KR20 will therefore be directly dependent on the correlation between the item in the test, i.e. on the extent to which the items measure the same variable. The more homogeneous the items are, the greater the numerical value of KR20 will be for a given number of items in the test. (p. 117)

Ryan and Hamm support this position as well: ". . . generally the highest reliability is achieved after deletion of items that do not fit the Rasch model" (Ryan-Hamm, 1976, p. 5).

Empirical evaluation of theory and applications of the Rasch model is generally supportive. For example, 21 doctoral dissertations, which were characterized as being potentially relevant to the objectives of this research, were done between 1965 and 1979, and the abstracts of only five of these evidenced any reservations about the efficacy of the model. And, every one of these reservations was qualified. The support for the model in the professional journals has been even more positive, and the content in which the model has been applied surprisingly diverse:

Achievement testing: writing (Wells, 1973); nutrition (Passmore, 1974); military training (Epstein, 1975);

Analysis of simulated test data: non-normal data (Cypress, 1972); discrimination (Cartledge, 1975); item performance (Forster, 1976); item discrimination (Dinero, 1976);

Psychometric theory: test model comparison (Hamm, 1977); effects of teaching on item calibration (Luska, p. 979);

Qualification examination: military (Anderson, 1969);

Test design: best test design (Douglas, 1975); item bias (Draba, 1978).

Other applications in reading achievement were previously noted. The overall impression given by the literature concerned with the application of the Rasch model is that its appropriateness in testing measurement and evaluation is very broad. It appears to be limited more

by the imagination of the psychometrician or researcher than anything else. Certainly its application to the analysis of MEAP reading tests, as proposed in this research, seems to be appropriate. It may even be that application of the Rasch model in this instance is especially important given the fact that MEAP tests are criterion-referenced tests. Epstein (1975) thought so: "This independence from the item set puts the major emphasis on the individual's ability. Thus [the Rasch model] seems philosophically attuned to criterion-referenced testing."

#### RELATED LITERATURE

The material in this section is primarily comprised of articles and official publications directly related to the Michigan Educational Assessment Program; its history, philosophy, testing format, and target population. Two sources have produced most of this material: 1) Most of the articles have been written by the former State Superintendent of Education, John W. Porter; 2) The Michigan Department of Education has published a profusion of pamphlets, booklets, and studies specifically related to the program. Since the majority of the state material was developed and published during the tenure of Porter, it reflects his very supportive view of the MEAP Program.

The rapid and sometimes controversial growth of the Michigan Educational Assessment Program since its inception in 1969 has generated numerous paths of debate which, regretfully, must be ignored in this research in an effort to concentrate on the test development aspects of the program.

Porter wrote eight articles on many of the more important aspects of the MEAP program (Porter, 1968, 1972, 1973, 1975a, 1975b, 1976, 1977, 1978). These articles deal rather comprehensively with MEAP abuses,

assumptions, consequences, criteria for item selection and test evaluation, definition of terminology and program concepts, demonstrable results, description of objectives, history, issues and problems, objectives, philosophy, procedure, reports, theory, and uses. The development of this major testing program has been dramatic and far reaching in the State of Michigan. There are probably very few state testing programs anywhere approaching the magnitude of this one, and it is even less likely that any testing program had greater political impact in education than this one in most troubled times. The MEAP program has been a landmark in mass state testing.

The Michigan Department of Education has dissected the MEAP program into relatively small and digestible components by publishing numerous pamphlets and monographs which concentrate on very specific aspects of the program. An excellent historical summary to the MEAP program was found in two state publications: "Report of the Michigan Educational Assessment Program's External Advisory Panel on Evaluation" (1977) and "The Status of Basic Skills Attainment in Michigan Public Schools" (1979). The first of these publications is especially interesting in that it is a more or less objective evaluation of MEAP done by third parties under contract to the State Department of Education. It deals with a number of basic questions relevant to the program. Reference to test objective selection and item validity are especially pertinent to this study.

Helpful discussions on various aspects of MEAP philosophy and policy were found in a number of state publications. Three of these are worth special note within the context of this study: "First Report, Objectives & Procedures of the Michigan Educational Assessment Program,

1974-1975" (1975), "Questions about the Michigan Educational Assessment Program" (no date), and "A Staff Response to the Report: An Assessment of the Michigan Accountability System" (1974). This last publication is unlike the other two in that it was prepared in response to an appraisal done by interests external to MEAP, the National Education Association and the Michigan Education Association, who saw their interests threatened by the MEAP program. The two groups published jointly a report on April 12, 1974 entitled An Assessment of the Michigan Accountability System (1974). This report and the "Staff Response" published one month later by the Michigan Department of Education provided good perspective on some of the more heated issues raised by MEAP in its earliest development. Most of these issues have no relevance to this investigation, but these two references do treat interesting issues regarding test validity and compensatory education.

Material dealing with MEAP test item development, test format and administration, and target population are covered in the following publications put out by the Michigan Department of Education: "Grades 4 and 7 Item and Objective Handbook" (no date), "First Report Objectives and Procedures 1974-75" (1975); "Communications Skills Objectives - - Reading--Speaking/Listening--Writing" (1979), "Questions and Answers about the Michigan Educational Assessment Program" (no date), and "Student Performance Expectations" (no date), "Interpretive Manual 1978-1979" (1979).

#### SUMMARY

A review of the background and related literature has demonstrated the general applicability of Rasch analysis theory to the criterion-referenced reading tests used in connection with the Michigan

Educational Assessment program for the fourth grade and seventh grade levels.

While the Rasch model appears to be moderately challenging on first acquaintance, application of the model also seems to be straightforward and certainly no more difficult than established classical test item analysis methods.

## CHAPTER III

### RESEARCH DESIGN - METHODS, PROCEDURES, AND LIMITATIONS

#### PREPARATION OF DATA

##### INTRODUCTION

Fourteen samples, comprised of 1,000 student records each, are used in this analysis. They were drawn from "Sample-of-5,000," data which was developed from MEAP test results for 1973 through 1979. Each Sample-of-5,000 was drawn, at random, from the records of every student taking the test each of these years.

Though MEAP test data is the property of the State of Michigan, no one in Michigan has direct access to complete records once tests are scored, not even State employees directly involved in the assessment program. All tests are scored in Iowa City, Iowa by Westinghouse Datascoresystems. This organization maintains the only complete files of MEAP test results that exist. Westinghouse Datascoresystems prepares Sample-of-5,000 data as part of the MEAP scoring and analysis services it provides to the State of Michigan. However, all of the information which would make it possible to identify individual students has been removed from each Sample-of-5,000 record. Actual use of this edited data is carefully controlled by the Michigan Department of Education, Michigan Educational Assessment Office in Lansing, Michigan.

##### DESCRIPTION OF THE SAMPLE-OF-5,000

The designation "Sample-of-5,000" is used by Michigan Educational Assessment Program staff personnel (Roerber, 1980) to identify the sample of fourth grade and seventh grade students taking the MEAP Test each year since the 1973/1974 test. These samples have been developed

specifically for research purposes for use by state offices and others having appropriate research interest in MEAP Tests. The sampling procedure employed is the same each year. The following description of sampling the 1976/1977 test is typical:

With certain exceptions, all students in the fourth and seventh grades receiving regular classroom instruction (i.e. instruction including mathematics and reading) were tested during the period of September 13 - October 1, 1976. Make-up tests occurred between October 4 and 8, 1976. A total of 291,647 public school students completed the mathematics and reading tests: 136,472 were in the fourth grade and 155,175 were in the seventh grade. In addition, 13,345 non-public school students completed the seventh grade MEAP tests without cost to their schools under Title III Elementary and Secondary Education Act State Plan, with the approval of the Michigan State Board of Education. . . .

#### SAMPLING PROCEDURES

The technical characteristics of the MEAP mathematics and reading tests were based on a sample of approximately 5,000 student scores drawn from each grade level. A replicated systematic sampling procedure was employed to select the sample. The procedure, on the average, yields estimators as precise as those yielded by a simple random sampling procedure when the population of scores is in random order. . . .

Spacing factors in the "every Kth" systematic samples were:

$$\text{Grade 4: } k = 10(136,472)/5,000 = 272,944$$

$$\text{Grade 7: } k = 10(155,175)/5,000 = 310,350$$

At the time the sample scores were selected, 136,472 and 155,175 were the number of student assessment booklets for the fourth and seventh grades respectively. Additional assessment booklets received after the samples were drawn increased these numbers to 136,858 for the fourth grade and 155,632 for the seventh grade.

The spacing factors were rounded to 273 and 310 for the fourth and seventh grades respectively. Ten random numbers were chosen from a table of random numbers for each of the grades. For the fourth grade, the numbers were 15, 49, 59, 131, 137, 180, 232, 268, 269, and 272; for the seventh grade, the numbers were 12, 19, 52, 79, 88, 153, 204, 222, 271, and 274. These random numbers were originally chosen for the 1974-75 MEAP data. These numbers were the first elements in each of the ten systematic samples for the respective grades.

The second elements were obtained by adding the spacing factor (K) to each of the first elements. For example, with Grade 4:  $15 + 273$



= 288, 49 + 273 = 322, 59 + 273 = 332, and so on. . . The replicated systematic sample of 5,000 was then obtained by combining the ten samples for each of the two grades. (Michigan Department of Education, 1977)

#### OBTAINING SAMPLE-OF-5,000 DATA

To obtain access to Sample-of-5,000 data, approval must first be obtained from the supervisor of the Michigan Educational Assessment Program. The request for data used in this research had to be made on the "Michigan Department of Education Data Request and Assurances Agreement" form, RA-2969-A. Requests for this form may be directed to: Edward Roeber, Supervisor, Michigan Educational Assessment Program, Box 30008, Michigan Department of Education, Lansing, Michigan 48909. See Appendix A.

The Sample-of-5,000 data actually had to be obtained from three sources:

1. The (Michigan) State Systems and Programming Unit in Lansing, Michigan.
2. The Computer Laboratory, Michigan State University, East Lansing, Michigan.
3. Westinghouse Datascoresystems, Iowa City, Iowa.

The 1973 through 1976 samples had to be purchased from Westinghouse Datascoresystems.

The 1977 through 1979 samples were in State of Michigan files at two locations. The 1977 and 1979 data was available in Lansing, Michigan directly through the State Systems and programming Unit. The 1978 samples were at the Michigan State University Computer Lab in East Lansing, Michigan.

There are 53 data elements in every student record in each sample. The variables which have relevance to this study had to be anticipated and their physical location in the file ascertained so that a uniform record, for every sample, could be prepared for input to the Rasch analysis computer program.

Since all of the data used in this analysis were on computer tapes, it was necessary to have detailed tape layouts to access the variables needed. The necessary tape formats, showing complete data organization and coding information, were available for only six of the seven years. There was, unfortunately, no tape format available for the 1978 sample. Though the number and sequence of variables in all of the samples did not change to any appreciable degree, their physical location in terms of actual columns did change substantially from year to year. By careful study of the tape formats that were available (i.e., 1973 to 1977 and 1979), and the tape dumps (i.e., printouts of a few complete records on the tapes) which were available for every year, it was possible to reconstruct a tape format of the 1978 data. Variable locations and coding were verified by comparing location and code information provided by the tape formats directly to the partial printouts (dumps) of each file. A summary table was then prepared showing the column locations and coding specifications for the ten variables being considered in this study.

A key step to this analysis was the definition of those test questions which corresponded to the MEAP test objectives, by year. There were 23 objectives in the 1973 MEAP test for both fourth grade and seventh grade students. But, from 1974 through 1979, the number of objectives was reduced to 19 objectives for fourth graders and 20

objectives for seventh graders. The reduced number of fourth grade and seventh grade objectives that were retained were originally part of the 1973 MEAP tests. Once established for the 1974 test, there were no further changes in test objectives through, and including, the 1979 tests. The source of this information is the INDIVIDUAL STUDENT REPORT FORM which shows items by objective. Examples of the 1973 - 74 and 1979 - 80 forms appear in the Appendix (See APPENDIX B: "GRADE 4 INDIVIDUAL STUDENT REPORT FORM and GRADE 7 INDIVIDUAL STUDENT REPORT FORM").

With the exception of 1974-75, a complete set of these forms was obtained from the Michigan Department of Education Assessment Office. The items in the 1974 tests which matched specific objectives were determined by comparing 1974 test items to 1973 and 1975 test items and matching related objectives. Through this process of matching actual items as they appear in their respective tests, and then verifying that the same objectives applied both in the 1973 and 1975 tests, it was possible to identify which objectives belong to the 1974 items.

Once the variables of interest in this study had been identified, actual coding of the computer routines could begin which would culminate in the application of the Rasch analysis computer program developed by the Measurement and Statistical Analysis Laboratory, The Department of Education, The University of Chicago.

#### REDUCING THE DATA SAMPLES AND SUBSEQUENT ANALYSIS

Once the Sample-of-5,000 data had become available and the relevant variables had been identified, it became necessary to devise an orderly approach to selecting appropriate samples from the enormous amount of MEAP data available; approximately 70,000 student records. Actually none of the Sample-of-5,000 files contained exactly 5,000 records. The

smallest sample was 4806 fourth graders for 1976. The largest sample was 5557 seventh graders for the year 1974.

The decision was made to randomly select 14 samples of 1,000 students each using the SPSS procedure SAMPLE and to format these samples with the ten variables chosen for this analysis in identical locations using the SPSS procedure WRITE CASES.

#### LIMITATIONS OF STUDY

##### A TWO PARAMETER MODEL

The key assumption associated with the Rasch model is that item-difficulty and person-ability are the only two variables worth considering in test measurement situations. "There has been a running debate for at least fifteen years as to whether or not there is any useful way by which some kind of estimates of item parameters like item discrimination and item 'guessing' can be obtained" (Stone & Wright, 1979, p. ix).

The inevitable resolution of this debate has been implicit ever since Fischer's invention of sufficient estimation in the 1920's and Nyman and Scott's work on the consistence of conditional estimators in the 1940's. Rasch (1968), Anderson (1973, 1977) and Barndorff-Nielsen (1978) each prove decisively that only item-difficulty may be estimated consistently and sufficiently from the right/wrong item response data available from item analysis. These proofs make it clear that dichotomous response data available for item analysis can only support the estimation of item difficulty and that attempts to estimate any other individual item parameters are necessarily doomed. (p. ix)

There are five important assumptions associated with the Rasch model:

1. Test calibrations are independent of the sample of persons used to estimate parameters. (Panchapakesan & Wright, 1969, p. 23).

2. Person measurements, the transformation of test scores into estimates of person-ability, are independent of the selection of items used to obtain the scores. (p. 23)
3. Variation in additional characteristics like guessing and discrimination must be dealt with during the construction and selection of items for the final sample-free pool (p. 24)

In those circumstances when the above three assumptions apply, then it follows that:

4. No assumptions need be made about the distribution of ability in the target population or in the calibration sample. (p. 24)
5. The number of correct responses to a given set of items is a sufficient statistic for estimating person-ability. (p. 24)

#### THE RASCH "FIT STATISTIC" DETERMINES MEAP ITEM FIT

Determination of test item fit to the Rasch model is the primary procedural objective in this investigation. While a fit statistic has been used as the sufficient statistic for this purpose, a number of questions had to be resolved before the decision was made to do so.

In the early stages of this investigation, determination of item fit seemed to be rather simple. The literature on Rasch measurement gives the impression that the process of determining item fit is no more difficult than the application of an appropriate statistic to the data. To facilitate the process, there is a computer program available to refine test data into a series of neat tables and charts. However, it has become apparent that determination of item fit may not be, for the

present at least, such a simple matter. The program is efficient, and it produces great quantities of useful information. It is based on theory that is apparently sound in every important respect. However, the program does not determine item fit to the Rasch model; it only provides statistics which may be used by an informed test analyst to estimate fit probability. Further, though a number of familiar terms have been applied to the fit statistics produced by this program, such as MS,  $\chi^2$ , F-statistic, and t-statistic, the fit statistic is really none of these despite the label in current use. All four of these designations have been applied to the fit statistic used in Rasch analysis over the past five years. Despite comparisons to traditional statistical measures, it is more likely true that the Rasch fit statistic is none of these despite any legitimate basis of comparison which may be employed to suggest that it is. The Rasch fit statistic is very likely to be a new statistic in its own right.

#### EASY ITEMS THAT DO NOT FIT THE RASCH MODEL SHOULD BE RETAINED IN MEAP TESTS

The statistically critical value which sets the lower limit of the area of rejection for interpreting the fit statistic in this study was set at alpha equal to 0.05.

Items that do not fit the Rasch model do not do so either because they are judged to be too easy or to be too hard. No item is judged to be too easy in this study. This analysis is concerned only with items which do not fit the Rasch model because they are considered to be too difficult. Easy items will not be rejected at any level of significance here because of the underlying assumption that all items in a criterion referenced test have content validity, even the easiest item possible.

This assumption follows from the fact that items in MEAP tests measure specific learning objectives. If the learning objective is easy, an easy item should be used to measure it. This principle applies to every criterion referenced test. Norman E. Grunlund makes this point in his text Measurement and Evaluation in Teaching (Grunlund, 1976, p. 153):

The difficulty of the test items in a criterion referenced mastery test is determined by the nature of the specific learning tasks to be measured. If the learning tasks are easy, the test items should be easy. If the learning tasks are of moderate difficulty, the test items should be of moderate difficulty. No attempt should be made to modify item-difficulty, or to eliminate easy items from the test, in order to obtain a range of test scores. On a criterion-referenced test, we should expect all, or nearly all, pupils to obtain perfect scores when the instruction has been effective. (Grunlund, 1976, p. 153)

It is assumed that easy items measure an appropriate aspect of the reading trait without any consideration being given to the possibility that students with low scores may get them right. It could be argued, perhaps, that the same assumption should be applied to difficult items found in this analysis. This argument may not follow however. Items which are too difficult for the level of ability being measured by MEAP reading tests probably do not belong in these tests unless a clearly defensible reason can be shown for retaining them in spite of their great difficulty.

#### TYPE I ERROR AND TYPE II ERROR

The consequences of TYPE I or TYPE II errors were carefully considered in setting the limits of the rejection area:

TYPE I ERROR: Failure to accept a true null hypothesis (i.e., observed item difficulty equals predicted item difficulty).

TYPE II ERROR: Failure to reject a false null hypothesis (i.e., observed item difficulty equals predicted item difficulty).

Commission of a TYPE I error in this study amounts to rejection of an item on the grounds that it is too difficult, when it is really not too difficult. As the research design entails re-scoring wrong answers on items judged to be too difficult, the consequences of this type of error would be to give a test taker credit for a wrong answer when that wrong answer should have been allowed to stand.

Commission of a TYPE II error in this study amounts to accepting an item on the grounds that it is not too difficult, when it really is too difficult. Again, as the research design entails re-scoring wrong answers on items judged too difficult, the consequences of this type of error would be to fail to give a test taker credit for a wrong answer when the wrong answer should not have been allowed to stand. This permits a test item to remain in the test which is a false measure of the underlying knowledge variable, and the test taker is materially harmed as a result.

The consequences of a TYPE II error outweigh the consequences of a TYPE I error. Students penalized for getting an item in the MEAP reading tests wrong when that item does not belong in the test in the first place are penalized unreasonably. To pass a learning objective in a MEAP test, the student must get four out of five items that measure the objective right. When one item is too difficult, the odds against passing the objective increase substantially. When two items are too difficult, there is no chance of passing the objective at all.



Therefore, TYPE II errors were guarded against in this analysis by attempting to increase the statistical power of the analysis two ways: 1) use a large sample in an effort to reduce the size of the standard error; and 2) set the alpha level equal to 0.05.

Alpha designates the probability of committing a TYPE I error. Beta, on the other hand, is the probability of committing a TYPE II error. Anything which can be done to reduce the size of Beta will reduce the chance of a TYPE II error, and consequently serve the objectives of this study.

Statistical power is represented by the complement of Beta (i.e.,  $1 - \text{Beta}$ ). Consequently, any step which reduces the size of Beta has the concurrent effect of increasing the statistical power of a test. There are basically only three things which can be done to increase statistical power: 1) increase the distance between the observed score and the expected score; 2) reduce the amount of sampling variability by either, a) increasing sample size, or b) reducing the source of extraneous error; or 3) increase the size of alpha.

There is no opportunity to control discrepancies between observed and expected responses in this investigation. Consequently, the first approach to increase statistical power was not available.

Statistical power could be increased by reducing sampling variability because there was opportunity for extensive control over sample size. Therefore, in an effort to reduce the size of the standard error in this study, a moderately large sample-of-1,000 test takers was chosen. Such a sample would certainly be considered "large" by Wright.

Sample sizes of four hundred and (sic) eight hundred persons were used (in a simulation study) because in principle, given the model, four hundred suitably chosen persons should be enough to determine

characteristics effectively, and eight hundred persons should be more than enough (Wright & Mead, 1980, p. 25).

There was no possibility of reducing sources of extraneous variability in this study, so this second alternative approach to reducing the size of the standard error could not be employed in this study.

Finally, the largest possible alpha value was chosen to increase statistical power, consonant with accepted practice in good research design. Alpha values of 0.05 and 0.01 are commonly found in social research. Both were considered carefully before settling on the value 0.05. Probably this level is sufficiently sensitive as not to subvert the intentions of the MEAP test designers too greatly by identifying false nulls. This value affords only one chance in twenty of incorrectly identifying an item's fit to the Rasch model. The unfavorable consequences of decreasing alpha to 0.01, or even lower perhaps, far outweigh any perceived benefits.

#### STATISTICAL ANALYSIS

##### SELECTING COMPUTER SOFTWARE

Six computer software systems were employed in this investigation. Each was chosen from the great variety of such systems maintained by the Wayne State University Computer Services Center for specific properties which would contribute to the objectives of this research. The computer software systems chosen were:

1. The "Michigan Terminal System" (MTS): This software package was developed by the University of Michigan to provide a "language" by which users could communicate with the University's computer system. Through MTS the user can

create computer files; store data; write and store computer programs; and utilize the great variety of other software systems maintained at the Computer Center for data manipulation and statistical analysis.

2. \*FS: The "File Save" software system is one of a number of "utility programs" developed and maintained by Wayne Computer Services Center personnel. Such systems facilitate frequently used file handling and data manipulation computer procedures which, though they are likely to be useful in connection with a great variety of more specialized applications, represent a relatively minor role in relation to those applications. This study is a good example. At various times, twelve different computer tapes were involved. Frequently the organization of the data on these tapes differed, and the quantity of data was too large, in every case, to be stored on disk files. Before records could be selected and organized for presentation to the Rasch analysis computer program, which was the primary data processing objective in this study, a means had to be found for manipulating and storing the massive quantities of data involved. It would have been extremely difficult and time consuming to write the individual computer programs needed. The \*FS system makes this unnecessary. The system was used exclusively in this investigation to move data between computer tape and temporary disk file storage facilities and to preserve the results of numerous computer analysis runs.

3. The "Statistical Package for the Social Sciences" (SPSS):  
This is one of several software systems maintained by Wayne Computer Center personnel which is especially designed for the manipulation and statistical analysis of large quantities of data. This particular system has been leased from SPSS Incorporated of Chicago, Illinois. While SPSS is technically a batch oriented system, it is interactive in the sense that success or failure of each computer run becomes immediately apparent during run execution. The user is then able to institute corrective measures which may ultimately salvage the investment being made in a particular run that results from using large data files. In this investigation, the quantity of data was so great that investment in permanent disk storage was neither practical nor cost efficient. Data had to be stored permanently on computer tapes. To provide a statistical analysis, these data would have to be restored from tape to temporary disk files which are automatically destroyed when the user signs off the system. Charges for mounting a computer tape and tape drive use often could exceed charges for the use of the central processing unit (CPU) of the computer. Therefore, it was essential to this investigation to be able to repair bad code during a computer run which would occasionally become apparent despite satisfactory runs of the same code on test data. SPSS was used extensively in this study to select the variables needed for analysis from the original MEAP Sample-of-5,000 tapes. The fourteen samples used in this investigation were selected

using SPSS routines. Each sample contains 1,000 students. The data set consists of 10 of the 53 variables in the original data. Each of the samples was drawn from records with varied formats and organized into a standardized format which was then input into the BICAL computer program. SPSS was also used to generate descriptive statistics for each of these samples.

4. BICAL and BICAL.3: These are two versions of a highly specialized computer program used to score objective tests and perform Rasch analysis on the results. This program has been under continual development by the Measurement and Statistical Laboratory (MESA), The Department of Education, The University of Chicago, since the early 1970's. There have been several versions of this program available over the years which have been released periodically as new aspects of Rasch analysis are developed by MESA. Two of these versions, BICAL and BICAL.3, were considered for use in this study. BICAL.3 was the latest version available at the time this was written. The older version of the Rasch analysis program, BICAL, was originally published in 1975. This was the version sold to Wayne State University in the Summer of 1978, and, it is assumed, constituted "state of the art" development up to that time. BICAL was used to develop the Rasch analysis statistics in this study after investigating the appropriateness of BICAL.3. The rationale for basing this research on BICAL is explained in APPENDIX C. All fourteen samples employed in this study were run against both

BICAL and BICAL.3, and some comparisons are drawn between the two versions of the program in APPENDIX C. But, briefly, BICAL was chosen as the basis for the Rasch analysis which was done in this investigation because it appeared to be more suitable to the data used and objectives sought. BICAL was employed in this study to measure test item-difficulty and to measure individual person ability as represented in each of the fourteen samples used in this study. Rasch analysis does not, strictly speaking, require a computer, but "large" numbers of items or tests subjects demand it in a practical sense. What constitutes a large number may be open to question in some cases, but not here. There are 14,000 subjects and up to 115 items encompassed by this study. The Rasch analysis involved would clearly be impossible without the aid of a computer. The computer program used in this study is the most important procedural component. On the one hand, the study would have been impossible without using either BICAL or BICAL.3, since they embody the only procedural implementation of Rasch analysis that appears to be available at this time. On the other, changes to underlying interpretation of statistics generated by the Rasch analysis program, observed while considering which version to use here, raise some interesting question about the "objectivity" of Rasch analysis. These questions too are discussed in the closing recommendations section of Chapter V and the conclusion of APPENDIX F.

5. \*TEXTEDIT: This is a software system which has great value in text processing and output. It was developed by Dan Fox, who is the Assistant Director of the University Computer Center at the University of Michigan. TEXTEDIT was implemented in the Spring of 1981 at the Wayne State University Computer Services Center. It constitutes the most recent of five text processors currently supported at Wayne University. It is very easy to use and has particularly powerful table generating capabilities which proved to be very useful in preparing the text for this dissertation. The revision and duplication power of this system was relied on heavily in the development of this manuscript at every point from inception to completion.
6. \*PGF: This is a software system which provides alternative means for formatting and printing computer output on the Xerox 9700 printer available at the Wayne State University Computer Services Center. All printed output generated in this study has been produced on the Xerox 9700 printer using \*PGF.

#### DEFINING THE VARIABLES

Ten variables were selected from the 53 elements of data available in the individual MEAP records. With the exception of 1973 when "GRADE" was not indicated, the number of variables is identical from 1973 through 1979. There were more test items in the 1973 test than in the tests used from 1974 through 1979. However, the items used in the later years also appeared in the 1973 tests. Ten of these variables have been identified as relevant in this investigation. Each one was assigned a

label that is subsequently used for reference purposes in all computerized data analysis. They are:

<u>VARIABLE NUMBER</u>	<u>VARIABLE LABEL</u>	<u>VARIABLE DESCRIPTION</u>
1	RECTYPE	Record type code.
2	GRADE	Grade 4 and Grade 7.
3	SEX	Girl or boy.
4	AGE	Student age in year/month order.
5	NUMPASS	Number of reading objectives passed.
6	NUMTRIED	Number of reading objectives attempted.
7	RESP###	Item responses where ### is a three digit number from 001 to 115 which corresponds to the question number.
8	OBJ###	Contribution of individual test item to its corresponding objective, where ### is a three digit number from 001 to 115 which corresponds to the question number.
9	OBSCOR##	Score in terms of the number of test items passed (i.e., from 1 to 5) by individual reading objective, where ## is a two digit number from 01 to 23 which corresponds to the objective number.
10	OBSTAT##	Objective status code: 1 = pass; 0 = fail; by individual reading objective, where ## is a two digit number from 01 to 23 which corresponds to the objective number.

The balance of the data in MEAP sample records is not relevant to fourth and seventh grade reading tests which are the focus of this study. Some of the unused portion of these records pertained to tenth grade reading and arithmetic tests initiated in recent years. Most of the data that were not used pertained to the arithmetic tests given the



same fourth and seventh grade students whose reading test results are the basis of this study.

#### DATA ANALYSIS STRATEGY

To obtain the 14 samples and complete the computer analysis required in the objectives of this study, thirteen steps had to be followed:

1. Code a series of SPSS control files which would read each MEAP tape format; randomly select slightly more than 1,000 cases from each file; and write the selected cases out in a fixed format on ten variables chosen for analysis. A typical control card set is shown in APPENDIX E. The control sets differed from each other in two respects: first - the columns designated in the DATA LIST control card had to conform to the MEAP tape formats; and second - the decimal fraction in the SAMPLE control card varied from one MEAP sample to another. The size of the fraction in the SAMPLE procedure was chosen in such a way that the procedure would select slightly more than 1,000 cases in each run. Since the samples varied in size, the denominator of this fraction differed accordingly with each run.
2. Code a series of SPSS control files which will read a sample-of-1,000 records and generate descriptive statistics. Use the SPSS procedure CONDESCRIPTIVE on the continuous variable AGE. Use the procedure FREQUENCIES on all of the other nine variables because they are discrete variables. See APPENDIX D for a typical control set example. The control sets differ from each other in only one significant respect: the number

of questions varied by year and grade. Question responses are individually labelled according to the following pattern: Qln. The capital letter "Q" indicates that the label pertains to a question response. The lower case letter "l" represents placement of a capital letter "A" through "W" in the label which corresponds to one of 23 learning objectives measured by MEAP tests. The lower case letter "n" represents placement of a digit "1" through "5" in the label which corresponds to the first to fifth item intended to measure that objective. For example, the label "QR4" identifies the response to the fourth question which measures the 18th objective in a MEAP test. The coding on QA1 through QW5 must take the following variations into account:

- a. Both 1973 fourth grade and seventh grade samples include 115 questions (i.e., QA1 through QW5).
- b. From 1974 through 1979, fourth grade samples include only 95 questions. Twenty items, designated QC1 through QC5, QN1 through QN5, and QW1 through QW5 which appeared in the 1973 fourth grade reading test were dropped in subsequent tests.
- c. From 1974 through 1979, seventh grade samples include only 100 questions. Fifteen items, designated QC1 through QC5 and QT1 through QT5, which appeared in the 1973 seventh grade reading test were dropped in subsequent tests.

3. Restore the 14 Sample-of-5,000 files to temporary disk files using \*FS. All of these files required over 1,000 disk pages. The largest was 1248 pages.
4. Run the appropriate SPSS WRITE CASES/SAMPLE control set against each Sample-of-5,000, which outputs 1,000+ reformatted records to temporary disk files.
5. Edit each output file from these runs to delete all cases beyond 1,000. The SAMPLE control statement used lacked the capability of selecting exactly 1,000 records, so it was set up to deliberately oversample the Sample-of-5,000 records. Since these files are random, and SAMPLE is a random procedure, which effectively accomplished a random sample from a random sample, this method was adopted to obtain an unbiased sample-of-1,000 records.
6. Run the appropriate SPSS FREQUENCIES and then CONDESCRIPTIVE control set against each Sample-of-1,000, which outputs results on a temporary disk file.
7. Save FREQUENCIES and CONDESCRIPTIVE run results on computer tape using \*FS.
8. Print FREQUENCIES and CONDESCRIPTIVE results using \*PGF.
9. Alter all special missing value coding in sample of 1,000 files set up especially for FREQUENCIES and CONDESCRIPTIVE runs to zeros. This step is necessary to prepare these samples for input to the BICAL program. BICAL does not accept negative values or special characters often used to identify missing values in SPSS.

10. Save the 14 sample-of-1,000 files on computer tape using \*FS.
11. Run the appropriate BICAL control set against each sample-of-1,000, outputting results on a temporary disk file.
12. Save BICAL results on computer tape using \*FS.
13. Print BICAL results using \*PAGEPR.

#### PLANNING COMPUTER TERMINAL SESSIONS

In the development and execution of the data reduction and analysis strategy, it became quickly apparent that considerable planning would be needed for each computer session to avoid inordinately high costs and processing errors. Early sessions ran more than three hours in duration; involved multiple tape and disk file processing; and, occasionally, use of two or more software systems. Costs of these early runs ran from \$50.00 to \$100.00 on several occasions, and the results were not always satisfactory.

Therefore, a method was devised for thinking through each computer session which is based on run logs and the practice of writing out all software instructions in detail, in advance of each run. By writing out instructions in advance, in complete detail, problems could be anticipated and dealt with effectively, if encountered, without danger of wasting costly computer resources. While this approach may appear to be unduly tedious, it becomes increasingly attractive in practice when large amounts of data are being processed infrequently, at odd hours, using costly computer software and hardware configurations. Mistakes, and their undesirable consequences, are reduced to a tolerable minimum.

This research involved repeated processing of similar files, using similar, but distinctly different software routines. Frequently minor

problems would occur during an extended processing sequence which would have been disastrous in a completely unstructured, interactive terminal session. Since a CRT terminal was used most often, when problems did occur the printed record of the session was not immediately available on which an appropriate restart point could be found after the problem was solved. Finally, these coded sequences, along with logs of previous sessions, proved invaluable in debugging problems in run results.

Efforts have been made in this research to generate a session log during each run, and, with few exceptions, this was accomplished. Often it was unnecessary to print session logs at the conclusion of the run, but the benefits of this procedure more than offset the modest trouble and cost entailed. On at least two occasions, two computer runs did not go well, but it would have been impossible to discern the reasons without the session logs.

#### FILE PROCESSING SUMMARY

For the most part \*FS procedures provided a completely satisfactory means for handling tape and disk files used in this study. All files were read from permanent storage on magnetic tape to temporary disk storage for processing. File modification and data analysis was done entirely through disk files. Results were then read from disk near the end of each run onto computer tape for storage and future reference.

SPSS was used exclusively to modify file contents and/or format. The original MEAP samples contained from 900 to 1250 characters of data. The first step toward reducing these files to a manageable format was to read each one individually, taking into consideration initial record format differences, and then writing a smaller sample set which contained only data of interest in this investigation. The layout, or

format, of each of these new samples is identical. SPSS randomly selected the records to be read from the original MEAP data for reformatting concurrently with the reformatting operation. Subsequently, SPSS was used to develop descriptive statistics on the ten variables retained in the new, much smaller, samples.

Two versions of a computer program designed to perform Rasch analysis, BICAL and BICAL.3, were applied to the 14 samples used in this study. Both versions worked well once the proper procedures for running them were worked out. Both versions of the program require a set of user prepared control cards which initiate the program, select various processing options which it offers, and eventually terminates it. Much of what is learned from this study about running BICAL or BICAL.3 had to be learned through trial and error methods. The documentation that is available to instruct a prospective user on control card preparation and program use occasionally is too abbreviated to be much help or it does not cover the topic at all. The procedure followed in this investigation for coding both the BICAL and BICAL.3 control cards, presented in APPENDIX F, will prove useful to the prospective user wishing to run the program for the first time. In addition, it is hoped that this material will provide sufficient information on the subject to enable every prospective user of the program to apply it to any conceivable set of test results with a minimum of difficulty. Reference to the BICAL documentation control card preparation is recommended for the additional perspective which it will provide. In particular, Memorandum 23, Chapter III (Wright & Mead, pp. 43 - 53, 1977c) and Memorandum 23.c, Chapter VIII (Wright & Mead, pp. 87 - 94, 1980) should prove helpful. While the original material is seldom quoted directly here, it is the

primary source of inspiration and factual information on the subject of BICAL control card preparation presented in this chapter.

#### UNDERSTANDING THE FIT STATISTIC

From as far back as 1969 to the present time there have been three approaches to conceptualizing the Rasch analysis fit statistic. In order of historical precedence they were the:

1.  $\chi^2$  interpretation.
2. F-statistic interpretation.
3. t-statistic interpretation.

Very likely, the shift in interpretation which is observable in the MESA literature is the result of refinements which have grown out of greater experience with Rasch analysis. It seems unfortunate that the reference material on using BICAL and interpreting the output of the program does not provide more insight into that experience. A more comprehensive record of the experience of MESA personnel would afford the opportunity to gain better historical perspective on the interpretation of the item fit statistic. This might promote a more complete understanding of the statistic and encourage greater confidence in its use.

All approaches to a fit statistic in Rasch item analysis have a common basis. They begin as measures of variance between expected and observed item performance. The Rasch model estimates expected performance on each item. Fit to the model is then determined from the difference between observed and expected performance. In principle, the process begins with the predicted proportion of test takers expected to get each item correct. These performance estimates are subtracted from the proportion actually observed getting the item correct. Different

predictions apply and different observations are realized depending on whether the reference is to individuals, sample subgroups, or the total sample. However this is the basic measurement concept no matter what the number of persons may be. These measured differences are called item residuals. Since they have both positive and negative values, depending on whether expectations fall short of or exceed actual results, the residuals are squared to eliminate sign. To obtain a standardized measure, the squared residuals are then divided by their standard deviation. The result is a statistic called a standardized residual, or, in Rasch analysis terminology, a FIT Z-SQUARED.

$$Z_G^2 = \text{FIT Z-SQUARED} = \sum_{x=1}^{L-1} \frac{\left( \begin{array}{c} X_{oi} - n_{G} P_{ei} \\ \end{array} \right)^2}{n_{G} P_{ei} \left( \begin{array}{c} 1 - P_{ei} \\ \end{array} \right)}$$

Development of this formula is explained in detail in APPENDIX C. The notation used here is a direct translation, term for term, of the notation used in MESA publications to represent the FIT Z-SQUARED. This translation was needed so that it would be possible to compare discussions of the same equations in these publications. The topic is discussed extensively in APPENDIX C. This is one of several key equations presented by MESA where different symbols were used each time to represent the same variables. APPENDIX C presents, in first person narrative form, notes developed during this investigation while searching for a definition of a "fit statistic" within the context of Rasch measurement. This effort proved to be far more difficult than expected at the outset of this study. Details presented in APPENDIX C



provide a more comprehensive view of this definition, as used in this chapter, than seemed advisable here in view of the primary objectives of this investigation.

This statistic is the basic building block of every conceptualization of the Rasch fit statistic. It was stated here earlier that this statistic could be computed just as readily using single person interactions with each item; subgroup interactions with each item; or total sample interaction with each item. Therefore, in theory, any statistic which is derived using the standardized residual as its basis is also applicable to individual, subgroup, and total. The orientation favored in much of the MESA documentation is toward a subgroup statistic. However, extensive research has been done with single person and total group statistics as well. Unfortunately the distinctions between reference group size have not been meticulously held in this material with the result that the material is often confusing. This distinction is considered to be very important in this discussion because this investigation pertains entirely to the interpretation of the Rasch fit statistic developed in the analysis of subgroup scores. References to an individual score statistic or to a total group statistic will be deliberately avoided whenever possible. The between group statistic emphasized in this study is the most useful. The other two have disadvantages which outweigh their supposed value.

While the standardized residual is the basic building block, it has never been seriously interpreted as a standard score. Other statistical points of reference have been used over the years to explain and interpret this statistic. At first, it was presented as a  $X^2$  statistic. Research Memorandum 18 pointed out that "Wright and Panchapakesan (1969)

proposed a Pearson chi-square statistic for testing if the item calibrations are person-free" (Wright & Mead, 1975, p. 9). The version of BICAL used in this investigation employs an average of the standardized residuals for the score subgroups produced by the program. This average is identified as the fit mean square. The fit mean square is the fit statistic employed in this investigation to determine item fit to the Rasch model. However the decision to employ this statistic for that purpose was not an easy one. Research Memorandum 18 was very difficult to follow. It seemed to talk at once about a  $Z^2$  statistic; a  $X^2$  statistic; the fit mean square; and the F-statistic. The reference to person; to subgroup; or to total sample was obscure as well, so there were quite a number of confusing issues to be resolved at this point. Research Memorandum 23 (Wright & Mead, 1977c) did little to clarify the confusion. Research Memorandum 23.c (Wright & Mead, 1980) introduced the concept of interpreting the Rasch fit statistic as a Student's t.

Attempts were made to resolve questions on determining item fit over a period of several months. While greater understanding of the process was acquired, it was impossible to gain closure on this problem until the Summer of 1981 at a meeting with MESA staff in Chicago during a series of "Rasch Analysis Workshops" conducted between Monday, June 6 through Friday, June 10, 1981.

#### CHOOSING A FIT STATISTIC

The Workshop offered a thorough discussion of the subject. The seminar program was excellent, but the most productive session was a one-on-one discussion which the author had with Richard Smith, Director of Testing, Mercer County Community College in Trenton, New Jersey. Smith was on sabbatical leave from Mercer College to study Rasch

Measurement under Wright. At this point in time he was a MESA staff member and served as a major presenter during the Seminar. On Monday evening, June 15, Smith spent two hours in an open discussion of the item fit statistic and its interpretation. Ronald Mead, Assistant Director of MESA and co-author of the three MESA publications so often referred to in this study (i.e., Research Memoranda 18, 23, and 23.c), joined this discussion for the last half hour.

Smith explained that the statistic labelled "FIT MN. SQ" in the output table of the version of BICAL used in this study constituted the fit statistic. He indicated that it should be interpreted as if it were an F-statistic with one and five degrees of freedom. He explained that if Wright had stopped at merely summing the FIT Z-SQUARED values, the result could be interpreted as a chi-square with one and infinite degrees of freedom. However, Wright did not stop there. He determined that an average of the group (i.e., ability group) FIT Z-SQUARED values provided a statistic which approximated the sum of squares between groups, in analysis of variance. This fit statistic is called the "between fit mean square" because of the similarity in the way it is derived to the sum of squares between groups in analysis of variance. This statistic, which is labelled "FIT MN SQ" in the BICAL tables produced in this study, is interpreted like an F-statistic with one and five degrees of freedom.

Excessively difficult items identified in this analysis are re-scored in favor of every test taker who got them wrong in the first pass of the data files. Then, the learning objectives are re-scored. Comparisons between the average number of objectives passed before and

after the re-scoring process indicates the probable consequences of including items that are too difficult for each test.

Throughout this study, the between fit mean value, labelled FIT MN. SQ in the printout produced by the Rasch analysis program used, has been interpreted as an F-statistic with one and five degrees of freedom at  $\alpha = 0.05$ . The critical value of the F-statistic is 6.61 at the degrees of freedom indicated. The region of rejection is entirely in the upper tail of the F-distribution. Execution of the BICAL runs was quickly accomplished once the data acquisition, file formatting, and computer run problems had been solved. Once they were resolved, an entirely new orientation was needed to the work which yet had to be done.

The remaining procedural steps in this investigation had direct bearing on the research objectives. Briefly, three things were needed to conclude this study once the Rasch analysis runs were completed:

1. Identify all items which did not fit the Rasch model.
2. Re-score items judged to be too difficult for the purposes of the test in favor of those who missed such items.
3. Compare and interpret the reading objective pass rate before re-scoring to the objective pass rate after re-scoring.

The following section deals with the first two procedural objectives. The third objective is the substantive purpose of this research.

#### RE-SCORING ITEMS THAT ARE TOO DIFFICULT

The critical F-value in this study is 6.61, with one and five degrees of freedom. From a statistical point of view, two item classes might fall into the region of rejection defined by this critical value:

1) items that are "too easy" to conform to the Rasch model, and 2) items that are "too difficult" to conform to the Rasch model. Conceptually, both classes of item might be interpreted as unsuitable on the grounds that, whatever they measure, they do not measure the underlying, or latent, trait which the test is intended to measure. However, as discussed in previous sections of this chapter, easy items that are valid, within the context of a criterion referenced test, are always suitable. Therefore, only those items judged to be too difficult are of direct concern in this investigation. The procedures outlined in this section pertain to the identification of those items having levels of observed difficulty exceeding expected difficulty by an amount which would occur by chance only five times in one hundred. These items are re-scored in a two-phase procedure. The first phase involves changing all incorrect answers to correct answers on the grounds that the test taker should not be penalized by items that are too difficult to measure the underlying trait which the MEAP reading test is intended to measure. The second phase employed the output from the first phase to produce a series of t-tests which were designed to show whether or not a significant change had occurred in average objective performance as a result of the re-scoring process.

#### PHASE ONE

In the BICAL output the FIT ORDER table, showing FIT MN. SQ data in ascending order, provides all the information necessary for the re-scoring operation. To begin, all items having FIT MN. SQ values larger than 6.61 are identified. Those having negative difficulty estimates are more difficult than predicted by the model. And since they also have fit mean squares exceeding the critical value, these are the items

which will be re-scored on the assumption that they are too difficult to be appropriate in a MEAP reading test. The next step in this procedure involves preparation of an SPSS computer routine. This routine will concurrently generate a revised sample file and a t-statistic to evaluate changes in the number of passed learning objectives, before and after the re-scoring process, for each of the 14 samples used in this investigation. In fact, 14 similar but distinctly different SPSS control card sets were used.

The logical design of these SPSS routines was identical. Each one was to perform four identical functions:

1. Read the original sample.
2. Re-score items judged to be too difficult by giving credit for all wrong answers for those items.
3. Re-score learning objectives, based on the re-scored items.
4. Perform a series of t-tests comparing the average number of passed learning objectives before and after the re-scoring process.

The SPSS procedure WRITE CASES was used to create the revised sample file, including some new variables, and the SPSS procedure T-TEST was used to determine if a statistically significant change had occurred in the average number of objectives passed as a result of the scoring process.

However, while the basic logic and procedures of the 14 SPSS routines were identical, each had to be customized to process only the specific items judged to be too difficult for each test. The strategy employed to accomplish this objective entailed the design of two model routines; one for fourth grade tests and one for seventh grade tests.

These model routines incorporated all items being re-scored without regard to test year. Some items were too difficult in every test where they were used. Other items were too difficult in only one year. Every conceivable pattern in between these extremes occurred. So, of course, the model routines were not really suitable for any specific test. They were created to include every conceivable re-scoring situation. Since these SPSS control sets considered every re-scoring possibility, they were easier to test for coding errors than individual routines would have been. When the necessary testing was completed, it was a simple matter to build the individual routines for each of the fourteen tests from the two original control sets using text editing procedures. Once these complex models had been successfully tested, it was a relatively simple matter to copy specific portions - portions which applied only to a specific test and year. In reality, this way only two routines had to be designed from inception rather than fourteen. The economy of effort was considerable and the approach virtually eliminated any prospects of error since the differences could be confined to specific, and very limited, areas. These were fairly extensive routines. The fourth grade model routine contained 271 lines of code. The seventh grade model routine contained 239 lines of code. The fourth grade model appears in APPENDIX G. All 14 of the routines derived from these models had fewer lines of code than the models themselves. The difference was the result of using only references to items judged too difficult in a specific test. All of these routines involved eight functional steps:

1. Read the original sample-of-1,000 students. This entails reading 169 variables from the files original three card format. Each item is considered a variable in this count.

2. Create item counters for (only) those items which are judged to be too difficult. For example, for item QA1:

```
COMPUTE      CA1=0
```

3. Create learning objective re-scoring counters for (only) those objectives involving items which are judged to be too difficult. For example, for objective A:

```
COMPUTE.    SCOREQA=0
```

4. Create learning objective status counters for (only) those objectives involving items which are judged to be too difficult. For example, for objective A:

```
COMPUTE      OBSTAT=0
```

5. Count each time an individual item is re-scored from 0 to 1 and also count the total number of times any item is re-scored. Incorrect items, coded 0, are recoded 1, indicating credit is given for the item. If the item is already coded 1, the code is left unchanged and no record is made of the occurrence of this type of transaction. There are three lines of code for each item at this point in the routine. For example, for item QA1:

```
IF          (QA1 EQ 0) CA1=CA1+1
IF          (QA1 EQ 0) COUNT=COUNT+1
IF          (QA1 EQ 0) QA1=1
```

The first line of code counts the number of times item QA1 is recoded from 0 to 1. Similar coding is used for each item being recoded. The second line of code appears in every recode sequence so that a count can be taken of every occurrence when any item is recoded from 0 to 1. The third



line of code re-scores the item from 0 to 1. This sequence of instructions is set up for those items judged too difficult. Other items, not so judged, are not re-scored.

6. Add the item count to those learning objective scoring counters created for (only) those objectives involving items which are judged too difficult. Then test each counter to determine if four or more items, including re-scored items, have been passed. If so, the corresponding learning objective status counter is set to 1, indicating that the learning objective has been passed. Six lines of code are required at this point in the procedure. For example, for objective A:

```

COMPUTE      SCOREQA=SCOREQA+QA1
COMPUTE      SCOREQA=SCOREQA+QA2
COMPUTE      SCOREQA=SCOREQA+QA3
COMPUTE      SCOREQA=SCOREQA+QA4
COMPUTE      SCOREQA=SCOREQA+QA5
IF           (SCOREQA GE 4) OBSTAT=1

```

Again, these objectives are set up only for those objectives involving items judged to be too difficult. No other objectives are re-scored.

7. Add the objective count to the counter NEWPASS. The pattern of additions to this counter involves the addition of one learning objective status counter for every objective in the test, but they are a mixture of original and recoded status counters. If an objective did not involve an item judged to be too difficult, the original objective status counter was

used. These have numeric tags in the last two positions of their label. For example, the original objective status counter for the third learning objective was labelled OBSTAT03. But, if an objective did involve an item that was judged to be too difficult, the re-scored objective status counter was used. These have a letter in the last position of their label. For example, the re-scored objective status counter for the third objective would be OBSTATC.

Consequently, the contents of NEWPASS reflect the changed, or re-scored, objective status count.

8. Create a revised sample file, using the WRITE CASES procedure card, on logical unit 9. This entails writing the 169 original variables from the original input file plus new variables created specifically for recoding and tracking the re-scoring procedure for items judged to be too difficult, in a four card format.
9. Compare the mean of the original number of objectives passed, represented by the variable NUMPASS, to the mean of the re-scored number of objectives passed, which are represented by the variable NEWPASS. The T-TEST procedure card, calling for a pairwise comparison, is used.

A maximum of only 30 permanent disk pages was available with the computer account on which this sequence of routines was run. Consequently, temporary disk files and computer tape had to be employed as extensively in this phase of the analysis as they were in preceding runs. Two hours, twelve minutes, and 21 seconds were required to run this sequence of fourteen recode routines. It was necessary to execute

125 instructions, manually, at a terminal on line to the computer system. These instructions were written out in advance of the session to minimize the possibilities of execution error and to reduce the amount of time needed to complete the sequence. The log for the entire session is presented in APPENDIX H.

#### PHASE TWO

The t-test entailed in the last step, step 9, in the first phase of this procedure had inadvertently produced erroneous results because the original score data, by learning objective and items within objective, was not reformatted to comply with the input requirements of the t-tests. The following primary objectives of this first series of computer runs had been accomplished successfully, however:

1. Re-score items judged to be too difficult under the criteria established in this investigation.
2. Produce a new sample file for each of the 14 MEAP reading tests used which reflects the result of the re-scoring procedure.

A second series of computer runs was implemented to produce the series of t-tests intended to evaluate the results of the re-scoring procedure. The re-scored files produced in the first phase were used as input to two SPSS control sets designed for the purpose: one specifically for the fourth grade tests and the other for the seventh grade tests. Both control sets involved 308 instructions designed to re-tabulate the objective score data under a revised format compatible to the t-tests produced. The statistical output for each of the 14 runs was comprised of a set of descriptive statistics for the original and the recoded score data plus the product of seven t-tests. The logical

design of these routines was basically identical to the SPSS control sets used in the first phase of the item re-score procedure outlined above. But, since the unique aspects of the input data had already been accounted for in the first phase, an even more straightforward series of control statements, identical though somewhat longer for each grade, could be used at this point, because the input files now could be processed in an identical fashion. Each of these runs accomplished five purposes:

1. Count the re-scored items by learning objective.
2. Compute the number of objectives passed.
3. Store the recoded item count and objective pass/fail record in a format which made these data suitable as input to the statistical tests incorporated within each run.
4. Produce descriptive statistics and seven t-tests on item and objective data.
5. Produce an SPSS system file for future use.

The t-test results produced in the output of this second phase of the item re-scoring procedure were employed extensively in the analysis done in this study. One final series of computer runs was necessary, however, before this analysis could be given full attention. The statistical tests referred to briefly in this discussion were designed to answer specific questions about item and objective performance measured by individual MEAP reading tests given each year from 1973 through 1979. In addition to an interest in item and objective performance measures each of these years, this research addressed the question of whether or not there was a statistically significant change in overall performance between the tests which evidenced the greatest

rate of failure in learning objectives and the failure rate evidenced in the 1979 tests.

COMPARING 1973 FOURTH GRADE AND 1974 SEVENTH GRADE LEARNING  
OBJECTIVE PERFORMANCE TO 1979 LEARNING OBJECTIVE PERFORMANCE

Four additional  $\chi^2$ -tests were devised to compare performance on items and on learning objectives between the 1973 and 1979 tests; and also between the 1974 through 1978 (as a group) and 1979 tests. It was anticipated that performance in the 1973 tests for both grades would be lower than any other year. In absolute terms, the 1974 tests, for both fourth and seventh grades, showed minimum performance on items and learning objectives at first. However, when the items that were not carried forward from the 1973 test were eliminated from the 1973 test, it was fully expected that 1973 would demonstrate the lowest levels of performance. However, when the items and objectives that were not carried forward to subsequent years were eliminated from the 1973 test, it did so only in the case of the fourth grade students. The seventh grade students evidenced slightly better performance in 1973 than they did in 1974. Consequently, 1973 was the low performance year for fourth graders compared to the 1979 test, and 1974 was the low performance year for seventh graders compared to the 1979 test. There appears to be a steady increase in level of performance measured from these base years, for both fourth and seventh graders, each year until 1979. The bases of comparison are the items and objectives common to the tests each year.

The objective of the analysis at this point of the investigation was to determine if there was a statistically significant difference between the item and objective performance on the 1979 tests and the year the least success was measured.

There were 20 items, involving four objectives, dropped from the 1973 fourth grade test and 15 items, involving three objectives dropped from the 1973 seventh grade test. The procedure employed to make the 1973 tests comparable to later years for comparison purposes, was to drop items QC1 to QC5, QN1 to QN5, QO1 to QO5, and QW1 to QW5 from the 1973 fourth grade test; and to drop items QC1 to QC5, QO1 to QO5, and QT1 to QT5 from the 1973 seventh grade test. Each of these tests, then became directly comparable, in terms of content, with all later tests from 1974 to 1979. To accomplish this objective, the 1973 tests were re-scored, less the dropped items, according to procedures outlined in phase two of the re-scoring discussion above. An item score count, before and after re-scoring, was made during this step which would be directly comparable to the corresponding counts done for the years 1974 through 1979 accomplished in the preceding, phase two, procedure.

The next step was to create a single file for fourth and seventh graders, respectively, comprised of 1979 data concatenated to the low performance year data so that these files could be input simultaneously to an SPSS run which would compare differences in levels of performance between the two benchmark years for each grade. The  $X^2$ -tests performed during this phase evaluated changes between learning objective performance means between the years and within the years before and after items judged to be too difficult had been re-scored.

#### SUMMARY

The limitations and assumptions which have been imposed on this investigation by design or circumstances may be summarized as follows:

1. Fourteen random samples, prepared by the Division of Research and Assessment, the State Department of Education, Lansing,

Michigan, comprise the data used in the analysis. Each was an official Sample-of-5000 drawn from MEAP tests given over the seven year span from 1973 through 1979. The State of Michigan has provided access for research purposes to samples drawn from all students who take the Reading Test and the Mathematics test each year since the inception of the Michigan Educational Assessment program.

2. Exactly 1,000 students were drawn at random from each Sample-of-5,000. The entire Sample-of-5,000 for each of the grades for each of the seven years encompassed by this study is considered too large. Such a sample is likely to produce statistically significant results in connection with relatively small score differences simply because it is so large. In addition, because of size, the Sample-of-5,000 would be too costly to process by computer with the funds available in this investigation. Consequently each sample was reduced to 1,000 students. This number is compatible to the available data processing budget yet the appropriate degree of statistical power is retained. Larger samples, it was felt, would likely be wasteful of resources if not completely unnecessary.
3. The analysis conducted in this study pertains entirely to fourth grade and seventh grade Reading Test samples. Mathematics samples are not considered. Therefore, all inferences drawn from these analyses are confined to the students who have taken the Michigan Educational Assessment Program criterion-referenced reading test. Accordingly, no

inferences are drawn concerning any other test administered in Michigan, or elsewhere.

4. Individual student scores on test items constitute the basis for determining the effectiveness of MEAP reading tests as measures of reading achievement.
5. The percentage of Reading Test objectives, passed is employed to determine the proportion of students taking the test who are qualified for remedial reading instruction. Students scoring below 40% of the objectives should be considered for remedial education under guidelines established by the Michigan Department of Education. Each test has either 19, 20, or 23 objectives depending on which grade level and year it was administered. Every objective is measured by five test items. To pass an objective, the student must answer four of the five items correctly.
6. The cost of computer time is a major consideration in the conduct of this analysis. Even at the reduced sample size employed in the analysis, 14,000 individuals were considered, on aggregate, in this study. Computer processing was, therefore, limited to the analysis of item fit to the Rasch model, and to t-tests and chi-square tests of differences in the number and proportion of students passing less than 40% of the MEAP Reading Test objectives before and after items which do not fit the Rasch model are removed from the analysis.
7. The computer program used to perform analysis of fit to the Rasch model, BICAL, will not process either a perfect score



or a zero score. Such scores are eliminated from the analysis on the grounds that they contain no useful information respecting the measure of knowledge an individual has of the latent variable intended to be measured by the test. Persons using the program may specify even more restrictive maximum and minimum score levels for elimination from the analysis should they choose.

8. The decision was made in this analysis to eliminate scores equal to 20% of the total test score possible in an effort to eliminate from consideration those scores which could be the result of random guessing by the test taker. No adjustment was made to the upper score limit included in this analysis. Only the default limit of a perfect score, set by the BICAL program, was dropped from the analysis at the top end of the scoring range.
9. Under Rasch measurement theory:
  - a. an individual's test score is viewed as a sufficient statistic. That is, the test score constitutes an appropriate measure of the latent trait covered by the test.
  - b. item-difficulty and person-ability are viewed as the only two variables worth considering in test measurement situations. Therefore, other variables such as propensity for guessing and item discrimination, often major concerns in classical measurement theory, are considered doomed to fail and probably irrelevant.

- c. calibration of item-difficulty is independent of persons taking the item, and, conversely, measures of ability are independent of the items selected to measure that ability.
  - d. items which do not fit the Rasch model may yet be retained if it appears that some factor other than item-difficulty or person-ability was the cause. In the process of calibrating test items to determine degree of fit to the Rasch model, factors like guessing and discrimination which could cause unexplained score variance should be considered in the decision to keep or reject an item. The objective, always, is to build a pool of items which "fit" the model.
10. In this study, it has been assumed that:
- a. the MEAP Reading Test score alone will stand as the measure of person-ability. No factors other than person-ability and item-difficulty need be considered.
  - b. no MEAP test item should be rejected on the grounds that it is too easy, even though that item clearly does not fit the Rasch model. This or any other grounds for rejecting an easy item must fail in the case of any criterion referenced test because there is a presumption of content validity for such test items which negates rejecting them because they are easy. It is assumed in this study that all easy items belong in the MEAP Reading Test because they are presumed to have

content validity, no matter how easy these items may appear to be.

- c. every MEAP test item that is too difficult to fit the Rasch model should be rejected. It is assumed that items which are too difficult do not belong in a criterion referenced test. Therefore the presumption of content validity which saves easy items from rejection in criterion referenced tests is not applied to difficult items which do not fit the Rasch model. These items are assumed to be too difficult.
- d. students should be given credit for items that are too difficult. It was felt that for these particular tests that it is unreasonable to penalize students with items that are so difficult that they do not fit the Rasch model. It is assumed that students would get an item of appropriate difficulty for the test correct were such an item present in place of the one, it seems fair to say, did not belong in the test from the beginning.
- e. students should retain credit for MEAP test items that are easy. It was felt that easy items which measure test objectives in a criterion referenced test such as this, ought to be kept in the item pool despite their lack of fit to the model.

The assumption that students should have credit for difficult MEAP test items that do not fit the Rasch model has played a key role in the design of this study. Proceeding on this assumption, tests were re-scored in the analysis phase of this study to reflect credit for such

items. Then a comparison is drawn between the proportion of students passing less than 40% of the test objectives before this adjustment and after this adjustment. The next chapter presents the results and description of the data analysis performed in this study in the execution of this overall design.

## CHAPTER IV

### RESULTS - DISCUSSION OF RESEARCH QUESTIONS AND PRESENTATION OF ANALYSIS RESULTS

#### INTRODUCTION

##### THE RESEARCH QUESTIONS

Results of the statistical analysis performed in this investigation are presented in this chapter. Five research questions are explored in an attempt to evaluate the impact of MEAP Reading Test items which do not fit the Rasch model. Since performance on MEAP tests is measured in terms of passed learning objectives rather than passed items, item failure is interpreted in this study in terms of effect on the probability for passing the objective associated with that item. Each learning objective is measured by five items. The student is required to get four of the five right to receive credit for the objective. Since an optimum test item, in terms of Rasch measurement theory, should correspond to reasonable expectations of person-ability, there are two item-fit possibilities: one possibility is that an item is so difficult as to be outside the acceptable limit of a hard item; the second possibility is that an item is so easy as to be considered an unworthy challenge of person-ability. Therefore, while items could potentially be too easy as well as too hard in the normal course of item evaluation, this is not the case here. Only those items which are too difficult to fit the Rasch model receive special attention in this analysis. Students who get these items wrong receive credit for them in this analysis and then a comparison is made to determine the affect of this procedure on the proportion of fourth and seventh grade students in Michigan that would be eligible for remedial reading instruction. The

five research questions developed in this chapter had to be resolved in the order of precedence that is given to them in the discussion which follows.

#### THE STATISTICAL OR NULL HYPOTHESIS

In the process of seeking answers to these questions, four statistical, or null, hypotheses are presented. The data developed in the analysis has been organized in this chapter into thirteen tables. Four of these tables support hypothesized findings. The other nine are discussed in connection with unhypothesized findings or serve to provide descriptive statistics associated with the samples used in this investigation. All results reported in any of these tables pertain to an N of 1000, the number in each of the 14 samples employed in this investigation.

#### DESCRIPTIVE STATISTICAL ANALYSIS

Question 1: How many items in the MEAP Reading Test for the years 1973 through 1979 fit the Rasch model?

#### STATISTICAL ANALYSIS

Four tables were developed in the process of seeking an answer to this question (i.e., Table 1 through Table 4). Each table presents an aspect of item-difficulty or item easiness encountered. The first three tables in this group are summarized directly from the results printed at the conclusion of each BICAL run. Table 2 shows the number from 1000 fourth grade and 1000 seventh grade students from 1973 through 1979 at every score level from one item correct to a maximum, on the 1973 test only, of 115 items correct. Table 3 shows the proportion which these same students represent of the sample. Table 4 depicts the estimates of

item difficulty computed by BICAL in log-odds terms. Table 1 presents item fit in tabular form. Each item is unique. Whenever an entry appears in Table 1 (i.e., "NO" or "Y"), the same item was administered that year. A number of items were administered only in 1973. Items not used from 1974 through 1979 are identified by a dash (i.e., "-").

#### HYPOTHESIZED FINDINGS

No hypothesis was posed in connection with this research question.

#### UNHYPOTHESIZED FINDINGS

In this group of four tables, Table 1 is the one which has a direct bearing on the answer to the first research question. The entry "Y" identifies those items which fit the Rasch model. Included in this group are easy items which, in point of fact, do not fit the Rasch model but were retained under the presumption of content validity. The entry "NO" identifies those items which do not fit the Rasch model under the criteria established in this study. The criteria for establishing a fit statistic have been discussed at length in APPENDIX C of this discussion. Briefly, the determination has been made here to reject all difficult items having FIT MN SQ values produced in the BICAL analysis of 6.61 or larger. The value 6.61 has been interpreted as if it were an F-statistic with one and five degrees of freedom at an alpha level of 0.05.

TABLE 1

DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL

TOTAL ITEMS	FOURTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO							TOTAL ITEMS	SEVENTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
	115	95	95	95	95	95	95		115	100	100	100	100	100	100
ITEM LABEL								ITEM LABEL							
QA1 .	NO	NO	NO	NO	NO	Y	Y	QA1 .	Y	Y	Y	Y	Y	Y	Y
QA2 .	Y	Y	Y	Y	Y	Y	Y	QA2 .	Y	Y	Y	Y	Y	Y	Y
QA3 .	Y	Y	Y	Y	Y	Y	Y	QA3 .	Y	Y	Y	Y	Y	Y	Y
QA4 .	NO	NO	NO	Y	Y	Y	Y	QA4 .	Y	Y	Y	Y	Y	Y	Y
QA5 .	Y	Y	Y	Y	Y	Y	Y	QA5 .	Y	Y	Y	Y	Y	Y	Y
QB1 .	Y	Y	Y	Y	Y	Y	Y	QB1 .	Y	Y	Y	Y	Y	Y	Y
QB2 .	NO	Y	Y	Y	Y	Y	Y	QB2 .	Y	Y	NO	Y	Y	Y	Y
QB3 .	NO	NO	NO	NO	Y	Y	Y	QB3 .	NO	NO	NO	NO	NO	Y	Y
QB4 .	NO	Y	Y	Y	Y	Y	Y	QB4 .	Y	Y	Y	Y	Y	Y	Y
QB5 .	Y	Y	Y	Y	Y	Y	Y	QB5 .	Y	Y	Y	Y	Y	Y	Y
QC1 .	Y	-	-	-	-	-	-	QC1 .	Y	-	-	-	-	-	-
QC2 .	Y	-	-	-	-	-	-	QC2 .	Y	-	-	-	-	-	-
QC3 .	Y	-	-	-	-	-	-	QC3 .	Y	-	-	-	-	-	-
QC4 .	Y	-	-	-	-	-	-	QC4 .	Y	-	-	-	-	-	-
QC5 .	Y	-	-	-	-	-	-	QC5 .	Y	-	-	-	-	-	-
QD1 .	Y	Y	Y	Y	Y	Y	Y	QD1 .	Y	-	-	-	-	-	-
QD2 .	Y	Y	Y	Y	Y	Y	Y	QD2 .	Y	-	-	-	-	-	-



TABLE 1 (continued)

DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL

TOTAL ITEMS	FOURTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO							TOTAL ITEMS	SEVENTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
	115	95	95	95	95	95	95		115	100	100	100	100	100	100
QD3 .	Y	Y	Y	Y	Y	Y	Y	QD3 .	Y	-	-	-	-	-	-
QD4 .	Y	Y	Y	Y	Y	Y	Y	QD4 .	Y	-	-	-	-	-	-
QD5 .	Y	Y	Y	Y	Y	Y	Y	QD5 .	Y	-	-	-	-	-	-
QE1 .	Y	Y	Y	Y	Y	Y	Y	QE1 .	Y	Y	Y	Y	Y	Y	Y
QE2 .	NO	Y	NO	NO	NO	NO	NO	QE2 .	Y	Y	Y	Y	Y	Y	Y
QE3 .	Y	Y	Y	Y	Y	Y	Y	QE3 .	Y	Y	Y	Y	Y	Y	Y
QE4 .	Y	Y	Y	Y	Y	Y	Y	QE4 .	Y	Y	Y	Y	Y	Y	Y
QE5 .	Y	Y	Y	Y	Y	Y	Y	QE5 .	Y	Y	Y	Y	Y	Y	Y
QF1 .	Y	NO	Y	Y	NO	NO	NO	QF1 .	Y	NO	Y	Y	Y	Y	Y
QF2 .	Y	Y	Y	Y	Y	Y	Y	QF2 .	Y	Y	Y	Y	Y	NO	Y
QF3 .	Y	Y	Y	Y	Y	Y	Y	QF3 .	Y	Y	Y	Y	Y	Y	Y
QF4 .	Y	Y	Y	NO	Y	Y	NO	QF4 .	NO	NO	NO	Y	Y	Y	Y
QF5 .	NO	Y	Y	Y	Y	Y	Y	QF5 .	Y	Y	Y	Y	Y	Y	Y
QG1 .	Y	Y	Y	Y	Y	Y	Y	QG1 .	Y	Y	Y	Y	Y	Y	Y
QG2 .	Y	Y	Y	Y	Y	Y	Y	QG2 .	Y	Y	Y	Y	Y	Y	Y
QG3 .	NO	NO	Y	NO	Y	Y	Y	QG3 .	Y	Y	Y	Y	Y	Y	Y
QG4 .	Y	Y	Y	Y	Y	Y	Y	QG4 .	Y	NO	Y	Y	Y	Y	Y
QG5 .	Y	Y	Y	Y	Y	Y	Y	QG5 .	Y	Y	Y	Y	Y	Y	Y
QH1 .	Y	Y	Y	Y	Y	Y	Y	QH1 .	Y	NO	Y	Y	Y	Y	Y

TABLE 1 (continued)

DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL

TOTAL ITEMS	FOURTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO							TOTAL ITEMS	SEVENTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
	115	95	95	95	95	95	95		115	100	100	100	100	100	100
QH2	NO	NO	NO	Y	Y	Y	Y	QH2	NO	NO	NO	Y	Y	Y	Y
QH3	Y	Y	Y	Y	Y	Y	Y	QH3	Y	Y	NO	Y	Y	Y	Y
QH4	NO	Y	Y	Y	Y	Y	NO	Y	Y	Y	Y	Y	Y	Y	Y
QH5	Y	Y	Y	Y	Y	Y	Y	QH5	Y	NO	NO	Y	Y	Y	Y
QI1	Y	Y	Y	Y	Y	Y	Y	QI1	Y	Y	Y	Y	Y	Y	Y
QI2	Y	Y	Y	Y	Y	Y	Y	QI2	Y	Y	Y	Y	Y	Y	Y
QI3	Y	Y	Y	Y	Y	Y	Y	QI3	Y	Y	Y	Y	Y	Y	Y
QI4	NO	NO	Y	Y	Y	Y	Y	QI4	Y	Y	Y	Y	Y	Y	Y
QI5	Y	Y	Y	Y	Y	Y	Y	QI5	Y	Y	Y	Y	Y	Y	Y
QJ1	Y	Y	Y	Y	Y	Y	Y	QJ1	Y	Y	Y	Y	Y	Y	Y
QJ2	Y	Y	Y	Y	Y	Y	Y	QJ2	Y	Y	Y	Y	Y	Y	Y
QJ3	Y	Y	Y	Y	Y	Y	Y	QJ3	Y	Y	Y	Y	Y	Y	Y
QJ4	Y	Y	Y	Y	Y	Y	Y	QJ4	Y	Y	Y	Y	Y	Y	Y
QJ5	Y	Y	Y	Y	Y	Y	Y	QJ5	Y	Y	Y	Y	Y	Y	Y
QK1	Y	Y	Y	Y	Y	Y	Y	QK1	Y	Y	Y	Y	Y	Y	Y
QK2	Y	Y	Y	Y	Y	Y	Y	QK2	Y	Y	Y	Y	Y	Y	Y
QK3	Y	Y	Y	Y	Y	Y	Y	QK3	Y	Y	Y	Y	Y	Y	Y
QK4	NO	Y	Y	Y	Y	Y	Y	QK4	Y	Y	Y	Y	Y	Y	Y
QK5	Y	Y	Y	Y	Y	Y	Y	QK5	Y	Y	Y	Y	Y	Y	Y

TABLE 1 (continued)

DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL

TOTAL ITEMS	FOURTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO							TOTAL ITEMS	SEVENTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
	115	95	95	95	95	95	95		115	100	100	100	100	100	100
QL1 .	Y	Y	Y	Y	Y	Y	Y	QL1 .	Y	Y	Y	Y	Y	Y	Y
QL2 .	Y	Y	Y	Y	Y	Y	Y	QL2 .	Y	Y	Y	NO	NO	Y	Y
QL3 .	Y	Y	Y	Y	Y	Y	Y	QL3 .	Y	NO	Y	Y	Y	Y	Y
QL4 .	Y	Y	Y	Y	Y	Y	Y	QL4 .	Y	Y	NO	Y	Y	Y	Y
QL5 .	Y	Y	Y	Y	Y	Y	Y	QL5 .	Y	Y	Y	Y	Y	Y	Y
QM1 .	Y	Y	Y	Y	Y	Y	Y	QM1 .	Y	Y	Y	Y	Y	Y	Y
QM2 .	Y	Y	Y	Y	Y	Y	Y	QM2 .	Y	Y	Y	Y	Y	Y	Y
QM3 .	Y	Y	Y	Y	Y	Y	Y	QM3 .	Y	Y	Y	Y	Y	Y	Y
QM4 .	Y	Y	Y	Y	Y	Y	Y	QM4 .	Y	Y	Y	Y	Y	Y	Y
QM5 .	Y	Y	Y	Y	Y	Y	Y	QM5 .	Y	Y	Y	Y	Y	Y	Y
QN1 .	Y	-	-	-	-	-	-	QN1 .	NO	Y	Y	Y	Y	Y	Y
QN2 .	Y	-	-	-	-	-	-	QN2 .	Y	Y	Y	Y	Y	Y	Y
QN3 .	Y	-	-	-	-	-	-	QN3 .	Y	Y	Y	Y	Y	Y	Y
QN4 .	Y	-	-	-	-	-	-	QN4 .	Y	Y	Y	Y	Y	Y	Y
QN5 .	Y	-	-	-	-	-	-	QN5 .	Y	Y	Y	Y	Y	Y	Y
QO1 .	Y	-	-	-	-	-	-	QO1 .	Y	Y	Y	Y	Y	Y	Y
QO2 .	Y	-	-	-	-	-	-	QO2 .	Y	Y	Y	Y	Y	Y	Y
QO3 .	Y	-	-	-	-	-	-	QO3 .	Y	Y	Y	Y	Y	Y	Y
QO4 .	Y	-	-	-	-	-	-	QO4 .	Y	Y	Y	Y	Y	Y	Y

TABLE 1 (continued)

DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL

TOTAL ITEMS	FOURTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO							TOTAL ITEMS	SEVENTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
	115	95	95	95	95	95	95		115	100	100	100	100	100	100
Q05	Y	-	-	-	-	-	-	Q05	Y	Y	Y	Y	Y	Y	Y
QP1	NO	Y	Y	Y	Y	Y	Y	QP1	Y	Y	Y	Y	Y	Y	Y
QP2	Y	Y	Y	Y	Y	Y	Y	QP2	NO	NO	NO	Y	Y	Y	Y
QP3	Y	Y	Y	Y	Y	Y	Y	QP3	Y	Y	Y	Y	Y	Y	Y
QP4	Y	Y	Y	Y	Y	Y	Y	QP4	Y	Y	Y	Y	Y	Y	Y
QP5	Y	Y	Y	Y	Y	Y	Y	QP5	Y	Y	Y	Y	Y	Y	Y
QQ1	Y	Y	Y	NO	NO	Y	Y	QQ1	Y	Y	Y	Y	Y	Y	Y
QQ2	Y	Y	Y	Y	Y	Y	Y	QQ2	Y	Y	Y	Y	Y	Y	Y
QQ3	Y	Y	Y	Y	Y	Y	Y	QQ3	Y	Y	Y	Y	Y	Y	Y
QQ4	Y	Y	Y	Y	Y	Y	Y	QQ4	Y	Y	Y	Y	Y	Y	Y
QQ5	Y	Y	Y	Y	Y	Y	Y	QQ5	Y	Y	Y	Y	Y	Y	Y
QR1	Y	Y	Y	Y	Y	Y	Y	QR1	Y	Y	Y	Y	Y	Y	Y
QR2	NO	NO	Y	Y	Y	Y	Y	QR2	Y	Y	Y	Y	Y	Y	Y
QR3	Y	Y	Y	Y	Y	Y	Y	QR3	Y	Y	Y	Y	Y	Y	Y
QR4	Y	Y	Y	Y	Y	Y	Y	QR4	Y	Y	Y	Y	Y	Y	Y
QR5	Y	Y	Y	Y	Y	Y	Y	QR5	Y	Y	Y	Y	Y	Y	Y
QS1	Y	Y	Y	Y	Y	Y	Y	QS1	Y	Y	Y	Y	Y	Y	Y
QS2	Y	Y	Y	Y	Y	Y	Y	QS2	NO	NO	NO	Y	Y	Y	Y
QS3	Y	Y	Y	Y	Y	Y	Y	QS3	Y	Y	Y	Y	Y	Y	Y

TABLE 1 (continued)

DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL

TOTAL ITEMS	FOURTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO							TOTAL ITEMS	SEVENTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
	115	95	95	95	95	95	95		115	100	100	100	100	100	100
QS4 .	Y	Y	Y	Y	Y	Y	Y	QS4 .	Y	Y	Y	Y	Y	Y	Y
QS5 .	Y	Y	Y	Y	Y	Y	Y	QS5 .	Y	Y	Y	Y	Y	Y	Y
QT1 .	Y	Y	Y	NO	Y	Y	Y	QT1 .	Y	-	-	-	-	-	-
QT2 .	Y	Y	Y	Y	Y	Y	Y	QT2 .	Y	-	-	-	-	-	-
QT3 .	NO	NO	NO	NO	Y	Y	Y	QT3 .	Y	-	-	-	-	-	-
QT4 .	Y	Y	Y	Y	Y	Y	Y	QT4 .	Y	-	-	-	-	-	-
QT5 .	Y	Y	Y	Y	Y	Y	Y	QT5 .	Y	-	-	-	-	-	-
QU1 .	Y	Y	Y	Y	Y	Y	Y	QU1 .	Y	Y	Y	Y	Y	Y	Y
QU2 .	Y	Y	Y	Y	Y	Y	Y	QU2 .	Y	Y	Y	Y	Y	Y	Y
QU3 .	Y	Y	Y	Y	Y	Y	Y	QU3 .	Y	Y	Y	Y	Y	Y	Y
QU4 .	Y	Y	Y	Y	Y	Y	Y	QU4 .	Y	Y	Y	Y	Y	Y	Y
QU5 .	Y	Y	Y	Y	Y	Y	Y	QU5 .	Y	Y	Y	Y	Y	Y	Y
QV1 .	Y	Y	Y	Y	Y	Y	Y	QV1 .	Y	Y	Y	Y	Y	Y	Y
QV2 .	Y	Y	Y	Y	Y	Y	Y	QV2 .	Y	Y	Y	Y	Y	Y	Y
QV3 .	Y	Y	Y	Y	Y	Y	Y	QV3 .	Y	Y	NO	NO	NO	NO	NO
QV4 .	Y	Y	Y	Y	Y	Y	Y	QV4 .	NO	Y	Y	Y	Y	Y	Y
QV5 .	Y	Y	Y	Y	Y	Y	Y	QV5 .	Y	Y	Y	Y	Y	Y	Y
QW1 .	Y	-	-	-	-	-	-	QW1 .	Y	Y	Y	Y	Y	Y	Y
QW2 .	Y	-	-	-	-	-	-	QW2 .	Y	NO	NO	NO	Y	Y	Y

TABLE 1 (continued)

DESIGNATION OF ITEM FIT WHERE "YES" IDENTIFIES ITEMS THAT EITHER FIT THE RASCH MODEL OR THOSE VERY EASY ITEMS THAT DO NOT FIT THE RASCH MODEL WHICH ARE RETAINED UNDER CRITERIA ESTABLISHED IN THIS STUDY AND WHERE "NO" IDENTIFIES VERY DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL

		FOURTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO									SEVENTH GRADE ITEMS WHICH FIT THE RASCH MODEL: YES (i.e., "Y") OR NO						
		1973	1974	1975	1976	1977	1978	1979			1973	1974	1975	1976	1977	1978	1979
TOTAL	ITEMS	115	95	95	95	95	95	95	TOTAL	ITEMS	115	100	100	100	100	100	100
QW3	.	NO	-	-	-	-	-	-	QW3	.	Y	Y	Y	Y	Y	Y	Y
QW4	.	NO	-	-	-	-	-	-	QW4	.	Y	Y	Y	Y	Y	Y	Y
QW5	.	Y	-	-	-	-	-	-	QW5	.	Y	Y	Y	Y	Y	Y	Y
		NUMBER OF VERY DIFFICULT FOURTH GRADE ITEMS THAT DO NOT FIT THE RASCH MODEL AND THE PERCENTAGE THESE ITEMS REPRESENT OF THE TOTAL NUMBER OF ITEMS									NUMBER OF VERY DIFFICULT SEVENTH GRADE ITEMS THAT DO NOT FIT THE RASCH MODEL AND THE PERCENTAGE THESE ITEMS REPRESENT OF THE TOTAL NUMBER OF ITEMS						
		1973	1974	1975	1976	1977	1978	1979			1973	1974	1975	1976	1977	1978	1979
TOTAL		17	9	6	8	4	3	3	TOTAL		7	11	11	4	3	2	1
% ALL	ITEMS	14.8	9.5	6.3	8.4	4.2	3.2	3.2	% ALL	ITEMS	6.1	11.0	11.0	4.0	3.0	2.0	1.0

Table 1 reveals an overall decline in the number of difficult items that do not fit the Rasch model during the seven year period of this study from 1973 through 1979. This is generally true for the fourth grade, though a temporary jump occurred in 1976. However, the downward trend in fourth grade results resumed in 1977. In 1973, the seventh grade results showed a lower number of items that did not fit the Rasch model than the next two succeeding years. A decline in seventh grade results began in 1976. The downward trend in fourth grade results, which resumed in 1977, and the seventh grade results beginning with 1976, dropped below any of the preceding years in each grade. Thus both grades demonstrate a successive, though small, decline in the incidence of difficult items from 1977 through the 1979 tests. The incidence of these very difficult items ranged from a high of 15% in 1973 to a low of 3% in 1978 and 1979, for the fourth graders, and from a high of 11% in 1974 and 1975 to a low of 1% in 1979, for the seventh graders. The total number of difficult items which did not fit the criteria established in this investigation from 1973 through 1979 were 17, 9, 6, 8, 4, 3, and 3, respectively, for the fourth grade and 7, 11, 11, 4, 3, 2, and 1, respectively, for the seventh grade.

Table 2 suggests that no one score group has very many of the students from the 1000 sampled, but a perceptible shift appears to occur in the number of students at the higher scoring levels in both fourth and seventh grades as time passes. This shift toward higher scores with the passage of time is more apparent in Table 2A which presents the quartile scores of the fourth grade and the seventh grade students. The predominance of higher scores among seventh grade students in comparison to fourth graders, that is revealed in this table, may be due to the

fact that there are five more questions in the seventh grade test than were used in the fourth grade test from 1974 through 1979. It is interesting to note that the first quartile scores in the fourth grade are higher for 1975 and 1976 than the seventh grade counterparts despite the use of fewer questions in the fourth grade test. There is probably very little value in a direct comparison of fourth grade and seventh grade scores as there may be little real difference between them at any level. However, there is a pronounced increase in first quartile scores for both grades between 1974 and 1979, the years tests within each grade had the same number of questions, that is interesting. These increases exceed 20 points in both cases. The second quartile scores also increased, but by less than half as much as the first quartile scores. Third quartile scores increased by approximately four points and two points, respectively, for the fourth and seventh graders from 1974 through 1979.



TABLE 2

NUMBER OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH  
SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

SCORE LEVEL	FOURTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
1	0	1	1	0	0	0	1	1	0	0	0	1	0	0	
2	1	0	0	2	1	0	0	2	0	1	0	0	0	0	
3	0	0	1	2	1	0	0	3	0	0	1	0	0	0	
4	1	0	0	0	0	0	0	4	0	0	0	0	0	0	
5	1	0	0	1	1	0	0	5	0	0	0	0	0	0	
6	0	0	0	0	2	0	0	6	0	2	0	0	0	0	
7	1	1	0	0	0	0	0	7	0	1	0	0	1	0	
8	0	0	0	1	0	0	0	8	0	0	0	0	0	0	
9	0	0	1	0	1	0	0	9	0	0	0	1	0	0	
10	0	0	0	0	0	0	0	10	0	0	0	0	0	0	
11	0	0	0	0	0	0	1	11	0	0	0	0	0	0	
12	0	0	0	0	0	0	0	12	0	0	0	0	0	0	
13	1	1	0	0	0	0	0	13	2	0	1	0	0	0	
14	0	2	0	0	1	0	1	14	2	0	2	0	0	0	
15	0	2	1	0	0	1	0	15	2	0	0	2	0	0	
16	0	0	0	0	2	0	0	16	1	1	0	1	0	0	
17	0	3	4	4	0	0	0	17	1	1	1	1	0	0	
18	1	0	3	0	1	0	1	18	0	2	1	1	0	0	
19	2	3	2	5	1	0	1	19	0	5	0	2	0	1	
20	0	4	2	4	2	1	0	20	0	4	8	6	3	0	
21	1	6	3	5	6	2	2	21	0	4	6	4	1	1	
22	2	3	6	7	2	6	0	22	0	4	1	2	0	1	
23	4	10	4	4	4	5	5	23	3	2	5	5	2	1	

TABLE 2 (continued)

NUMBER OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

SCORE LEVEL	FOURTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
24	6	12	6	11	9	5	4	24	3	8	8	4	3	3	3
25	4	3	10	12	4	4	1	25	2	11	7	7	2	3	1
26	7	11	9	13	4	4	2	26	6	8	5	7	5	1	0
27	8	10	8	11	5	3	3	27	10	5	12	5	2	1	1
28	10	10	6	10	4	3	5	28	3	7	7	8	0	0	1
29	9	15	13	5	1	2	6	29	12	10	10	8	1	4	1
30	17	7	7	5	5	5	5	30	6	8	8	4	1	2	1
31	7	13	9	2	4	6	0	31	5	9	10	1	3	2	4
32	15	6	8	7	6	6	6	32	6	11	8	7	3	6	2
33	7	11	4	9	13	9	7	33	7	10	9	4	2	2	5
34	10	9	5	9	10	5	3	34	9	7	4	6	3	0	2
35	11	8	9	6	4	7	2	35	8	8	7	7	3	3	3
36	12	8	8	7	4	9	0	36	7	9	7	7	4	1	3
37	12	10	8	5	7	3	0	37	7	8	3	9	2	4	1
38	7	6	6	8	2	6	4	38	7	11	11	3	3	4	4
39	9	9	4	7	3	3	2	39	2	11	10	9	3	1	3
40	11	9	3	1	4	1	2	40	10	3	2	4	3	3	2
41	4	6	2	5	1	2	3	41	5	9	5	8	4	4	1
42	8	3	7	5	2	1	2	42	2	5	6	6	8	3	4
43	4	3	4	1	2	4	5	43	13	7	8	9	5	1	2
44	8	7	8	2	4	6	3	44	3	4	3	3	2	2	3
45	9	6	3	3	4	1	1	45	7	4	7	7	3	2	2
46	8	3	8	2	4	4	4	46	5	6	6	8	5	2	5
47	2	7	6	7	4	3	5	47	6	2	5	6	1	6	4

TABLE 2 (continued)

NUMBER OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH  
SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

SCORE LEVEL	FOURTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
48	5	6	6	3	4	4	2	48	7	6	10	4	5	3	4
49	8	8	6	4	3	7	2	49	3	9	8	4	6	2	2
50	6	5	3	3	6	7	6	50	3	8	7	5	5	3	2
51	6	6	5	4	4	4	5	51	4	4	6	12	5	6	3
52	5	7	2	6	5	4	8	52	4	2	6	6	5	5	4
53	4	5	4	4	4	5	4	53	3	7	1	8	7	4	3
54	6	4	5	6	6	5	8	54	3	2	3	8	8	5	4
55	5	10	6	5	9	1	6	55	4	7	3	8	7	8	3
56	7	7	6	5	3	3	4	56	4	3	4	9	5	5	6
57	7	7	5	3	6	7	1	57	5	8	7	8	9	5	2
58	3	6	3	11	7	7	3	58	8	6	7	10	4	3	4
59	8	3	2	2	7	7	4	59	5	12	2	6	4	8	4
60	6	5	7	4	14	4	6	60	6	10	12	5	5	4	6
61	7	6	12	7	7	7	13	61	4	6	4	10	4	10	6
62	13	6	5	5	2	8	2	62	7	11	10	3	6	6	4
63	6	9	7	6	4	10	7	63	4	12	1	8	6	5	4
64	8	9	5	10	10	8	4	64	7	6	5	5	7	4	7
65	3	12	9	7	6	5	5	65	6	9	10	7	3	9	7
66	7	16	7	7	13	12	3	66	8	9	6	8	6	8	5
67	7	13	11	8	12	7	12	67	5	10	3	6	7	8	9
68	6	15	9	11	8	6	9	68	7	11	8	6	9	5	5
69	8	7	8	7	4	5	13	69	7	7	9	3	5	10	9
70	7	15	7	10	7	10	5	70	9	10	8	4	5	7	9
71	9	14	13	10	11	10	9	71	4	6	4	5	5	7	7

TABLE 2 (continued)

NUMBER OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH  
SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

SCORE LEVEL	FOURTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
72	6	12	15	13	10	12	12	72	3	7	5	11	7	14	13
73	9	15	17	9	13	15	11	73	3	6	11	9	10	10	10
74	7	19	14	19	20	8	10	74	8	12	15	12	9	13	12
75	10	22	20	15	16	5	13	75	5	12	7	12	12	6	10
76	10	11	10	22	9	15	18	76	6	12	13	10	12	16	16
77	7	16	20	18	17	18	17	77	7	3	7	13	18	8	7
78	5	25	22	20	12	17	14	78	11	9	10	17	13	15	10
79	5	15	27	20	24	13	20	79	7	4	16	8	7	11	16
80	10	24	16	15	22	17	18	80	2	15	11	9	17	11	20
81	9	23	24	22	21	20	16	81	10	17	9	16	15	22	13
82	12	28	32	18	26	26	30	82	12	18	12	18	14	10	15
83	12	42	29	29	32	29	26	83	11	13	15	11	16	22	13
84	11	31	25	23	33	36	33	84	3	12	18	13	19	22	18
85	9	26	35	26	27	38	36	85	7	19	13	15	13	17	27
86	13	34	37	30	38	50	43	86	10	19	23	19	25	15	16
87	11	41	52	45	34	41	43	87	3	27	18	17	28	27	28
88	7	34	42	46	51	65	53	88	8	19	28	24	28	27	22
89	8	31	46	48	48	37	60	89	16	24	29	27	26	19	30
90	8	32	49	49	62	52	60	90	13	22	31	31	28	33	29
91	10	34	30	39	50	57	68	91	13	26	37	38	46	39	44
92	16	28	39	49	53	50	50	92	22	29	38	40	44	42	47
93	15	15	28	48	37	56	48	93	14	46	35	37	51	55	45
94	18	10	18	21	37	34	42	94	22	51	44	48	57	61	62
95	15	1	11	18	10	18	21	95	13	45	41	50	53	54	68

TABLE 2 (continued)

NUMBER OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH  
SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

SCORE LEVEL	FOURTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE NUMBER OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
96	20	..	..	..	..	..	..	96	13	44	49	48	60	67	65
97	13	..	..	..	..	..	..	97	11	30	46	48	52	53	66
98	21	..	..	..	..	..	..	98	23	26	39	32	53	51	45
99	25	..	..	..	..	..	..	99	14	30	30	15	35	40	43
100	24	..	..	..	..	..	..	100	18	7	6	7	13	16	8
101	23	..	..	..	..	..	..	101	8	..	..	..	..	..	..
102	28	..	..	..	..	..	..	102	27	..	..	..	..	..	..
103	19	..	..	..	..	..	..	103	28	..	..	..	..	..	..
104	27	..	..	..	..	..	..	104	31	..	..	..	..	..	..
105	26	..	..	..	..	..	..	105	20	..	..	..	..	..	..
106	21	..	..	..	..	..	..	106	29	..	..	..	..	..	..
107	22	..	..	..	..	..	..	107	40	..	..	..	..	..	..
108	20	..	..	..	..	..	..	108	36	..	..	..	..	..	..
109	22	..	..	..	..	..	..	109	..	..	..	..	..	..	..
110	14	..	..	..	..	..	..	110	27	..	..	..	..	..	..
111	30	..	..	..	..	..	..	111	48	..	..	..	..	..	..
112	10	..	..	..	..	..	..	112	32	..	..	..	..	..	..
113	5	..	..	..	..	..	..	113	23	..	..	..	..	..	..
114	7	..	..	..	..	..	..	114	13	..	..	..	..	..	..
115	2	..	..	..	..	..	..	115	4	..	..	..	..	..	..

**TABLE 2A**  
**SCORES OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH**  
**QUARTILE LEVEL**

Q LEVEL	FOURTH GRADE STUDENT SCORES AT EACH QUARTILE LEVEL							Q LEVEL	SEVENTH GRADE STUDENT SCORES AT EACH QUARTILE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
Q1	52.45	50.18	59.92	60.00	65.22	69.45	72.34	Q1	64.50	54.31	57.23	57.03	74.24	74.91	79.96
Q2	84.93	75.25	79.43	80.23	82.16	83.47	84.40	Q2	93.50	81.34	84.92	84.07	88.37	89.08	89.39
Q3	101.01	85.10	87.19	88.23	88.49	89.11	89.34	Q3	105.15	92.31	93.06	92.95	94.17	94.30	94.32
N	999	998	1000	998	999	999	990	N	960	989	984	986	998	1000	999

Table 3 presents basically the same information found in Table 2 but in terms of proportions of the total sample within each score group. The most striking impression gained from this table is the information that no score category contains more than 7% of the total sample and that most, by far, contain 3% or less.

PROPORTION OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH QUARTILE LEVEL WHO CORRECTLY ANSWER EACH ITEM

TABLE 3

SCORE LEVEL	FOURTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL					SCORE LEVEL	SEVENTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL							
	1973	1974	1975	1976	1977		1978	1979	1973	1974	1975	1976	1977	1978
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00



PROPORTION OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

TABLE 3 (continued)

SCORE LEVEL	FOURTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
25	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00
26	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00
27	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00
28	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00
29	0.01	0.02	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00
30	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
31	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00
32	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.00
33	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.00
34	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
35	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
36	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
37	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
38	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
39	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
40	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
41	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
42	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00
44	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
45	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
46	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.00
47	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00
48	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00



TABLE 3 (continued)  
 PROPORTION OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

SCORE LEVEL	FOURTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
73	0.01	0.02	0.02	0.01	0.01	0.02	0.01	73	0.00	0.01	0.01	0.01	0.01	0.01	0.01
74	0.01	0.02	0.01	0.02	0.02	0.01	0.01	74	0.01	0.01	0.02	0.01	0.01	0.01	0.01
75	0.01	0.02	0.02	0.02	0.02	0.01	0.01	75	0.01	0.01	0.01	0.01	0.01	0.01	0.01
76	0.01	0.01	0.01	0.02	0.01	0.02	0.02	76	0.01	0.01	0.01	0.01	0.01	0.02	0.02
77	0.01	0.02	0.02	0.02	0.02	0.02	0.02	77	0.01	0.00	0.01	0.01	0.02	0.01	0.01
78	0.01	0.03	0.02	0.02	0.01	0.02	0.01	78	0.01	0.01	0.01	0.02	0.01	0.02	0.01
79	0.01	0.02	0.03	0.02	0.02	0.01	0.02	79	0.01	0.00	0.02	0.01	0.01	0.01	0.02
80	0.01	0.02	0.02	0.02	0.02	0.02	0.02	80	0.00	0.02	0.01	0.01	0.02	0.01	0.02
81	0.01	0.02	0.02	0.02	0.02	0.02	0.02	81	0.01	0.02	0.01	0.02	0.02	0.02	0.01
82	0.01	0.03	0.03	0.02	0.03	0.03	0.03	82	0.01	0.02	0.01	0.02	0.01	0.01	0.02
83	0.01	0.04	0.03	0.03	0.03	0.03	0.03	83	0.01	0.01	0.02	0.01	0.02	0.02	0.01
84	0.01	0.03	0.03	0.02	0.03	0.04	0.03	84	0.00	0.01	0.02	0.01	0.02	0.02	0.02
85	0.01	0.03	0.04	0.03	0.03	0.04	0.04	85	0.01	0.02	0.01	0.02	0.01	0.02	0.03
86	0.01	0.03	0.04	0.03	0.04	0.05	0.04	86	0.01	0.02	0.02	0.02	0.03	0.02	0.02
87	0.01	0.04	0.05	0.05	0.03	0.04	0.04	87	0.00	0.03	0.02	0.02	0.03	0.03	0.03
88	0.01	0.03	0.04	0.05	0.05	0.07	0.05	88	0.01	0.02	0.03	0.02	0.03	0.03	0.02
89	0.01	0.03	0.05	0.05	0.05	0.04	0.06	89	0.02	0.02	0.03	0.03	0.03	0.02	0.03
90	0.01	0.03	0.05	0.05	0.06	0.05	0.06	90	0.01	0.02	0.03	0.03	0.03	0.03	0.03
91	0.01	0.03	0.03	0.04	0.05	0.06	0.07	91	0.01	0.03	0.04	0.04	0.05	0.04	0.04
92	0.02	0.03	0.04	0.05	0.05	0.05	0.06	92	0.02	0.03	0.04	0.04	0.04	0.04	0.05
93	0.02	0.02	0.03	0.05	0.04	0.06	0.05	93	0.01	0.05	0.04	0.04	0.05	0.06	0.05
94	0.02	0.01	0.02	0.02	0.04	0.03	0.04	94	0.02	0.05	0.04	0.05	0.06	0.06	0.06
95	0.02	0.00	0.01	0.02	0.01	0.02	0.02	95	0.01	0.05	0.04	0.05	0.05	0.05	0.07
96	0.02	.....	.....	.....	.....	.....	.....	96	0.01	0.05	0.05	0.05	0.06	0.07	0.07

TABLE 3 (continued)

PROPORTION OF FOURTH AND SEVENTH GRADE STUDENTS AT EACH SCORE LEVEL WHO CORRECTLY ANSWER EACH ITEM

SCORE LEVEL	FOURTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL							SCORE LEVEL	SEVENTH GRADE PROPORTION OF STUDENTS AT EACH SCORE LEVEL						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
97	0.01	.....	.....	.....	.....	.....	.....	97	0.01	0.03	0.05	0.05	0.05	0.05	0.07
98	0.02	.....	.....	.....	.....	.....	.....	98	0.02	0.03	0.04	0.03	0.05	0.05	0.05
99	0.03	.....	.....	.....	.....	.....	.....	99	0.01	0.03	0.03	0.02	0.04	0.04	0.04
100	0.02	.....	.....	.....	.....	.....	.....	100	0.02	0.01	0.01	0.01	0.01	0.02	0.01
101	0.02	.....	.....	.....	.....	.....	.....	101	0.01	.....	.....	.....	.....	.....	.....
102	0.03	.....	.....	.....	.....	.....	.....	102	0.03	.....	.....	.....	.....	.....	.....
103	0.02	.....	.....	.....	.....	.....	.....	103	0.03	.....	.....	.....	.....	.....	.....
104	0.03	.....	.....	.....	.....	.....	.....	104	0.03	.....	.....	.....	.....	.....	.....
105	0.03	.....	.....	.....	.....	.....	.....	105	0.02	.....	.....	.....	.....	.....	.....
106	0.02	.....	.....	.....	.....	.....	.....	106	0.03	.....	.....	.....	.....	.....	.....
107	0.02	.....	.....	.....	.....	.....	.....	107	0.04	.....	.....	.....	.....	.....	.....
108	0.02	.....	.....	.....	.....	.....	.....	108	0.04	.....	.....	.....	.....	.....	.....
109	0.02	.....	.....	.....	.....	.....	.....	109	0.03	.....	.....	.....	.....	.....	.....
110	0.01	.....	.....	.....	.....	.....	.....	110	0.03	.....	.....	.....	.....	.....	.....
111	0.03	.....	.....	.....	.....	.....	.....	111	0.05	.....	.....	.....	.....	.....	.....
112	0.01	.....	.....	.....	.....	.....	.....	112	0.03	.....	.....	.....	.....	.....	.....
113	0.01	.....	.....	.....	.....	.....	.....	113	0.02	.....	.....	.....	.....	.....	.....
114	0.01	.....	.....	.....	.....	.....	.....	114	0.01	.....	.....	.....	.....	.....	.....
115	0.00	.....	.....	.....	.....	.....	.....	115	0.00	.....	.....	.....	.....	.....	.....

Table 4 provides some ready indication of how consistently each item administered from 1973 through 1979 retains a measure of difficulty. For the most part, easy items, identified by positive logit values, and hard items, identified by negative logit values, tend to retain this easy or hard characteristic throughout the term of the study. Some items, close to an ideal fit where the difference between the difficulty value computed by the model and that expected under the model would be zero, change in sign occasionally as might be expected. These items tend to have relatively small logit values in absolute terms, indicating that they are very close to fitting the predicted value, which implies a near perfect fit. In fact no item does fit the model perfectly. The zero values shown in the table identify those items that were dropped from the 1973 test.

ESTIMATES OF DIFFICULTY FOR EACH TEST ITEM MEASURED IN TERMS OF LOG-ODDS

TABLE 4

ITEM LEVEL	FOURTH GRADE										ITEM LEVEL	SEVENTH GRADE									
	READING TEST ITEM DIFFICULTY IN LOG-ODDS											READING TEST ITEM DIFFICULTY IN LOG-ODDS									
	1973	1974	1975	1976	1977	1978	1979	1973	1974	1975		1976	1977	1978	1979						
QA1	-0.85	-0.90	-0.92	-0.72	-1.00	-0.97	-1.24	0.09	0.54	-0.37	-0.53	-0.72	-1.26	-1.16							
QA2	-0.30	-0.33	-0.32	-0.15	-0.07	-0.07	-0.39	1.47	1.43	1.43	1.39	1.61	2.03	1.87							
QA3	0.44	0.61	0.39	0.15	0.42	0.25	0.43	0.43	0.42	0.29	0.35	0.35	0.21	0.31							
QA4	-0.26	-0.26	-0.21	-0.29	-0.19	-0.31	-0.32	0.43	0.63	0.34	0.35	0.46	0.42	0.45							
QA5	0.06	0.14	0.09	0.08	-0.03	-0.06	-0.07	0.68	0.84	0.72	0.62	0.61	0.68	0.69							
QB1	-0.29	-0.19	-0.25	-0.23	-0.37	-0.42	-0.35	-0.53	-0.34	-0.34	-0.41	-0.35	-0.20	-0.37							
QB2	-0.12	-0.02	0.07	0.11	0.26	0.15	-0.07	-0.91	-0.88	-0.67	-0.95	-0.98	-0.99	-1.13							
QB3	-0.90	-0.68	-0.72	-0.70	-0.81	-0.91	-1.26	-0.58	-0.76	-0.63	-0.64	-1.16	-1.26	-1.00							
QB4	-0.18	-0.02	-0.15	0.00	0.00	0.12	0.12	0.67	0.56	-0.08	-0.19	-0.25	-0.48	-0.21							
QB5	-0.33	-0.22	0.04	0.27	0.24	0.23	-0.11	0.33	0.63	0.60	1.05	0.98	0.94	1.01							
QC1	1.70	0.00	0.00	0.00	0.00	0.00	0.00	-0.41	0.00	0.00	0.00	0.00	0.00	0.00							
QC2	0.71	0.00	0.00	0.00	0.00	0.00	0.00	-1.04	0.00	0.00	0.00	0.00	0.00	0.00							
QC3	1.03	0.00	0.00	0.00	0.00	0.00	0.00	-0.96	0.00	0.00	0.00	0.00	0.00	0.00							
QC4	-0.29	0.00	0.00	0.00	0.00	0.00	0.00	-0.63	0.00	0.00	0.00	0.00	0.00	0.00							
QC5	1.12	0.00	0.00	0.00	0.00	0.00	0.00	-0.88	0.00	0.00	0.00	0.00	0.00	0.00							
QD1	-0.35	-0.73	-0.56	-0.29	-0.29	-0.15	-0.30	0.45	0.00	0.00	0.00	0.00	0.00	0.00							
QD2	-0.30	-0.35	-0.17	-0.03	0.26	0.06	0.10	-1.04	0.00	0.00	0.00	0.00	0.00	0.00							
QD3	-0.58	-0.61	-0.69	-0.43	-0.39	-0.32	-0.45	0.31	0.00	0.00	0.00	0.00	0.00	0.00							
QD4	0.48	0.11	0.21	0.48	0.69	0.64	0.76	1.09	0.00	0.00	0.00	0.00	0.00	0.00							
QD5	-0.04	-0.10	-0.04	0.09	0.40	0.29	0.19	1.56	0.00	0.00	0.00	0.00	0.00	0.00							
QE1	-2.25	-1.71	-2.12	-2.07	-1.92	-2.12	-1.81	0.14	0.08	0.57	0.62	0.87	0.87	0.70							
QE2	-0.84	-0.49	-0.40	-0.64	-0.34	-0.29	-0.14	0.43	0.47	0.72	0.83	1.01	1.00	0.91							
QE3	-0.94	-0.77	-0.89	-0.86	-1.10	-1.17	-1.00	-0.76	-0.59	-0.07	-0.05	0.27	0.30	0.15							
QE4	-1.72	-1.28	-1.33	-1.28	-1.62	-1.85	-2.26	-0.64	-0.55	-0.09	-0.07	-0.15	0.24	-0.16							

TABLE 4 (continued)

ESTIMATES OF DIFFICULTY FOR EACH TEST ITEM MEASURED IN TERMS OF LOG-ODDS

ITEM LEVEL	FOURTH GRADE							ITEM LEVEL	SEVENTH GRADE						
	READING TEST ITEM DIFFICULTY IN LOG-ODDS								READING TEST ITEM DIFFICULTY IN LOG-ODDS						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
QE5	-1.50	-1.34	-1.35	-1.49	-1.18	-1.74	-1.68	QE5	-1.18	-1.07	-0.58	-0.61	-0.56	-0.26	-0.64
QF1	-0.69	-0.41	-0.63	-0.36	-0.48	-0.45	-0.48	QF1	-0.94	-0.88	-1.05	-0.99	-0.70	-0.98	-1.16
QF2	0.31	0.58	0.39	0.29	0.53	0.49	0.55	QF2	-1.18	-1.03	-0.82	-0.72	-0.55	-0.69	-0.51
QF3	0.64	0.83	0.81	0.54	0.57	0.64	0.61	QF3	-1.33	-1.28	-1.41	-1.22	-1.57	-1.88	-1.78
QF4	0.08	0.28	0.41	-0.00	0.16	0.24	-0.01	QF4	-0.42	-0.30	-0.10	-0.18	0.24	0.31	0.21
QF5	-0.15	0.01	0.11	0.21	0.11	0.15	0.23	QF5	-0.88	-0.97	-0.57	-0.68	-0.77	-0.77	-0.69
QG1	-0.44	-0.69	-0.41	-0.33	-0.36	-0.52	-0.44	QG1	-0.41	-0.64	-0.51	-0.55	-0.80	-0.70	-0.55
QG2	-0.77	-0.80	-0.67	-0.76	-0.89	-0.54	-0.82	QG2	-0.58	-0.42	-0.55	-0.50	-0.33	-0.47	-0.52
QG3	-1.07	-1.42	-0.95	-1.03	-1.32	-1.03	-1.46	QG3	-0.86	-0.99	-0.95	-1.39	-1.60	-1.51	-1.69
QG4	-1.35	-1.24	-1.12	-0.96	-1.39	-1.21	-1.40	QG4	-0.87	-1.04	-0.82	-0.91	-1.67	-1.91	-1.81
QG5	-1.09	-1.29	-1.03	-0.87	-1.23	-0.92	-1.10	QG5	-1.14	-1.13	-0.93	-0.99	-1.55	-1.95	-1.99
QH1	-0.72	-0.69	-0.56	-0.37	-0.42	-0.65	-0.34	QH1	-0.26	-0.43	-0.25	0.20	0.14	0.16	0.39
QH2	-0.55	-0.68	-0.77	-0.49	-0.49	-0.62	-0.68	QH2	-0.54	-0.50	-0.48	-0.65	-0.72	-0.96	-1.07
QH3	-0.76	-0.79	-0.76	-0.60	-0.56	-0.74	-0.82	QH3	-0.26	-0.24	-0.28	-0.19	-0.40	-0.54	-0.45
QH4	-0.49	-0.40	-0.59	-0.40	-0.36	-0.49	-0.44	QH4	-0.60	-0.62	-0.49	-0.71	-0.58	-0.62	-0.92
QH5	-0.83	-0.96	-0.88	-0.72	-0.54	-0.73	-0.82	QH5	-0.42	-0.46	-0.75	-0.13	-0.42	-0.00	0.18
QI1	-0.21	0.02	-0.00	0.11	-0.11	-0.29	-0.07	QI1	0.76	0.63	0.73	0.67	0.57	0.63	0.73
QI2	-0.26	-0.36	-0.00	-0.31	-0.18	-0.14	-0.05	QI2	-0.26	-0.26	-0.34	-0.53	-0.92	-1.13	-1.18
QI3	-0.49	-0.51	-0.45	-0.65	-0.63	-0.65	-0.58	QI3	0.17	0.11	0.11	-0.06	0.19	0.21	0.20
QI4	-0.40	-0.25	-0.57	-0.38	-0.54	-0.20	-0.16	QI4	0.23	0.38	0.20	0.18	0.48	0.46	0.49
QI5	0.21	0.67	0.75	1.85	1.78	1.79	1.98	QI5	0.29	0.45	0.13	0.31	0.31	0.27	0.46
QJ1	-0.73	-0.73	-0.69	-0.61	-0.94	-0.76	-0.51	QJ1	-0.39	0.07	-0.06	-0.25	-0.11	-0.04	0.09
QJ2	-0.28	-0.28	-0.09	-0.19	-0.13	-0.06	0.00	QJ2	-0.15	-0.09	-0.11	0.20	0.36	0.49	0.56
QJ3	-0.17	-0.34	-0.26	-0.34	-0.19	-0.10	-0.35	QJ3	0.11	0.16	0.19	-0.06	0.23	0.05	0.22

TABLE 4 (continued)  
ESTIMATES OF DIFFICULTY FOR EACH TEST ITEM MEASURED IN TERMS OF LOG-ODDS

ITEM LEVEL	FOURTH GRADE							ITEM LEVEL	SEVENTH GRADE						
	READING TEST ITEM DIFFICULTY IN LOG-ODDS								READING TEST ITEM DIFFICULTY IN LOG-ODDS						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
QJ4	1.51	1.71	0.61	0.85	0.95	0.84	0.95	QJ4	0.30	0.35	0.23	0.21	0.48	0.44	0.50
QJ5	0.48	0.76	2.00	1.35	1.21	1.23	1.48	QJ5	0.37	0.32	0.22	0.03	0.08	0.11	0.10
QK1	-0.07	0.29	0.33	0.17	0.45	0.33	0.33	QK1	-1.51	-1.43	-1.31	-1.60	-2.36	-2.51	-1.81
QK2	0.51	0.69	0.79	1.02	1.24	1.29	1.31	QK2	-0.08	-0.06	0.19	0.23	0.50	0.33	0.11
QK3	0.16	0.06	0.34	0.40	0.31	0.47	0.32	QK3	-0.56	-0.64	-0.63	-0.37	-0.76	-0.47	-0.54
QK4	-0.25	0.20	0.01	0.16	1.33	1.49	1.58	QK4	-0.75	-0.53	-0.48	-0.45	-0.80	-0.64	-0.78
QK5	0.48	0.70	0.74	0.26	0.35	0.43	0.61	QK5	0.79	0.95	1.20	1.21	1.30	1.42	1.39
QL1	0.42	0.16	1.22	1.07	1.18	1.23	1.27	QL1	-0.21	0.05	-0.20	-0.10	-0.20	-0.36	-0.44
QL2	1.37	1.60	0.31	0.26	0.42	0.36	0.46	QL2	-0.48	-0.30	-0.43	-0.50	-0.73	-1.11	-1.20
QL3	0.44	0.57	0.01	0.31	0.21	0.01	0.11	QL3	-0.75	-0.40	-0.12	-0.19	-0.33	-0.22	-0.40
QL4	0.96	1.16	1.75	1.97	1.58	1.95	2.27	QL4	0.04	0.22	-0.51	0.58	0.55	0.70	0.62
QL5	1.46	1.46	0.93	1.55	1.44	1.47	1.60	QL5	1.13	0.96	1.04	0.88	1.12	1.02	1.14
QM1	-0.49	-0.20	-0.18	0.31	0.14	0.24	0.19	QM1	-0.40	-0.44	-0.39	0.91	1.07	1.02	1.12
QM2	-0.57	-0.37	-0.39	-0.26	-0.21	-0.24	-0.49	QM2	0.88	0.84	0.83	0.97	1.23	1.49	1.40
QM3	0.74	1.09	1.16	1.52	1.26	1.41	1.63	QM3	1.66	1.31	0.93	0.52	0.61	1.01	0.68
QM4	0.49	0.62	0.68	1.00	0.93	0.88	1.13	QM4	0.28	0.22	0.17	0.07	0.21	0.28	0.22
QM5	1.47	1.58	1.96	2.12	2.14	2.28	2.35	QM5	0.47	0.50	0.47	0.58	1.33	1.15	1.20
QN1	0.06	0.0	0.0	0.0	0.0	0.0	0.0	QN1	-0.23	0.30	-0.19	-0.22	-0.35	-0.17	-0.19
QN2	0.55	0.0	0.0	0.0	0.0	0.0	0.0	QN2	-0.46	-0.39	-0.48	-0.55	-0.42	-0.32	-0.59
QN3	1.80	0.0	0.0	0.0	0.0	0.0	0.0	QN3	0.73	1.00	0.72	0.79	1.09	1.10	1.10
QN4	0.81	0.0	0.0	0.0	0.0	0.0	0.0	QN4	-0.00	0.11	0.12	-0.07	-0.11	-0.07	0.17
QN5	0.74	0.0	0.0	0.0	0.0	0.0	0.0	QN5	0.50	0.41	0.32	0.30	0.32	0.35	0.26
QO1	0.33	0.0	0.0	0.0	0.0	0.0	0.0	QO1	0.41	0.59	0.43	0.48	0.57	0.64	0.58
QO2	1.70	0.0	0.0	0.0	0.0	0.0	0.0	QO2	0.23	0.39	0.22	0.18	0.46	0.55	0.48



TABLE 4 (continued)  
ESTIMATES OF DIFFICULTY FOR EACH TEST ITEM MEASURED IN TERMS OF LOG-ODDS

ITEM LEVEL	FOURTH GRADE										ITEM LEVEL	SEVENTH GRADE									
	READING TEST ITEM DIFFICULTY IN LOG-ODDS											READING TEST ITEM DIFFICULTY IN LOG-ODDS									
	1973	1974	1975	1976	1977	1978	1979	1973	1974	1975		1976	1977	1978	1979						
Q03	1.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Q03	0.51	0.88	0.38	0.30	0.22	0.22	0.42			
Q04	1.76	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Q04	0.45	0.16	0.13	0.03	0.31	0.21	0.19			
Q05	0.60	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Q05	0.95	0.89	1.20	1.02	1.29	1.30	1.19			
QP1	-1.06	-0.87	-0.93	-0.73	-1.00	-1.03	-1.46	-1.03	-1.46	-1.46	QP1	-0.77	-0.72	-0.64	-0.74	-1.38	-1.43	-1.41			
QP2	-0.89	-0.73	-0.76	-0.65	-0.81	-0.86	-0.70	-0.86	-0.70	-0.70	QP2	-0.64	-0.35	-0.31	-0.03	-0.06	-0.05	-0.11			
QP3	-0.14	-0.18	-0.15	-0.21	-0.04	0.31	0.08	0.31	0.08	0.08	QP3	0.75	0.65	0.92	0.73	0.96	0.91	1.24			
QP4	0.26	0.55	0.54	0.54	0.53	0.71	0.63	0.71	0.63	0.63	QP4	1.05	0.92	1.03	0.83	1.01	1.03	1.16			
QP5	0.24	0.48	0.51	0.45	0.48	0.67	0.56	0.67	0.56	0.56	QP5	0.74	0.93	0.87	0.95	1.12	1.24	1.19			
Q01	0.21	0.76	0.14	-0.51	-0.23	-0.45	-0.31	-0.45	-0.31	-0.31	Q01	-0.95	-0.82	-1.03	-0.86	-1.57	-1.48	-1.43			
Q02	0.49	0.68	0.04	-1.15	-1.09	-1.30	-1.17	-1.30	-1.17	-1.17	Q02	-0.69	-0.43	-0.39	0.54	0.76	0.79	0.91			
Q03	0.21	0.72	0.20	-0.07	0.11	0.17	0.41	0.17	0.41	0.41	Q03	-0.34	-0.41	-0.24	-0.43	-0.47	-0.47	-0.41			
Q04	0.34	0.57	0.41	-1.36	-1.46	-1.50	-1.31	-1.50	-1.31	-1.31	Q04	0.65	-0.45	0.82	0.55	0.57	0.70	0.75			
Q05	0.05	0.16	-0.11	-0.84	-1.00	-0.72	-0.80	-1.00	-0.80	-0.80	Q05	0.60	0.71	0.59	0.62	0.62	0.76	0.65			
QR1	-0.33	-0.15	-0.07	-0.07	0.19	-0.09	0.24	0.19	-0.09	0.24	QR1	0.36	0.56	0.52	0.48	0.23	0.31	0.47			
QR2	-0.81	-0.87	-0.75	-0.61	-0.87	-0.92	-0.89	-0.92	-0.89	-0.89	QR2	-0.27	-0.41	-0.22	-0.26	-0.15	-0.08	-0.17			
QR3	-0.11	-0.07	0.04	0.12	0.24	0.44	0.16	0.24	0.44	0.16	QR3	-0.69	-0.51	-0.59	-0.76	-0.55	-0.82	-0.78			
QR4	0.13	0.39	0.15	0.25	0.15	0.07	0.05	0.15	0.07	0.05	QR4	0.50	0.62	0.45	0.59	0.73	0.52	0.70			
QR5	0.13	0.14	0.35	0.52	0.35	0.51	0.63	0.35	0.51	0.63	QR5	0.75	0.79	1.01	1.00	1.01	1.06	0.82			
QS1	0.37	0.74	0.77	0.85	0.83	0.90	0.90	0.83	0.90	0.90	QS1	0.28	0.38	0.31	0.35	0.40	0.26	0.28			
QS2	-0.19	0.06	0.21	0.38	0.31	0.65	0.74	0.31	0.65	0.74	QS2	-0.67	-0.36	-0.37	0.57	0.28	0.56	0.61			
QS3	0.72	0.98	0.99	1.18	0.94	1.13	1.18	0.94	1.13	1.18	QS3	0.11	0.33	0.49	0.29	0.10	0.41	0.39			
QS4	0.40	0.73	0.50	0.72	0.48	0.64	0.64	0.48	0.64	0.64	QS4	0.53	0.40	0.39	0.27	0.34	0.31	0.51			
QS5	0.15	0.17	0.14	0.37	0.26	0.34	0.20	0.26	0.34	0.20	QS5	0.33	0.32	0.31	0.38	0.35	0.60	0.45			
QT1	1.04	1.14	1.35	-0.65	-0.59	-0.89	-0.91	-0.59	-0.89	-0.91	QT1	1.35	0.0	0.0	0.0	0.0	0.0	0.0			

TABLE 4 (continued)

ESTIMATES OF DIFFICULTY FOR EACH TEST ITEM MEASURED IN TERMS OF LOG-ODDS

ITEM LEVEL	FOURTH GRADE							ITEM LEVEL	SEVENTH GRADE						
	READING TEST ITEM DIFFICULTY IN LOG-ODDS								READING TEST ITEM DIFFICULTY IN LOG-ODDS						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
QT2	-0.09	0.32	0.35	0.34	0.41	0.31	0.41	QT2	1.42	0.0	0.0	0.0	0.0	0.0	0.0
QT3	-1.27	-0.83	-0.87	-0.89	-1.23	-1.32	-1.19	QT3	1.73	0.0	0.0	0.0	0.0	0.0	0.0
QT4	1.28	1.74	1.87	0.73	0.77	0.68	0.80	QT4	1.46	0.0	0.0	0.0	0.0	0.0	0.0
QT5	-0.81	-0.66	-0.58	-0.31	-0.32	-0.32	-0.42	QT5	1.13	0.0	0.0	0.0	0.0	0.0	0.0
QU1	-0.50	-0.11	0.01	0.11	-0.05	-0.29	-0.34	QU1	-0.38	-0.12	0.32	-0.10	0.11	0.02	0.29
QU2	-1.13	-0.88	-0.86	-0.63	-0.48	-0.85	-0.89	QU2	-1.07	-0.78	-0.77	-0.91	-0.85	-0.87	-1.24
QU3	1.01	1.07	1.29	1.57	1.67	1.78	1.66	QU3	0.07	0.10	0.38	-0.29	-0.26	-0.32	-0.55
QU4	-0.22	-0.12	-0.03	0.12	0.24	0.01	-0.06	QU4	0.14	0.19	-1.06	-1.22	-1.47	-1.68	-1.60
QU5	-1.10	-0.77	-0.84	-0.58	-0.64	-0.62	-0.78	QU5	-1.29	-0.99	0.29	0.03	-0.05	0.09	0.39
QV1	-0.33	-0.12	-0.02	-0.03	0.01	0.21	0.34	QV1	-1.45	-1.34	-1.56	-1.62	-2.36	-2.57	-2.21
QV2	-0.85	-0.64	-0.63	-0.80	-0.77	-0.76	-1.02	QV2	1.31	0.45	0.64	0.97	1.86	1.55	1.73
QV3	0.47	0.45	0.51	0.81	0.71	0.69	0.97	QV3	-0.35	-0.33	-0.32	-0.26	-0.16	-0.06	-0.23
QV4	0.96	0.89	0.28	0.60	0.54	0.65	0.43	QV4	-0.97	-1.10	-1.15	-1.20	-1.55	-1.81	-2.07
QV5	0.65	0.81	1.00	0.97	0.99	0.82	1.05	QV5	1.86	1.93	2.05	1.06	1.23	1.42	1.15
QW1	-1.10	0.0	0.0	0.0	0.0	0.0	0.0	QW1	-0.74	-0.65	-0.86	-0.71	-1.10	-0.96	-0.98
QW2	0.35	0.0	0.0	0.0	0.0	0.0	0.0	QW2	-0.50	-0.46	-0.62	-0.57	-0.60	-0.92	-0.85
QW3	-0.77	0.0	0.0	0.0	0.0	0.0	0.0	QW3	0.46	0.45	0.49	0.44	0.61	0.50	0.48
QW4	-0.05	0.0	0.0	0.0	0.0	0.0	0.0	QW4	-0.05	0.16	0.19	0.14	0.28	0.40	0.49
QW5	0.54	0.0	0.0	0.0	0.0	0.0	0.0	QW5	1.34	1.27	1.27	1.38	1.51	1.86	1.81

Question 2: Is there a statistically significant increase in the measurement efficiency of MEAP reading tests after items which do not fit the Rasch model, because they are too difficult, have been credited to students who get these items wrong?

#### STATISTICAL ANALYSIS

Four tables were developed in the process of seeking an answer to this research question (i.e., Table 5 through Table 8). Each table presents a comparison of selected statistical properties associated with the number of test learning objectives passed before and after credit is given for missed items that are so difficult that they do not fit the Rasch model. The first three tables in this group are summarized directly from the results printed at the conclusion of each SPSS run applied to the fourteen individual samples. Table 6 shows the effect which the re-scoring process has on the average number of MEAP Reading Test learning objectives passed. Table 7 shows the effect of this adjustment on objective score variance, and Table 8 shows the effect on the standard error of the objective scores. Table 5 shows the t-statistic developed on the basis of a directional comparison between Reading Test learning objective score means. With the exception of the 1979 seventh grade comparison, the t-statistics all exceed the critical value. In 1979, there was one seventh grade item that did not fit the Rasch model. It did not affect the corresponding learning objective score. The 1973 learning objective scores include scores on those objectives which are not actually administered all of the seven years from 1973 through 1979.

**HYPOTHESIZED FINDINGS****Hypothesis H:O<sub>2</sub>:**

There is no statistically significant increase in the measurement efficiency of MEAP reading tests after items which do not fit the Rasch model, because they are too difficult, have been re-scored and credited to students who get these items wrong at a probability level of 0.05, or less.

Table 5 indicates this hypothesis must be rejected.

TABLE 5

t-STATISTIC AND ONE-TAIL PROBABILITY LEVEL (i.e., ALPHA LEVEL) DERIVED ON COMPARISON OF AVERAGE SCORE OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST

t OR ALPHA VALUE	FOURTH GRADE t-STATISTIC AND ONE-TAIL PROBABILITY							t OR ALPHA VALUE	SEVENTH GRADE t-STATISTIC AND ONE-TAIL PROBABILITY						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
SAMPLE SIZE .	1,000	1,000	1,000	1,000	1,000	1,000	1,000	SAMPLE SIZE .	1,000	1,000	1,000	1,000	1,000	1,000	1,000
COMP. t-VALUE	23.28	13.90	10.04	13.00	09.86	06.54	10.25	COMP. t-VALUE	13.71	17.08	17.51	11.23	09.87	09.45	00.00
TABLE t-VALUE AT ALPHA 0.05 . .	1.71	1.73	1.73	1.73	1.73	1.73	1.73	TABLE t-VALUE AT ALPHA 0.05 . .	1.71	1.72	1.72	1.72	1.72	1.72	1.72

OBJECTIVES DROPPED FROM LATER TESTS ARE INCLUDED IN THE COMPUTATION OF 1973 VALUES.

In this group of four tables, Table 5 is the one which has a direct bearing on the answer to this research question. The computed t-values in this table were compared to critical t-values at an alpha level of 0.05: 1.71 for all 1973 tests; 1.72 for 1974 to 1979 seventh grade tests; and 1.73 for 1974 to 1979 fourth grade tests. Excepting the 1979 seventh grade results, they all exceed the critical t-values which would define areas of rejection under the alpha criteria established in this study at 0.05, or five chances in one hundred occurrences. Therefore, Table 5 demonstrates, with the one exception noted, that there was a statistically significant shift in the average number of passed learning objectives. No change, whatever, is indicated in the 1979 objective score for seventh graders.

#### UNHYPOTHESIZED FINDINGS

Table 6 reveals that reported average objective scores for the fourth grade tests increased each succeeding year from 1974 through 1979. Excepting a drop from 1975 to 1976, a similar pattern is revealed in the reported average objective scores for the seventh grade tests from 1974 through 1979.

TABLE 6

THE AVERAGE SCORE OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST

BEFORE/ AFTER ITEM DELETION	FOURTH GRADE AVERAGE SCORE ON READING OBJECTIVES							BEFORE/ AFTER ITEM DELETION	SEVENTH GRADE AVERAGE SCORE ON READING OBJECTIVES						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
BEFORE	12.6	11.5	12.7	12.9	13.6	14.0	14.7	BEFORE	14.5	12.5	13.1	13.0	15.2	15.5	15.8
AFTER	13.3	11.7	12.8	13.0	13.7	14.1	14.8	AFTER	14.7	12.8	13.4	13.2	15.3	15.6	15.8

OBJECTIVES DROPPED FROM LATER TESTS ARE INCLUDED IN THE COMPUTATION OF 1973 VALUES.

The average number of objectives passed after re-scoring is shown in Table 6. Objective score behavior closely parallels the decline in the incidence of difficult items which do not fit the Rasch model. Of course it would seem reasonable to expect the total number of passed items to increase as the incidence of very difficult items falls off, and this is precisely what these data would suggest. The re-scoring process has increased the average in every case but one; the 1979 seventh grade objective scores. However, since there were no very difficult items in that test, no change in objective scores could be anticipated in the re-scoring process. The most noteworthy condition suggested by this table is that the average objective scores before and after the re-scoring process tend to converge, and the 1979 seventh grade scores before and after re-scoring did converge completely. For the most part, the re-scoring process resulted in little change of the average, reported objective score. Most of the differences were one or two tenths of a point. The improvement in fourth grade objective scores was 2.1 points overall while the seventh grade scores improved by only 1.3 points during the period of this study. Fourth grade scores were nearly two points below their seventh grade counterparts in 1973, but closed the gap by eight tenths of a point at the end of the study period in 1979. However, since there is one less objective in the 1974 through 1979 fourth grade tests than the seventh grade tests for those years, there is probably no real difference between the grades on objective performance.

Table 7 shows a general decrease in objective score variance over the seven years of this study. It also shows that the general effect of



the re-scoring process results in rather consistent reduction in that variance.

TABLE 7

THE SCORE VARIANCE OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST

BEFORE/ AFTER ITEM DELETION	FOURTH GRADE SCORE VARIANCE ON READING OBJECTIVES							BEFORE/ AFTER ITEM DELETION	SEVENTH GRADE SCORE VARIANCE ON READING OBJECTIVES						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
BEFORE	63.5	41.5	40.4	41.3	35.5	32.5	27.6	BEFORE	63.5	51.2	49.8	49.2	33.9	30.7	28.1
AFTER	55.9	39.5	38.9	39.4	34.5	31.9	26.9	AFTER	61.2	46.7	45.9	48.3	33.0	30.0	28.1

OBJECTIVES DROPPED FROM LATER TESTS ARE INCLUDED IN THE COMPUTATION OF 1973 VALUES.

Again, the tendency for the data to converge is observed in Table 7 to follow much the same pattern that was observed in Table 6. Re-scoring seemed to have its most pronounced effect on the variance for 1973 fourth grade objective scores and for the 1974 and 1975 seventh grade scores. The re-scoring process tended to reduce score variance in every test but the 1979 seventh grade test which had no excessively difficult items.

Table 8 shows a general decrease in the standard error over the seven years of this study. It also shows that the general effect of the re-scoring process is a consistent reduction in the magnitude of the standard error. Once again, the tendency for the statistical results produced in the analysis of the data to converge is demonstrated in Table 8. Here the standard error of estimate follows the same pattern as the average reading objective score in Table 6 and the reading objective score variance in Table 7. That is, with the passage of time, the process of re-scoring very difficult items in this study that do not fit the Rasch model has less impact on data from the last year covered in this study, 1979, than it did on data from the first year, 1973. The differences which result in the objective score mean, variance, and standard error as a result of this re-scoring process tend to be small.

TABLE 8

THE STANDARD ERROR OF ESTIMATE ON SCORES OF FOURTH AND SEVENTH GRADE STUDENTS ON MEAP READING OBJECTIVES BEFORE AND AFTER DIFFICULT ITEMS THAT DO NOT FIT THE RASCH MODEL HAVE BEEN DELETED FROM THE TEST

BEFORE/ AFTER ITEM DELETION	FOURTH GRADE STANDARD ERROR OF ESTIMATE							BEFORE/ AFTER ITEM DELETION	SEVENTH GRADE STANDARD ERROR OF ESTIMATE						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
BEFORE	.252	.204	.201	.203	.188	.180	.166	BEFORE	.252	.226	.223	.222	.184	.175	.168
AFTER	.236	.199	.197	.199	.186	.179	.164	AFTER	.247	.216	.214	.220	.182	.173	.168

OBJECTIVES DROPPED FROM LATER TESTS ARE INCLUDED IN THE COMPUTATION OF 1973 VALUES.

Question 3: Is there any change in pattern regarding item fit to the Rasch model which would suggest either an increase or decrease in difficulty as items are used over time?

#### STATISTICAL ANALYSIS

Two tables were developed in the process of seeking an answer to this research question (i.e., Table 9 and Table 10). Each table presents information on the prospect of passing MEAP Reading Test learning objectives over time in terms of the number of items passed which make up each objective. Table 10 shows a tabulation, by objective, of items that are too difficult to fit the Rasch model from 1973 through 1979. This table includes a summary total and average of these items, by year. Table 9 shows the  $X^2$  statistic developed on the basis of a comparison between the proportion of too-difficult items in the last year of the sequence, 1979, and the preceding year in which the largest proportion of too-difficult items occurred. Appropriate adjustment was made to the 1973 item score for the fourth grade students, the year most too-difficult items occurred for this group, to delete items not carried forward in succeeding years so that only the scores on items actually administered each of the years from 1973 to 1979 are compared. The seventh grade comparison is drawn between 1979 and 1974, the year when most too-difficult items occurred for this group.

#### HYPOTHESIZED FINDINGS

Hypothesis H:0<sub>3</sub>:

There is no statistically significant decrease in the average number of difficult items in the MEAP Reading Test over time,

measured between the 1979 Test and the earlier Test that contains the largest number of items that are too difficult to fit the Rasch model, at a probability level of 0.05, or less.

Table 9 indicates this hypothesis must be rejected.

TABLE 9

X<sup>2</sup>-STATISTIC DERIVED ON COMPARISON  
 OF THE LARGEST PROPORTION OF DIFFICULT ITEMS TO THE 1979 PROPORTION OF DIFFICULT ITEMS IN  
 THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHICH DO NOT FIT THE RASCH MODEL

CATEGORY	FOURTH GRADE		CATEGORY	SEVENTH GRADE	
	1973	1979		1974	1979
NUM. ITEMS	95	95	NUM. ITEMS	100	100
HARD ITEMS	15	3	HARD ITEMS	11	0
PROPORTION	.16	.03	PROPORTION	.11	.00
1973 COMPARED TO 1979			1974 COMPARED TO 1979		
COMPUTED X <sup>2</sup> -VALUE	570.751		COMPUTED X <sup>2</sup> -VALUE	706.360	
TABLE X <sup>2</sup> -VALUE/ ALPHA = .05	27.59		TABLE X <sup>2</sup> -VALUE/ ALPHA = .05	19.68	

TABLE 9 (continue)

X<sup>2</sup>-STATISTIC DERIVED ON COMPARISON  
 OF THE LARGEST PROPORTION OF DIFFICULT ITEMS TO THE 1979 PROPORTION OF DIFFICULT ITEMS IN  
 THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHICH DO NOT FIT THE RASCH MODEL

CATEGORY	FOURTH GRADE	CATEGORY	SEVENTH GRADE
1973 COMPARED TO 1979		1974 COMPARED TO 1979	
CRAMER'S V-STATISTIC/ ALPHA=.05	0.53421	CRAMER'S V-STATISTIC/ ALPHA=.05	0.59429

ITEMS DROPPED FROM LATER TESTS ARE NOT INCLUDED IN THE COMPUTATION OF THESE VALUES.



In this group of two tables, Table 9 is the one which has a direct bearing on this research question. The critical value for the fourth grade  $\chi^2$  statistic is 27.59 at 17 degrees of freedom. The critical value for the seventh grade  $\chi^2$  statistic is 19.68 at 11 degrees of freedom. Both of these values are determined at an alpha level of 0.05. The "COMPUTED  $\chi^2$ " values shown in the table are critical at 1 in 10,000 occurrences. Therefore, the  $\chi^2$ -statistics computed for both fourth grade and seventh grade students exceed the critical  $\chi^2$  values which would define areas of rejection under the alpha criteria established in this study at 0.05, or five chances in one hundred occurrences. Table 9 demonstrates a statistically significant shift in the proportion of difficult items which occurred in MEAP reading tests between the last year covered in this analysis, 1979, and the preceding year, back to 1973 or 1974, in which the greatest number of difficult items occurred.

The Cramer's V statistic which corresponds to the respective fourth grade and seventh grade  $\chi^2$ -value is also presented in Table 9. Cramer's V provides a means for determining the strength of relationship measured by the  $\chi^2$  value. Cramer's V corresponding to the fourth grade  $\chi^2$  is 0.53421, and Cramer's V corresponding to the seventh grade  $\chi^2$  is 0.59429. Both values suggest a moderate degree of association does exist between the proportion of difficult items that do not fit the Rasch model in the years compared.

The number of items which fit the Rasch model under the criteria established in this study for the entire sample-of-1000 fourth grade students was 90,894 out of a possible 95,000, or 95.68%, in the 1973 test. The number of these items in the 1979 fourth grade test rose to 94,577, or 99.55%. The number of items which fit the Rasch model under

these same criteria for the entire sample-of-1000 seventh grade students was 97,680 out of a possible 100,000, or 97.68%, in the 1974 test. The number of these items in the 1979 seventh grade test rose to 100,000, or 100%.

#### CRAMER'S V

The  $\chi^2$ -statistic determines whether or not two variables are dependent or independent, but even when a significant relationship is demonstrated, the statistic does not provide any indication of the strength of this relationship (i.e., correlation).

By itself, chi-square helps us only decide whether our variables are independent or related. It does not tell us how strongly they are related. Part of the reason is that the sample size and table size have such an influence upon chi-square. Several statistics which adjust for these factors are available. When chi-square is thus adjusted it becomes the basis for assessing strength of relationship. (Nie et. al., 1975, p. 224)

Cramer's V is a modification of the phi statistic and constitutes a measure of correlation between two variables where one or both has more than two values. It ranges from 0 to 1. "Thus a large value of V merely signifies that a high degree of association exists, without revealing the manner in which the variables are associated" (Nie, et. al., 1975, p. 225). In this instance, Cramer's V is 0.53421 for the fourth grade sample and 0.59429 for the seventh grade sample. Both values suggest a moderate degree of association exists between the proportion of very difficult items in the 1979 test and the largest proportion found in any previous test as far back as 1073.

#### UNHYPOTHESIZED FINDINGS

Table 10 reveals that by far the greater number of items that were so difficult that they did not fit the Rasch model occurred in the first

four years of the fourth grade test and the first three years of the seventh grade test.

TABLE 10

DIFFICULT ITEMS IN THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHICH DO NOT FIT THE RASCH MODEL SHOWN BY QUESTION NUMBER 1 TO 5 WITHIN TEST OBJECTIVE

READING TEST OBJECTIVE	FOURTH GRADE NUMBER OF DIFFICULT ITEMS (i. e., 1 to 5)							READING TEST OBJECTIVE	SEVENTH GRADE NUMBER OF DIFFICULT ITEMS (i. e., 1 to 5)						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
A . . .	1..4.	1..4.	1..4.	1....	1....	.....	.....	A . . .	.....	.....	.....	.....	.....	.....	
B . . .	.234.	.3.	.3.	.3.	.....	.....	.....	B . . .	.3.	.3.	.23.	.3.	.3.	.....	
C . . .	.....	.....	.....	.....	.....	.....	.....	C . . .	.....	.....	.....	.....	.....	.....	
D . . .	.....	.....	.....	.....	.....	.....	.....	D . . .	.....	.....	.....	.....	.....	.....	
E . . .	.2.	.....	.2.	.2.	.2.	.2.	.2.	E . . .	.....	.....	.....	.....	.....	.....	
F . . .	.5	1.	.....	.4.	1.	1.	1..4.	F . . .	.4.	1..4.	.4.	.....	.....	.2.	
G . . .	.3.	.3.	.....	.3.	.....	.....	.....	G . . .	.....	.4.	.....	.....	.....	.....	
H . . .	.2.4.	.2.	.2.	.....	.....	.....	.4.	H . . .	.2.	12.5	.23.5	.....	.....	.....	
I . . .	.4.	.4.	.....	.....	.....	.....	.....	I . . .	.....	.....	.....	.....	.....	.....	
J . . .	.....	.....	.....	.....	.....	.....	.....	J . . .	.....	.....	.....	.....	.....	.....	
K . . .	.4.	.....	.....	.....	.....	.....	.....	K . . .	.....	.3.	.4.	.2.	.2.	.....	
L . . .	.....	.....	.....	.....	.....	.....	.....	L . . .	.....	.....	.4.	.2.	.2.	.....	
M . . .	.....	.....	.....	.....	.....	.....	.....	M . . .	.....	.....	.....	.....	.....	.....	
N . . .	.....	.....	.....	.....	.....	.....	.....	N . . .	1.	.....	.....	.....	.....	.....	
O . . .	.....	.....	.....	.....	.....	.....	.....	O . . .	.....	.....	.....	.....	.....	.....	
P . . .	1.	.....	.....	.....	.....	.....	.....	P . . .	.2.	.2.	.2.	.....	.....	.....	
Q . . .	.....	.....	.....	.....	.....	.....	.....	Q . . .	.....	.....	.....	.....	.....	.....	
R . . .	.2.	.2.	.....	1.	1.	.....	.....	R . . .	.....	.....	.....	.....	.....	.....	
S . . .	.....	.....	.....	.....	.....	.....	.....	S . . .	.2.	.2.	.2.	.....	.....	.....	
T . . .	.3.	.3.	.3.	1.3.	.....	.....	.....	T . . .	.....	.....	.....	.....	.....	.....	

TABLE 10 (continued)

DIFFICULT ITEMS IN THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHICH DO NOT FIT THE RASCH MODEL SHOWN BY QUESTION NUMBER 1 TO 5 WITHIN TEST OBJECTIVE

READING TEST OBJECTIVE	FOURTH GRADE NUMBER OF DIFFICULT ITEMS (i. e., 1 to 5)							READING TEST OBJECTIVE	SEVENTH GRADE NUMBER OF DIFFICULT ITEMS (i. e., 1 to 5)						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
U . . .	.....	.....	.....	.....	.....	.....	.....	U . . .	.....	.....	.....	.....	.....	.....	.....
V . . .	.....	.....	.....	.....	.....	.....	.....	V . . .	..4.	.....	..3.	..3.	..3.	..3.	.....
W . . .	..34.	.....	.....	.....	.....	.....	.....	W . . .	.....	..2.	..2.	..2.	.....	.....	.....
TOTAL: .	17	9	6	8	4	3	3	TOTAL: .	7	11	11	4	3	2	1
AVERAGE:	.15	.09	.06	.08	.04	.03	.03	AVERAGE:	.06	.11	.11	.04	.03	.02	.01

Examination of Table 10 reveals that the total of very difficult items was 29 for 1973 through 1975 for both the fourth grade and the seventh grade samples. For the last three years studied, 1977 through 1979, the fourth grade total was only ten hard items, and the seventh grade total was just four. For the fourth grade median year, 1976, there were eight hard items; for the seventh grade median year, 1976, there were only three hard items. This table reveals clearly that both grades encountered roughly the same proportion of hard items in the first three tests covered by this investigation. In the fourth year, the fourth grade students continued to experience about the same proportion of difficult items as this grade encountered in the first three years. In the last three tests studied, the incidence of hard items in fourth grade tests dropped by better than 60%. Overall, almost 7% of the items encountered by the fourth graders were so difficult that they did not fit the Rasch model. A decline of hard items in seventh grade tests occurred sooner and to an even greater extent than it did in the fourth grade tests. Overall, just over 5% of the items encountered by the seventh graders were so hard that they did not fit the Rasch model.

Question 4: Can a negative effect of items identified as being too difficult to fit the Rasch model be demonstrated on the probability that a student will pass the MEAP Reading Test learning objectives by re-scoring these items in favor of the student and treating the items as if they had been originally calibrated to fit the Rasch model?

#### STATISTICAL ANALYSIS

Two tables were developed in the process of seeking an answer to this research question (i.e., Table 11 and Table 12). These tables present information on the prospect of passing MEAP Reading Test learning objectives over time in terms of probability. Table 12 presents the probability of passing each objective incorporated in the MEAP Reading Test from 1973 to 1979. The table shows three possibilities: the probability of passing each learning objective when no item is too difficult (i.e., a probability of .33); when one item is too difficult (i.e., a probability of .17); and when two or more items are too difficult (i.e., a probability of 0). By definition, a passable objective is an objective having four or more items that are not too difficult to fit the Rasch model. Table 11 shows the  $\chi^2$ -statistic developed on the basis of a comparison between the proportion of passable objectives in the last year of the sequence, 1979, and the preceding year having the smallest proportion of passable objectives. Appropriate adjustment was made to the 1973 objective score for the fourth grade students, the year in which the smallest proportion occurred for this group, to delete objectives not carried forward in succeeding years so that only the scores on objectives actually administered each of the years from 1973 to 1979 are compared. The

seventh grade comparison is drawn between 1979 and 1974, the year when the smallest proportion of possible objectives occurred for this group.

#### HYPOTHESIZED FINDINGS

##### Hypothesis H:O<sub>4</sub>:

There is no statistically significant increase in the proportion of passable learning objectives in the MEAP Reading Test over time, measured between the 1979 Test and the earlier Test that contains the smallest number of passable learning objectives, at a probability level of 0.05, or less.

Table 11 indicates this hypothesis must be rejected.



TABLE 11

X<sup>2</sup>-STATISTIC DERIVED ON COMPARISON OF THE  
 SMALLEST PROPORTION OF PASSABLE OBJECTIVES TO THE PROPORTION OF PASSABLE OBJECTIVES IN THE 1979  
 FOURTH AND SEVENTH GRADE MEAP READING TESTS AFTER DELETING HARD ITEMS WHICH DO NOT FIT THE RASCH MODEL

CATEGORY	FOURTH GRADE		CATEGORY	SEVENTH GRADE	
	1973	1979		1974	1979
NUMBER OF OBJECTIVES	19	19	NUMBER OF OBJECTIVES	20	20
PASSABLE OBJECTIVES	16	17	PASSABLE OBJECTIVES	18	20
PROPORTION	.84	.89	PROPORTION	.90	1.00
1973 COMPARED TO 1979			1974 COMPARED TO 1979		
COMPUTED X <sup>2</sup> -VALUE	331.158		COMPUTED X <sup>2</sup> -VALUE	312.139	
TABLE X <sup>2</sup> -VALUE/ ALPHA = .05	11.07		TABLE X <sup>2</sup> -VALUE/ ALPHA = .05	7.81	

TABLE 11 (continued)

X<sup>2</sup>-STATISTIC DERIVED ON COMPARISON OF THE  
 SMALLEST PROPORTION OF PASSABLE OBJECTIVES TO THE PROPORTION OF PASSABLE OBJECTIVES IN THE 1979  
 FOURTH AND SEVENTH GRADE MEAP READING TESTS AFTER DELETING HARD ITEMS WHICH DO NOT FIT THE RASCH MODEL

CATEGORY	FOURTH GRADE	CATEGORY	SEVENTH GRADE
1973 COMPARED TO 1979		1974 COMPARED TO 1979	
CRAMER'S V-STATISTIC/ ALPHA=.05	0.40691	CRAMER'S V-STATISTIC/ ALPHA=.05	0.35506

OBJECTIVES DROPPED FROM LATER TESTS ARE NOT INCLUDED IN THE COMPUTATION OF THESE VALUES.

In this group of two tables, Table 11 is the one which has direct bearing on the answer to this research question. The critical value of the fourth grade  $\chi^2$  statistic is 11.07 at five degrees of freedom. The critical value of the seventh grade  $\chi^2$  statistic is 7.81 at three degrees of freedom. Both of these values were determined at an alpha value set at 0.05. The computed  $\chi^2$ -values shown in the table are significant at a probability less than or equal to 1 in 10,000 occurrences. Therefore, the  $\chi^2$ -statistic computed for both fourth grade and seventh grade students are far in excess of the critical  $\chi^2$  values which would define areas of rejection under the alpha criteria established in this study at 0.05, or five chances in one hundred occurrences. Therefore, Table 11 demonstrates that there is a statistically significant shift in the proportion of passable objectives in MEAP reading tests between the last year covered by this analysis, 1979, and the preceding year, back to 1973 or 1974, in which the smallest number of passable objectives occurred.

The Cramer's V statistic which corresponds to the respective fourth grade and seventh grade  $\chi^2$ -value is also presented in Table 11. Cramer's V provides a means for determining the strength of relationship measured by the  $\chi^2$  value. Cramer's V corresponding to the fourth grade  $\chi^2$  is 0.40691, and Cramer's V corresponding to the seventh grade  $\chi^2$  is 0.39506. Both values suggest a less than moderate degree of association exists between the proportion of passable objectives in the years compared.

The number of passable fourth grade objectives in the entire sample of 1000 students was 18,310 out of a possible 19,000, or 96.37%, in the 1973 test. The number of passable objectives in the 1979 test rose to

18,901, or 99.48%. The number of passable seventh grade objectives in the entire sample-of-1000 students was 19,638 out of a possible 20,000, or 98.19%, in the 1974 test. The number of passable objectives in the 1979 test rose to 20,000, or 100%.

#### UNHYPOTHESIZED FINDINGS

Table 12 reveals that the seventh grade MEAP Reading Test has a higher percentage of passable reading objectives than the fourth grade test for 1973, and for 1976 through 1979; the same percentage in 1975; and a lower percentage only one year: 1974.

TABLE 12

PROBABILITY OF PASSING TEST OBJECTIVES IN THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHERE IT IS ASSUMED THAT STUDENTS WILL ALWAYS GET DIFFICULT ITEMS WRONG WHEN THOSE ITEMS ARE SO DIFFICULT THAT THEY DO NOT FIT THE RASCH MODEL

READING TEST OBJECTIVE	FOURTH GRADE PROBABILITY OF PASSING READING OBJECTIVES: .33 WHEN 5 ITEMS FIT; .17 WHEN 4 ITEMS FIT; AND 0 WHEN 3 OR LESS ITEMS FIT							READING TEST OBJECTIVE	SEVENTH GRADE PROBABILITY OF PASSING READING OBJECTIVES: .33 WHEN 5 ITEMS FIT; .17 WHEN 4 ITEMS FIT; AND 0 WHEN 3 OR LESS ITEMS FIT						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
A	0	0	0	.17	.17	.33	.33	A	.33	.33	.33	.33	.33	.33	.33
B	0	.17	.17	.17	.33	.33	.33	B	.17	.17	0	.33	.33	.33	.33
C	.33	-	-	-	-	-	-	C	.33	-	-	-	-	-	-
D	.33	.33	.33	.33	.33	.33	.33	D	.33	-	-	-	-	-	-
E	.17	.33	.33	.17	.17	.17	.17	E	.33	.33	.33	.33	.33	.33	.33
F	.17	.17	.33	.17	.17	.17	0	F	.17	0	.17	.33	.33	.17	.33
G	.17	.17	0	.17	0	0	0	G	.33	.17	.33	.33	.33	.33	.33
H	0	.17	.17	.33	.33	.17	.33	H	.17	0	0	.33	.33	.33	.33
I	.17	.17	.33	.33	.33	.33	.33	I	.33	.33	.33	.33	.33	.33	.33
J	.33	.33	.33	.33	.33	.33	.33	J	.33	.33	.33	.33	.33	.33	.33
K	.17	.33	.33	.33	.33	.33	.33	K	.33	.33	.33	.33	.33	.33	.33
L	.33	.33	.33	.33	.33	.33	.33	L	.33	.17	.17	.17	.17	.33	.33
M	.33	.33	.33	.33	.33	.33	.33	M	.33	.33	.33	.33	.33	.33	.33
N	.33	-	-	-	-	-	-	N	.17	.33	.33	.33	.33	.33	.33
O	.33	-	-	-	-	-	-	O	.33	.33	.33	.33	.33	.33	.33
P	.17	.33	.33	.33	.33	.33	.33	P	.17	.17	.17	.33	.33	.33	.33
Q	.33	.33	.33	.17	.17	.33	.33	Q	.33	.33	.33	.33	.33	.33	.33
R	.17	.17	.33	.33	.33	.33	.33	R	.33	.33	.33	.33	.33	.33	.33
S	.33	.33	.33	.33	.33	.33	.33	S	.17	.17	.17	.33	.33	.33	.33
T	.17	.17	.17	0	.33	.33	.33	T	.33	-	-	-	-	-	-

TABLE 12 (continued)

PROBABILITY OF PASSING TEST OBJECTIVES IN THE FOURTH AND SEVENTH GRADE MEAP READING TESTS WHERE IT IS ASSUMED THAT STUDENTS WILL ALWAYS GET DIFFICULT ITEMS WRONG WHEN THOSE ITEMS ARE SO DIFFICULT THAT THEY DO NOT FIT THE RASCH MODEL

READING TEST OBJECTIVE	FOURTH GRADE PROBABILITY OF PASSING READING OBJECTIVES: .33 WHEN 5 ITEMS FIT; .17 WHEN 4 ITEMS FIT; AND 0 WHEN 3 OR LESS ITEMS FIT							READING TEST OBJECTIVE	SEVENTH GRADE PROBABILITY OF PASSING READING OBJECTIVES: .33 WHEN 5 ITEMS FIT; .17 WHEN 4 ITEMS FIT; AND 0 WHEN 3 OR LESS ITEMS FIT						
	1973	1974	1975	1976	1977	1978	1979		1973	1974	1975	1976	1977	1978	1979
U . . . . .	.33	.33	.33	.33	.33	.33	.33	U . . . . .	.33	.33	.33	.33	.33	.33	.33
V . . . . .	.33	.33	.33	.33	.33	.33	.33	V . . . . .	.17	.17	.17	.17	.17	.17	.17
W . . . . .	0	-	-	-	-	-	-	W . . . . .	.33	.17	.17	.17	.33	.33	.33
TOT.33 PROB:	11	11	14	12	14	15	16	TOT.33 PROB:	16	11	12	17	18	18	19
TOT.17 PROB:	8	7	3	6	4	3	1	TOT.17 PROB:	7	7	6	3	2	2	1
TOT.00 PROB:	4	1	2	1	1	1	2	TOT.00 PROB:	0	2	2	0	0	0	0
TOT DROPPED:	0	4	4	4	4	4	4	TOT DROPPED:	0	3	3	3	3	3	3
% PASSABLE:	87%	95%	90%	95%	95%	95%	90%	% PASSABLE:	100%	90%	90%	100%	100%	100%	100%

Table 12 reveals that the seventh grade test has demonstrated a fairly consistent tendency for a greater proportion of passable objectives than the fourth grade test during most of the seven tests studied. All of the test objectives were passable in the last four years the seventh grade test was given. At no time did the percentage of seventh grade test objectives fall below 90%. The percentage of passable test objectives in the fourth grade test never reached 100% during the seven year period covered by this study, but four years out of the seven, 1974, 1976, 1977, and 1978, 95% of the test objectives were passable. In 1973 the percentage of passable objectives was 87%, and for 1975 and 1979, the percentage was 90%.

Question 5: Do changes occur in the proportion of students who are "qualified" for remedial reading instruction between scores reported on MEAP Reading Test objectives compared to the proportion of students who would be qualified if scores were based solely on items which have been re-scored to compensate for the adverse affect perceived in this study by the method described? That is, does a change in proportion of qualified students occur when students are credited for too-difficult items they have missed?

#### STATISTICAL ANALYSIS

One table was developed in the process of seeking an answer to this research question: Table 13.



Table 13 presents information on the change in proportion of students qualified for remedial reading instruction in 1979 before and after adjusting the learning objective scores that year to compensate for items that are so difficult that they do not fit the Rasch model.

#### HYPOTHESIZED FINDINGS

##### Hypothesis H:0<sub>5</sub>:

There is no statistically significant decrease in the proportion of students passing less than 40% of the learning objectives in the MEAP Reading Test administered in 1979 between objective scores reported that year and the objective scores after the 1979 items which do not fit the Rasch model, because they are too difficult, have been re-scored and credited to students who get these items wrong at a probability level of 0.05, or less.

Table 13 indicates this hypothesis must be accepted.

Table 13 has direct bearing on the answer to this research question. This table presents a comparison of the proportion of students scoring below 40% on objectives before and after all items judged to be too difficult under criteria established in this study are re-scored in favor of the student. This is the group which is qualified for remedial reading instruction. Both the fourth grade and the seventh grade results are shown side by side to highlight the differences and similarities which this procedure has on these two entirely independent samples.

Only one  $\chi^2$  statistic, for the fourth grade is computed because only the objective scores of the fourth grade group changed on completion of the re-scoring process. Before re-scoring, 136 fourth

grade students passed less than 40% of the MEAP Reading Test objectives for the 1979 test. By crediting students with items that do not fit the Rasch model, which they got wrong, more students would be expected to pass more objectives with the result that fewer would fall in the remedial group. In the case of the fourth grade students, these expectations were met. The number in the 1979 sample qualified for remedial reading instruction dropped from 136 to 129. However, this change was not significant! For the seventh grade students, there was no change whatever following the re-scoring process. There were 129 students qualified for remedial instruction in this sample before and after the re-scoring process. Since there was no change in the proportion of seventh grade students qualified for remedial instruction, the computed  $\chi^2$  value was zero.

Cramer's V was not computed for data in Table 13 because this statistic is not appropriate to a 2 x 2 comparison as is the case here. Cramer's V is applicable only to comparisons involving three or more variables in either, or both, the row or column dimension.

#### UNHYPOTHEZIZED FINDINGS

The occurrence of hard items that do not fit the Rasch model has declined from 1973 through 1979. As the proportion of these items drops, the incidence of passable objectives can be expected to increase. It appears that the factors prompting the decline in hard items that do not fit the Rasch model have all but eliminated such items in the 1979 test. There were only three in the 1979 fourth grade test and none in the 1979 seventh grade test. For comparison, it is interesting to note that there were 17 items that did not fit the Rasch model in the 1973 fourth grade test and 7 in the 1973 seventh grade test.

## SUMMARY

Tables 1 through 4 are largely descriptive. That is, they provide information on certain aspects of individual score categories and item difficulty for the fourteen tests covered by this investigation. Certainly no clear conclusions may be drawn from the data which they contain. None is intended, but they give a feeling for the data which suggests that there may be shifts in the measurement potential of these tests over the years which, though perceptible perhaps, would be difficult to interpret without some form of inferential analysis. The shifts in meaning may exist in these figures, but they are subtle at best and require further statistical analysis in probability terms.

Tables 5 through 8 are, again for the most part, largely descriptive in that they are intended to provide descriptive information on certain aspects of individual score categories and item-difficulty. However, Table 5 does present, with one exception, t-values for the fourth grade and for the seventh grade in each of the seven years spanned in this study that are larger than any critical t-value which could be anticipated merely by chance. For 1002 cases, at an alpha level of 0.05, the critical t-value varies from year to year from 1.71 to 1.73. With the exception of the 1979 seventh grade sample, all of the computed t-values are larger. Since there were no items which were too difficult to fit the Rasch model in the 1979 seventh grade test, there could be no difference in average reading objective scores before and after item re-scoring in this instance, and there was none. Hence Table 5 portrays a statistically significant change in every test having items judged to be too difficult under criteria established in this study as a result of the adjustment process used to correct the effect of these

items. However, while the adjustment procedure adopted in this investigation to eliminate the influence of overly difficult test items produced statistically significant results which were in the direction anticipated, the effect appears to be minimal and in rather steady decline with each successive test. Table 6 shows these results. Table 7 and Table 8 tend to provide further confirmation. Therefore it seems apparent that there may be other factors at work which have far greater influence on average objective scores than the presence of items which are too difficult to fit the Rasch model. It can be demonstrated, in fact, that the incidence of items judged to be too difficult to fit the Rasch model declined significantly in both fourth grade and seventh grade MEAP Reading Test results with the passage of time.

Table 9 demonstrates that there is a statistically significant decrease in the number of difficult items, that is hard items that do not fit the Rasch model, between the last year the MEAP Reading Test was evaluated in this study and the preceding year having the most such items: 1973 for fourth grade and 1974 for the seventh grade.

The results presented in Table 10 suggest that more items fit the Rasch model in the seventh grade reading test overall while both fourth grade and seventh grade tests started out with comparable levels of difficulty. Overall the percentage of items that fit the Rasch model, or if they were so easy that they did not fit the model but enjoyed the presumption of validity, ran better than 93% for the period of this study. In the 1979 seventh grade test, only one item did not fit the Rasch model.

Table 11 reveals a statistically significant increase in the number of passable learning objectives between the last year the MEAP Reading

Test was evaluated in this study, 1979, and the preceding year having the least number of passable objectives: 1973 for the fourth grade and 1974 for the seventh grade. A moderate correlation between the year of the test and the number of passable objectives is demonstrated.

The results presented in Table 12 suggest that objectives were easier to pass, and the results more consistent, for the seventh grade students than the fourth graders. Overall the percentage of passable reading test objectives has been high for both grades; above 95% in most cases.

Table 13 supports the conclusion that there was no statistically significant increase in the number of passable learning objectives in the last year the MEAP Reading Test was evaluated in this study, 1979, after hard items that did not fit the Rasch model are re-scored. There was only one of these items in the seventh grade test and it had no impact on the number of students in the sample who passed 40%, or more, of the requisite learning objectives. Only seven of the 136 fourth grade students were affected by the re-scoring procedure, and this did not constitute a significant number.

While it seems clear that items which are so difficult that they do not fit the Rasch model have materially hampered student ability to pass MEAP Reading Test learning objectives in past tests, they had no such effect in the 1979 test. The incidence of hard items has apparently dropped off in MEAP reading tests to such a degree in 1973 to 1979 that they can no longer be considered cause for learning objective scores below 40% at the conclusion of this period of time.

## CHAPTER V

### CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

#### CONCLUSIONS

The statistical analysis presented in the preceding chapter demonstrates that the Rasch model may be used to evaluate learning objectives, measured by multiple test items, in large criterion referenced tests.

The first research question deals with the determination of the number of difficult items in 1973 through 1979 MEAP reading tests given to fourth grade and seventh grade students. At no time is the number of these items very large - never exceeding 10% in any specific year. The incidence of these very hard items is greatest in the early years of this series of tests, and an overall decline is apparent in both grades with the passage of time. Since the items which were evaluated in this study did not change in content over time (only order of presentation in each test changed) it seems likely that forces were at work which tended to improve the fit of these items. And since all MEAP items are, in this study, presumed valid in MEAP tests, lack of fit to the Rasch model should not be charged to the possibility that these very difficult items do not really measure the underlying trait sought to be measured by the test as a whole. Such a conclusion might follow if the items were not consistent in content, but this is not the case. Furthermore, because these are objective tests, the presumption must be that all MEAP test items should fit the Rasch model and that any tendency not to do so most likely is the result of factors unrelated to the items themselves. One of the most likely among such factors is insufficient instruction to prepare the student to answer related items. Since one of the primary

objectives of the MEAP testing program is to focus learning on objective performance which falls short of expectations, it is safe to say that this was done in successive years. Certainly there were many school districts where this actually did happen, and the decline in items which did not fit the Rasch model over time suggests that these efforts may have been worthwhile.

The second research question deals with the issue of whether the measurement efficiency of MEAP reading tests improves after hard items that do not fit the Rasch model are credited to the student who gets these items wrong. The question attempts to explore the effect which items may have upon the probability of passing MEAP learning objectives. The presence of items that do not fit the Rasch model does not necessarily preclude a student from passing an objective. There are relatively few such items and they may be distributed in such a way as to have little or no affect on passing four out of the five items designed to measure each objective. For example, the presence of one of these very difficult items would not affect the score on an objective for those students who already had four of the items right. However, a statistically significant change in the number of items passed does occur after hard items that do not fit the Rasch model have been re-scored. Therefore, it seems likely that the presence of hard items that do not fit the Rasch model do work to the disadvantage of students in terms of the number of objectives passed. This relationship holds consistently for all but one of the fourteen tests: the 1979 seventh grade test. These results suggest that the incidence of very hard items is detrimental, despite their decline in occurrence, throughout the series of tests. The only case where a significant change did not occur

was the single test in which there was only one hard item which did not fit the Rasch model--the 1979 seventh grade test.

This analysis has demonstrated that the presence of two or more items, for a given objective, that do not fit the Rasch model work to the disadvantage of students in that such items materially affect their ability to pass the learning objectives being measures.

The third research question deals with the possibility that a directional trend may be evident in item-difficulty over the period covered in this investigation. There is an apparent decline in the number of hard items that do not fit the Rasch model, over the years, which suggests that this may be the case. Overall, more and more items, which were once too difficult, have fit the Rasch model with the passage of time. The statistical analysis associated with this question leaves little doubt that, over the seven year interval, there has been a significant decline in the level of difficulty associated with a number of hard items that did not fit the Rasch model in the earlier years covered in this investigation. Since the presence of items which are too difficult to fit the Rasch model was shown to be materially disruptive to students' chances for passing MEAP reading objectives, it seems reasonable to conclude that the decline in the level of difficulty associated with such items did improve students' chances to a significant degree. As noted in the preceding discussion of question 2, it is apparent that as few as two very hard items that did not fit the Rasch model have a detrimental affect on prospects for passing a MEAP reading objectives. Though not measured directly, an opposite impact should probably be credited to those items which are no longer so difficult that they do not fit the Rasch model. It seems reasonable to



conclude that a decline in the level of difficulty formerly associated with very hard items would have a significant, positive effect on objectives passed.

A statistically significant decline in the difficulty level of items that were at one time too difficult to fit the Rasch model is presented in Table 9. Table 10 reveals that there were more of these items in the MEAP reading tests given in the first three years covered in this investigation compared to the last three years. As previously indicated, the items themselves did not change, only their apparent level of difficulty changed. An important objective of these tests is to focus instructional effort on those areas where objectives have not been met, and it seems reasonable to conclude from these data that this kind of corrective action was taken. Increased instructional emphasis is probably responsible for much of the decline in the rating of items considered in previous years to be so difficult that they did not fit the Rasch model.

The fourth research question deals with the effect items that are too hard to fit the Rasch model have on the probability for passing learning objectives between the year in which the smallest proportion of learning objectives were passed, and 1979, the last year in the analysis. There is a statistically significant increase in the proportion of objectives passed between the worst year and the last year for both fourth grade and seventh grade students. Since probability is, by definition, the proportion of possible occurrences of a particular event to all possible events, the statistically significant increase in the proportion of passed learning objectives denotes a significant increase in the probability that a MEAP learning objective will be

passed. The decline in the level of difficulty, which results in a corresponding decrease in the number of very difficult items, tends to be linear. That is, the greatest number of such items occurs toward the earliest years of the 1973 to 1979 interval, and a more or less steady drop in difficulty level for these items appears to occur throughout the time period. The fourth grade data demonstrated this tendency from 1973 to 1979, where most of the very hard items which did not fit the Rasch model in 1973 did so in 1979. The seventh grade data very nearly demonstrated the same tendency, but the year 1974 had the majority of these items in this case rather than 1973.

The fifth research question deals with the effect hard items which do not fit the Rasch model have on the proportion of students passing less than 40% of the MEAP reading objectives. The statistical analysis was intended to measure change in this proportion for the 1979 test. In fact a statistically significant change did not occur as anticipated at the outset of the analysis. In so far as the 1979 data is concerned, recoding these very hard items resulted in an insignificant change in the proportion of both the fourth grade and the seventh grade students who passed less than 40% of the learning objectives. There were only three of these items in the fourth grade data and only one in the seventh grade data. The re-scoring process resulted in a net decrease of only seven fourth grade students that passed less than 40% of the test's learning objectives and no change at all in the number of seventh grade students passing less than 40% of these learning objectives. The conclusion to be drawn as a result of this analysis is that the 1979 data had too few hard items to make any difference. Since there were only a few very difficult items in the 1979 test, nothing has been

gained toward understanding their impact upon the proportion of students in need of remedial instruction.

#### SUMMARY OF CONCLUSIONS

The presence of items that do not fit the Rasch model was anticipated in MEAP reading tests when this study was undertaken. The data have revealed that such items indeed were in the 1973 through 1979 tests as expected. The incidence of these items was expected, at the outset of the study, to remain more or less constant throughout the period investigated. However, contrary to expectations, this was not the case. Probably it may have been erroneous to assume that the level of difficulty for items encountered in MEAP tests which did not fit the Rasch model would, from the outset, remain constant. An important objective of the Michigan Educational Assessment Program is to focus additional attention on areas where learning achievement falls short of expectations. If this is done properly, the incidence of items that do not fit the Rasch model is likely to decline. If such items belong in these tests from the outset because they measure reasonable, though not yet attained learning objectives, then it is irrelevant that they do not fit the model at an earlier stage. Lack of fit to the Rasch model is an important consideration in the initial stages of designing a criterion referenced test. However, valid questions may be retained in such tests despite their lack of fit to the Rasch model.

Rasch measurement theory certainly implies that the proportion of items that do not fit the Rasch model should remain constant through repeated administrations of the same items and that the appropriate procedure is to replace them by calibrated items--items that fit the model. The initial presumption is that an item which does not fit the

model may measure something other than the underlying trait which the overall test is intended to measure. These items become candidates for rejection. They should be dropped if lack of fit is significant, and no rationale can be found in the circumstances of the test to explain that lack of fit. However, there is a controverting assumption used in this study that all MEAP test items are valid. This assumption forces a degree of tolerance of items which do not fit the model which the underlying theory possibly does not anticipate. Surely this is the case if the assumption of content validity is accorded all MEAP items. Once this assumption has been accepted in connection with criterion referenced tests, then it would follow that all items, even items that do not fit the Rasch model, must be retained without question. The underlying implications of all of these considerations are that every item in these objective tests measures the underlying trait and that lack of item fit probably means that persons taking the test are inadequately prepared. This is the major conclusion to be drawn from this investigation. Viewed from this perspective, Rasch measurement becomes a means for determining the effectiveness of remedial instruction programs over time.

Within such a context, Rasch measurement takes on a dual role when applied to objective tests. First, it may be employed to advantage in determining item fit in the initial stages of test development in reaching decisions on whether to use individual items. Its use, at this stage, would augment established procedures of objective test design. Second, if items that do not fit the Rasch model are allowed to remain in a test after item calibration, the assumption is that the test taker should know the item. It is irrelevant that the item is too difficult.

In these instances, Rasch measurement should be applied to determine the impact of programs designed to improve test performance over time. Both applications of the Rasch model are most appropriate, and desirable, in connection with objective testing efforts like the Michigan Educational Assessment Program. The first application focuses on items which may not measure the underlying trait intended by the test under development. Once it is determined that these items do measure the underlying trait intended, and the test taker ought to know the answers despite present difficulty, they may be retained in an objective test. The second application may then be undertaken as an active part of the test program, with greater confidence, as a measure of improvement or declining performance on a set of consistent objective items over time.

Therefore, the assumption used initially in this study that very difficult items do not belong in a criterion referenced test must be qualified. Such items do not belong when those items have not been carefully analyzed to determine content validity. Lacking this, such items unduly penalize the student. However valid, difficult items that do not fit the Rasch model do belong in a criterion referenced test. In such a case, use of the Rasch model shifts from a determination of possible inequity to a determination of performance improvement.

#### RECOMMENDATIONS FOR FUTURE RESEARCH

Rasch theory seems to promise objective measurement, but problems of interpretation associated with the decision that an item does, or does not, fit the Rasch model could withhold that promise. The issues encountered during attempts to define and use the item fit statistic in this research raised some important questions about the objectivity of Rasch measurement which are fundamental to the practical application of

this tool in test measurement. While Rasch measurement theory describes an objective test measurement tool, Rasch measurement application may entail too many subjective elements to make this possible in a practical sense. Interpretation of the item fit statistic has changed over the years. Conviction as to the appropriateness of a statistic in determining item fit to the Rasch model, independent of subjective considerations, seems to have softened over the years. Objectivity in test measurement may yet be possible, but the tendency seems to be growing to add subjective elements of interpretation to the use of an item fit statistic, or at least increase the complexity of using the statistic, so that the result is a potentially impractical measure. Further study of the need to augment Rasch measurement; the action required in doing so; and the circumstances under which such action may be required seems appropriate. The objective of this investigation would be to identify and evaluate modifications applied to the interpretation of Rasch theory in practice. An evaluation of the apparent increased use of subjective elements in the interpretation of the fit statistic in practice could be an important result from a study of this kind.

Another study might be done on a comparison of the results produced by BICAL and BICAL.3. The impact of the weighting procedures used in the current BICAL program, BICAL.3, poses some important questions. The current computer program employs a weighting algorithm which is designed to offset the effect of an unexpected response in extremely wide tests (i.e., tests having difficulty estimates greater than four logits in width). Some college entrance qualifying examinations, for example, are up to 12 logits in width. In such tests, unexpected responses have an

enormous impact on an individual's total fit statistic. The t-statistic produced by BICAL.3 is weighted in such a way as to compensate for the extreme impact of such items in these tests. However, since the average test in education is three or four logits in width. The weighting factor now used in the most recent version of BICAL may not be appropriate to these narrower tests. This possibility was the major reason for use of the earlier BICAL program rather than BICAL.3 in this investigation. BICAL may have less tendency to mask bad fit in narrow tests than BICAL.3. The objective of this proposed study might be to develop guidelines in the application of different fit statistic algorithms to tests of varying width; or, possibly, determination that the current weighting method used in BICAL.3 is indeed appropriate to tests of any width.

While the item fit statistic employed in this investigation appears to be enough like an F-statistic to be interpreted as if it were one, it is not truly an F-statistic. The Rasch measurement item fit statistic is probably unique, and though it may demonstrate distribution characteristics similar to one or more established statistical measures, its probability distributions should be established. There are likely to be a number of such distribution families with properties that vary according to the number of score groups in the analysis and the number of degrees of freedom involved. It would be a major undertaking to establish such distributions. Nevertheless, it is a legitimate question at this point in time to ask: What are the theoretical limitations which should be placed upon the Rasch fit statistic? Until this statistic becomes more clearly defined, prospective users of Rasch measurement are likely to be reluctant to proceed. Currently the most

knowledgeable proponents of Rasch measurement, Wright and his associates, urge caution in using a purely statistical interpretation of item fit analysis results. The prospective user may well ask "What circumstances warrant the use of a Rasch fit statistic at all?" The objective of this prospective research could advance understanding of the item fit statistic in Rasch measurement and promote its more general use or the invention of a more appropriate procedure.

Rasch measurement seems to offer a pragmatic means for objective evaluation of item-difficulty and person-ability. In early attempts to implement the Rasch model "in the field," a manual method for using the model on small tests was brought to the attention of prospective users. In more recent years, however, the emphasis has been directed almost entirely toward computerized analysis. At least this seems to be the area in which most of the developmental work is occurring. There is a real possibility that the Rasch measurement model may not, in fact, be easy to use. It does not help matters much to insist that the Rasch model is easy to use when the preponderance of documented illustrations emphasize the need for computer resources and considerable understanding of a series of subjective side issues which could affect its interpretation. There appears to be a considerable void in relatively simple and straight forward applications of the model which would serve to encourage its use. Future research into such applications would determine if the Rasch model is really a practical device or just a hope.



**APPENDICES**

**APPENDIX A**  
**DATA REQUEST AND ASSURANCES AGREEMENT**

Edward Roeber, Supervisor  
Michigan Educational Assessment Program  
Box 30008  
Michigan Department of Education  
Lansing, Michigan 48909

Tuesday  
July 8, 1980  
Detroit

Dear Dr. Roeber;

At last I am in a position to begin my research. My proposal was approved by my advisory committee three weeks ago. A copy of the proposal summary is enclosed for your information.

Also enclosed is a letter from my primary advisor, Dr. Donald Marcotte, which he was kind enough to write on my behalf. He understands, as I am sure you do also, that people in my position, attempting to complete a dissertation, need all the help, encouragement, and patience they can get.

Finally, I have completed and enclosed the "Data Request and Assurances Agreement" form you gave me so many months ago. I am requesting the use of the "Pupil Sample File" for the seven years 1973 through 1979 inclusive. I am making tentative arrangements to secure three tapes onto which the data can be copied following approval of my request. I understand that the data will be made available to me, through your good offices, at no cost as the result of the interest your office has had in working with graduate students at Wayne's College of Education. I am deeply appreciative that such an opportunity might be made available to me.

I hope that you will approve this request at your earliest convenience. I want to assure you that use of the data made available to me as a result will be confined to my dissertation. I am a friend of the Michigan Educational Assessment Program and hope that the Program too might in some way benefit from my efforts.

Please tell me what I must now do to actually obtain the data which I am requesting here. As always, I am prepared to come to Lansing on very short notice if such a step would expedite matters. I really look forward to hearing from you. I'm most anxious to get started. Thank you, again, for your time and attention.

Sincerely,



Donald J. McPherson  
Doctoral Candidate, Educational Evaluation and Research

DM/dm

RA-2000-A  
**MICHIGAN DEPARTMENT OF EDUCATION**  
**DATA REQUEST AND ASSURANCES AGREEMENT**

Return TWO copies to Michigan Educational Assessment Program, Box 30008, Lansing, MI 48909.

**DATA DESCRIPTION AND COST**

NEAP files are available on magnetic computer tape which is 9-track, 1600 bpi. The following files are available: District, School, Pupil (Anonymous) and Pupil Sample of 5000. The minimum cost of each file is \$150.00 plus postage; the anonymous pupil file costs an additional \$0.04 per pupil selected. The cost is payable directly to Westinghouse Learning Corporation upon receipt of their invoice. Each file selected will be accompanied by a tape format. A more complete description of each file is given on the back of this form. Please indicate file selection(s) below.

DISTRICT FILE	SCHOOL FILE	PUPIL FILE	PUPIL SAMPLE FILE
<input type="checkbox"/> 1973-74	<input type="checkbox"/> 1973-74	<input type="checkbox"/> SELECTED PUPILS	X 1973, '74, '75, and X 1976-77
<input type="checkbox"/> 1974-75	<input type="checkbox"/> 1974-75	<input type="checkbox"/> ALL PUPILS	X 1977-78 X 1978-79
<input type="checkbox"/> 1975-76	<input type="checkbox"/> 1975-76	<input type="checkbox"/> 1973-74	OTHER
<input type="checkbox"/> 1976-77	<input type="checkbox"/> 1976-77	<input type="checkbox"/> 1974-75	<input type="checkbox"/>
<input type="checkbox"/> 1977-78	<input type="checkbox"/> 1977-78	<input type="checkbox"/> 1975-76	
		<input type="checkbox"/> 1976-77	
		<input type="checkbox"/> 1977-78	

**ASSURANCES**

The following assurances are given to the Michigan Department of Education in return for access to data for educational research purposes:

- The data supplied will be used exclusively under the direction of the researcher whose name appears below, and will not be supplied to any other individual, agency, or organization.
- No school or school district, nor any individual staff member of any school or school district will be identified in any report of the research conducted with these data.
- The expense of obtaining a copy of the required assessment data will be borne by the researcher.
- The researcher will supply at least one copy of all completed research reports based upon these data to the Director of the Research, Evaluation and Assessment Service, Michigan Department of Education.

I certify the above assurances will be followed while using data provided by the Michigan Department of Education.

NAME Donald J. Mc Pherson		POSITION OR TITLE Doctoral Candidate, Wayne State U., Detroit, MI	
STREET ADDRESS 25 E. Palmer, Apt. #52		CITY Detroit	STATE AND ZIP CODE Michigan 48202
SIGNATURE <i>Donald J. McPherson</i>		DATE July 8, 1980	
<b>APPROVAL</b>			
This Data Request and Assurances is approved.			
NAME		SIGNATURE	
POSITION		DATE	

\*DO NOT USE ONLY\*

APPENDIX B  
FOURTH GRADE AND SEVENTH GRADE  
INDIVIDUAL STUDENT REPORT FORM  
FOR 1973-74 MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM  
AND CORRESPONDING LISTS OF ITEMS MEASURING EACH OBJECTIVE

PUPIL NAME:

SCHOOL:

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM 1973-74 (YEAR 5)

GRADE 4

TEACHER:

DISTRICT:

M A T H E M A T I C S	OBJECTIVE NO.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35		
	OBJECTIVE DESCRIPTION	[Diagonal text for objectives 1-35]																																				
OBJECTIVE ATTAINED!																																						
ITEM NUMBERS AND RESPONSES BY OBJECTIVE	96	91	116	106	101	126	131	161	36	71	56	171	66	11	11	11	61	46	51	76	136	21	1	146	6	141	26	31	151	156	111	176	166	86	16			
	97	92	117	107	102	127	132	162	37	72	57	172	67	12	12	12	62	47	52	77	137	22	2	147	7	142	27	32	152	157	112	177	167	87	17			
	98	93	118	108	103	128	133	163	38	73	58	173	68	13	13	13	63	48	53	78	138	23	3	148	8	143	28	33	153	158	113	178	168	88	18			
	99	94	119	109	104	129	134	164	39	74	59	174	69	14	14	14	64	49	54	79	139	24	4	149	9	144	29	34	154	159	114	179	169	89	19			
100	95	120	110	105	130	135	165	40	75	60	175	70	15	15	15	65	50	55	80	140	25	5	150	10	145	30	35	155	160	115	180	170	90	20				
NUMBER CORRECT																																						

R E A D I N G	OBJECTIVE NO.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
	OBJECTIVE DESCRIPTION	[Diagonal text for objectives 1-23]																						
OBJECTIVE ATTAINED!																								
ITEM NUMBERS AND RESPONSES BY OBJECTIVE	211	219	181	264	182	11	195	211	192	212	188	115	214	113	183	117	269	193	189	217	206	191	218	
	221	240	252	265	183	12	196	232	236	278	211	218	221	240	187	274	270	216	219	207	207	194	222	
	249	241	259	266	184	13	197	233	237	250	226	260	247	248	200	141	271	275	246	203	208	229	223	
	261	242	275	267	185	14	198	234	253	278	245	278	254	256	244	258	272	279	267	204	209	255	257	
286	243	282	248	186	15	199	235	271	281	274	283	260	265	263	287	273	284	290	275	210	288	289		
NUMBER CORRECT																								

INDIVIDUAL STUDENT REPORT FORM

WORD RELATIONSHIPS

RAW SCORE: STANDARD SCORE: PERCENT BELOW:

DATE OF TESTING: CHRONOLOGICAL AGE:

COMMENTS:

1979-80

**LIST OF ITEMS MEASURING EACH FOURTH GRADE OBJECTIVE**

<b>Reading</b>		<b>Mathematics</b>			
<b>Objective Number</b>	<b>Item Number</b>	<b>Objective Number</b>	<b>Item Number</b>	<b>Objective Number</b>	<b>Item Number</b>
1	45,52,73,81,92	1	196-200	18	171-175
2	83-87	2	101-105	19	211-215
3	65-69	3	241-245	20	251-255
4	16-20	4	231-235	21	106-110
5	6-10	5	226-230	22	161-165
6	27-31	6	136-140	23	1-5
7	35-39	7	176-180	24	206-210
8	24,32,33,76,98	8	246-250	25	126-130
9	41,53,74,89,97	9	111-115	26	201-205
10	21,40,51,70,96	10	166-170	27	141-145
11	34,43,80,90,99	11	116-120	28	186-190
12	42,48,72,77,88	12	156-160	29	216-220
13	47,49,75,79,93	13	151-155	30	221-225
14	11-15	14	146-150	31	256-260
15	23,44,50,91,100	15	236-240	32	181-185
16	22,46,71,82,95	16	191-195	33	131-135
17	55-59	17	121-125		
18	60-64				
19	25,26,54,78,94				





79/80

**LIST OF ITEMS MEASURING EACH SEVENTH GRADE OBJECTIVE**

<b>Reading</b>		<b>Mathematics</b>			
<b>Objective Number</b>	<b>Item Number</b>	<b>Objective Number</b>	<b>Item Number</b>	<b>Objective Number</b>	<b>Item Number</b>
1	30,79,80,90,92	1	106-110	24	251-255
2	70-74	2	1-5	25	256-260
3	101-105	3	126-130	26	176-180
4	6-10	4	131-135	27	271-275
5	24-28	5	111-115	28	306-310
6	46-50	6	136-140	29	311-315
7	15,31,35,54,83	7	141-145	30	276-280
8	13,51,66,81,97	8	161-165	31	281-285
9	59,60,61,68,93	9	156-160	32	121-125
10	12,32,33,55,76	10	166-170	33	261-265
11	84-88	11	196-200	34	301-305
12	63,64,78,89,91	12	181-185	35	286-290
13	11,29,53,62,94	13	186-190	36	266-270
14	14,52,65,98,99	14	291-295	37	191-195
15	17,57,67,75,96	15	201-205	38	316-320
16	16,34,58,77,100	16	216-220	39	146-150
17	18,56,69,82,95	17	226-230	40	171-175
18	36-40	18	236-240	41	116-120
19	19-23	19	241-245	42	151-155
20	41-45	20	231-235	43	206-210
		21	246-250	44	296-300
		22	221-225	45	321-325
		23	211-215		

79/80

## APPENDIX C

### THE ITEM FIT STATISTIC - EVOLUTIONARY CHANGES OF INTERPRETATION

#### Introduction

The fit statistic employed in this study is based upon the FIT MN. SQ (i.e., the fit mean square). The FIT MN. SQ is a measure of variance between the expected performance on an item, indicated by the Rasch model, and actual performance observed on an item. It is computed in a manner quite similar to that used to compute variance, except that each item score is subtracted from an "expected" score on the item, which is determined by the model, rather than from the overall score mean.

The interpretation given the FIT MN. SQ in this study hinges on expected values concepts. Technically, the FIT MN. SQ values constitute, in themselves, only the between group variance component of the F-ratio:

$$s_b^2 / s_w^2$$

where subscript b identifies between group variance and subscript w identifies within group variance. How, then, is it possible to interpret the numerator of this equation (i.e., the FIT MN. SQ) directly as an F-statistic? Is not there a value in the denominator which must be considered in this F-ratio? The answer to both of these questions lies in the fact that the denominator indeed does have a value. Its expected value is one. Remember that the FIT MN. SQ is really based on the standardized residual for each ability subgroup: the FIT Z-SQUARED value for each subgroup. Since it is a standardized value, the FIT Z-SQUARED has an expected mean of zero; an expected standard deviation of one; and an expected variance of one. Consequently, the value of the denominator in the F-ratio in this instance is always expected to be one. Because this particular F-ratio is always derived as a ratio of observed between group variance, whatever it may be, to the expected within group variance, which is always one, the denominator of this F-statistic is always one. Therefore, the FIT MN. SQ appears to be an F-statistic:

$$F_{calc} = \text{FIT MN.SQ}/1$$

### Is the Fit Statistic an F-statistic?

In a word - no. While it is derived in a way that is quite similar to that used for computing the ratio of between sum of squares to within sum of squares in analysis of variance, this fit statistic does not conform to the F distribution exactly. Reference should be made at this point to information obtained on this point in a meeting with Mr Richard Smith, Director of Testing, Mercer County Community College in Trenton, New Jersey and Ronald Mead, Assistant Director of MESA, at the University of Chicago the evening of June 6, 1981. The occasion was an informal evening session after the first day of a three day seminar on Rasch measurement at the College of Education, the University of Chicago. The substance of this meeting is referred to in Chapter III. It lasted nearly two hours and dealt entirely with the subject of the "item fit statistic" used in Rasch measurement and its interpretation. At one point in the discussion, both Smith and Mead gave assurances that the fit statistic used in Rasch measurement was close enough to the behavior of an F-statistic that F-tables could be used to interpret the statistic. However, when Smith was pressed to identify the statistic - to give it a name, he preferred just to call it a fit statistic.

. . . we can demonstrate the distribution of this statistic, and the distribution is fine. . .but it has to stand on its own. It can not develop out of the argument that it looks like (some statistic we are familiar with). . . these names (i.e.,  $X^2$ , F, t) lead people to believe this statistic is really those things and they're not. (Smith, 1981).

Comparison to more familiar statistics has been employed over the years to aid in understanding this new fit statistic. At times the distinction between a comparison and the real thing has become lost in the discussion of underlying theory.

Smith indicated that for a time, Wright concentrated on the fit  $z$ -squared. Wright tried to interpret it as a  $X^2$  with one degree of freedom. Most of the discussion in MESA Research Memorandum 18 (Wright & Mead, 1975) and Research Memorandum 23 (Wright & Mead, 1977c) on fit statistic interpretation treated the fit statistic as a  $X^2$ -like statistic based on the  $Z^2$ . However, there has been a shift over the years from emphasis on a  $X^2$  interpretation; to an F-statistic interpretation; to, most recently, a t-statistic interpretation. With each successive interpretation, a "weighting factor" has been introduced which modifies the value of the between fit mean square to compensate for some shortcoming which experience had revealed in successive interpretations of the fit statistic. There has also been a subtle change in conviction. Wright's thinking has changed, apparently, from a firm conviction in the sufficiency of Rasch fit analysis in terms of clear statistical probability to a more cautious view calling for the addition of further analysis and personal judgement on the psychometrician's part. A comparison of the following passages from MESA publications shows these changes of faith in a fit statistic. The first passage is from Research Memorandum 18, published in March of 1975.

The third section of the table contains  $Z^2$  statistics for testing the fit of each item in each score group. They are approximately distributed as chi-square statistics with one degree of freedom. The first column on the right "FIT MN SQ" contains a statistic for testing the fit of each item over all groups. Since the deviations for the model were standardized in computing the  $Z^2$  statistics, the mean squares have expected values of one, and can be evaluated as F-ratios. . . (Wright & Mead, p. 18, 1975)

There was no suggestion in this publication that the F-statistic interpretation given FIT MN SQ values output by the BICAL program needed any supplemental interpretation. This conviction in the adequacy of the fit statistic was reflected in Research Memorandum 23, published in January of 1977.

A primary benefit from having an explicit mathematical model for a process is the possibility of making rigorous tests of how well the observed data are predicted by the model. . . (Wright & Mead, p. 37, 1977c)

By implication, the Rasch model affords the possibility of making such "rigorous tests". However, the same publication introduces a note of caution respecting the interpretation of "a fit statistic".

All the test fit statistics presented in this section have the appearance of chi-square (or mean square) variables, but recent simulation studies (Mead, 1976) show that this distribution is not exactly correct. Hence, exact probability statistics about lack of fit are not possible. The chi-square distribution is a useful background against which to judge these statistics, however. (Wright & Mead, p. 42, 1977c)

So it would appear that the emphasis placed on the  $X^2$  interpretation of the fit statistic warranted some caution. The interpretation of this caution was left to the reader's judgement in this case, but subsequent material published by MESA revealed that a new approach to the fit statistic was in order. The concept of the "Between Fit t" statistic was cautiously introduced in Research Memorandum 23,c, published in June of 1980.

In the search of measurable variables, tentative estimation of item difficulties is only the first. In order for these estimates to be useful as item calibrations, it must be established that it is reasonable to treat the items in question as members of the same measuring class. If that is found to be reasonable, then the measurement of persons based on calibration of these items can proceed. If not, then the available data must be reconsidered to see if there are any subsets of items that may possibly belong to a single common measuring class. If that search fails, then no "Measurement" will be possible with these items.

The most natural fit statistic in BICAL, labelled "Between Fit t" is derived directly from the "sample-free" requirements of the model. (Wright & Mead, p. 10, 1980)

While the label is again different, this t-statistic is based principally upon the fit mean square concept. The evaluation of the fit statistic to its present form as a "t-like" statistic is interesting. Though no great measure of importance is placed in this study upon the current version of BICAL (i.e., BICAL.3) that is discussed in Memorandum 23.c, and the t-statistic that it produces, brief attention is given to the subject in the subsequent sections of this appendix on the reasons for using BICAL rather than BICAL.3 to determine item fit. For the present, however, the only purpose is to track the development of the fit statistic to its present state and comment where that development has bearing on this research. The new t-statistic has been introduced to MESA publications without fanfare. One who was not familiar with prior MESA publications on Rasch analysis might not appreciate that it is not an entirely new statistic. But, like its predecessors, it too is based on the standardized residual. It seems, now, that the Rasch fit statistic is in a state of flux. One of the concluding statements in the section on fit analysis in this publication points clearly to this possibility.

It is unrealistic to expect the results of simulations to match the ideal consequences exactly, but one can ask, "To what extent do the results, and hence the algorithm they document, approximate these ideals?"

This is an important question, as the ideals are the frame of reference from which an experimenter must judge the fit of any real data. . . (Wright & Mead, p. 51, 1980)

. . . As sample or test spread out beyond typical values, variations among the item mean squares for data simulated to fit the model increases to twice that expected by the model. At the same time, the total mean square falls slightly below its expected value of one. When judging the fit of real data for either a wide test ( $W > 4$ ) or a wide sample ( $S > 1.0$ ), it would seem reasonable to be tolerant of item mean square dispersions somewhat larger than expected, but to work toward average mean squares falling slightly below one. (Wright & Mead, p. 53, 1980)

Thus, in a sense, the discussions has turned from unqualified confidence in the fit statistic as the primary measure of fit to the introduction of judgment to supplement the fit statistic.

The most tangible difference between the forms of fit statistic developed by MESA over the years is in the weighting factor applied to the fit mean square component. In every case, the fit mean square is based upon the standardized residual. This is labelled the FIT Z-SQUARED in the version of BICAL used in this study. However, close examination of the output produced by BICAL and BICAL.3, as well as the computational formula offered to explain the fit statistic, reveals that a weighting factor has been introduced to make the fit statistic more tractable. The different versions of the fit statistic appear to have evolved with greater experience in using Rasch analysis.

The formula for the fit mean square produced by the version of BICAL used in this study was not fully developed in the documentation received with the program. The earliest representation of the fit statistic found in this investigation refers back to 1969. It is presented as a chi-square in Research Memorandum 18 on pages 9 and 10.

Wright and Panchapakesan (1969) proposed a Pearson chi-square statistic for testing if the item calibrations are person free. This goodness of fit involves dividing the sample of subjects into subgroups and using the model to predict the numbers in each group expected to answer each item correctly. The test statistic may be computed as

$$C^2 = \sum_{i=1}^k \sum_{r=1}^{k-1} \frac{\left( a_{ir} - n_r P_{ir} \right)^2}{n_r P_{ir} \left( 1 - P_{ir} \right)}$$

(Wright & Mead, pp. 9-10, 1975)

where:  $C^2$  = a chi-square-like fit statistic computed on the total sample for one score level (i.e., level  $r$ ).

$k$  = the number of test items.

$a_{ir}$ , subscript  $ir$  = the total number of persons at score level  $r$  who got item  $i$  right.

$n_r$ , subscript  $r$  = the total number of persons at score level  $r$ .

$P_{ir}$ , subscript  $ir$  = the probability that those persons at score level  $r$  will get item  $i$  right.

This form of the fit statistic pertains to the total sample. It tests "goodness of fit" for persons in the total sample, who are at a single score level, to one item. But - the statistic used in this study to test goodness of fit pertains to persons in separate score subgroups of the sample, who are at a single score level, to one item. This statistic is computed by first computing the  $C^2$  for each score level subgroup established by BICAL and then dividing by the number of subgroups. Since MESA notation is not consistent from Research Memorandum 18 through Research Memorandum 23c in the various computational formula presented for the fit statistic, The following notation was established to facilitate comparison between them:

TFS = Total sample fit statistic on one item.

BFS = Between subgroup (i.e., score subgroup) fit statistic on one item.

$Z^2$  = the standardized residual (i.e., FIT Z-SQUARED).

i = one item.  
 v = one person.  
 o = observed.  
 e = expected.  
 $S^2$  = variance of the sample.  
 s = standard deviation of the sample.  
 n = a number of persons that is less than all persons tested.  
 N = all persons tested.  
 x = a score on one item.  
 X = a score on all items.  
 L = all test items.  
 G = all score subgroups produced by BICAL.  
 P = a probability of success.  
 K = some weighting factor.  
 $\chi^2$  = chi-square statistic.  
 t = Student's t-statistic.

When notation from this set is applied to the MESA formula for  $C^2$ , it is transformed to:

$$TFS = \sum_{i=1}^L \sum_{x=1}^{L-1} \frac{\left( \begin{array}{c} x - N P \\ o_i \quad e_i \end{array} \right)}{N P \quad e_i \left( \begin{array}{c} 1 - P \\ \quad e_i \end{array} \right)}$$

This formula translates verbally to: The Total Sample Fit Statistic on one item is equal to the number of items times one less than the number of items times the observed score of all tested persons on the item, minus the expected score of all tested persons on the item, divided by the product of the expected score times one minus the expected score, where P, subscript ei is the expected probability of getting item i right.

The statistic which is used in this study to test the goodness of fit is computed by averaging the standardized score subgroup residuals produced by the version of BICAL used in this study. The total of these residuals is equal to the total fit mean square (i.e., the TFS). That is, summation of the standardized residuals for score subgroups is another method for computing the total fit statistic. This average is called the between fit mean square (i.e., the BFS). The formula for computing the standardized residual (i.e.,  $Z^2 = \text{FIT Z-SQUARE}$ ), which the between subgroup mean squares is based upon, is:

$$Z_G^2 = \text{FIT Z-SQUARED} = \sum_{x=1}^{L-1} \frac{\left( \begin{matrix} X_{oi} - n_{G} P_{ei} \\ \end{matrix} \right)^2}{n_{G} P_{ei} \left( \begin{matrix} 1 - P_{ei} \\ \end{matrix} \right)}$$

This formula translates verbally to: The standardized residual for a score subgroup on one item is equal to the sum of one less than the number of scores on item  $i$  times the quotient of the square of the difference between the observed score of all persons in the score subgroups on the item minus the number of persons in the subgroup times the expected probability of success on item  $i$ , all of which is divided by the product of one minus the expected probability of getting item  $i$  right times the product of the number of persons in the subgroup times the expected probability of getting item  $i$  right.

The theoretical base for the fit statistic used in this research lies in the concept that the sum of individual score subgroup residuals is equivalent to the total fit mean square. The underlying assumption is that this is so because item fit has been shown in Rasch analysis to be independent of individual ability. Hence, item fit is likewise presumed to be independent of subgroup ability and even the ability represented by the total sample. Therefore any measure of item difficulty is, conceptually, the same whether it is based on a single item-person interaction, or on the interaction of a subgroup, or the total sample with that item. Consequently, the measure of variability in item fit based on the total sample (i.e., TFS) is presumed equal to the sum of the separate measures of variability in item fit based on each score subgroup. This relationship can be expressed:

$$\text{TFS} = \sum_1^G Z^2$$

Therefore, an average  $Z^2$  value is presumed to be equivalent to the TFS when interpreted by the appropriate number of degrees of freedom. Since there were six subgroups produced for every sample used in this study, the general form of the subgroup fit mean square values divided by the number of subgroups (i.e., six subgroups) is:



$$BFS = TFS/G = \frac{\sum_1^G Z^2/G}{G} = \frac{\sum_1^G Z^2/6}{6}$$

This is the computational form of the between fit mean square applied in this research. It amounts to a simple average of standardized residuals developed for score subgroups in the analysis. Subsequent forms of the fit statistic have been weighted in various ways to reflect changes MESA staff considered to be appropriate in the interpretation of the fit statistic.

In Research Memorandum 23, on pages 37 through 39 Wright and Mead introduced an approach to weighting the  $Z^2$ , and consequently the between fit statistic, as a "correction factor" which would cause the  $Z^2$  value to more closely approximate a true chi-square with one degree of freedom. Notation is modified somewhat in the following formula from the original text to facilitate comparison:

$$TFS = \left( \sum_{i=1}^L \sum_{x=1}^{L-1} \frac{\left( \begin{matrix} x - N & P \\ o_i & e_i \end{matrix} \right)^2}{N P \begin{pmatrix} 1 - P \\ e_i \end{pmatrix}} \right) K$$

According to Wright and Mead:

We obtain a chi-square statistic with one degree of freedom. The multiplier  $K$  is a correction factor, usually near one, to inflate the statistic to the equivalent of one degree of freedom. (Haberman, 1973)

If all of the  $N$  are equal and  $P$ , subscript  $e_i * (1 - P$ , subscript  $e_i)$  is nearly constant for all  $x$  and  $i$ , then  $K$  can be shown to be: The intuitive motivation for this can be grasped easily by noting that since  $i$  goes from 1 to  $L$  and  $x$  goes from 1 to  $N - 1$ , there are  $L(N - 1)$  statistics  $TFS$ . But, having fit  $L - 1$  item parameters and  $N - 1$  person parameters, there are only  $(L - 1) (N - 2)$  degrees of freedom available." (Wright & Mead, p38, 1975)

$$K = \frac{L(N-1)}{(L-1)(N-2)}$$

Wright and Mead then apply an appropriate modification of K to the standardized residual to obtain a single "corrected" score subgroup  $Z^2$  as follows:

$$BFS = \left( \frac{\sum_{G=1}^G \sum_{i=1}^{L-1} x_i \left( \frac{\sum_{oi} X - n P}{G e_i} \right)^2}{\sum_{G=1}^G \sum_{ei} \left( \frac{1 - P}{G e_i} \right)} \right) K_{GX^2}$$

$$\text{where: } K_{GX^2} = \frac{L G}{(L-1)(M-1)}$$

These standardized score group residuals may be summed to obtain an expression equivalent to the total fit mean square statistic TFS. That is:

$$TFS = \sum_G Z_G^2$$

Therefore . . . TFS "specifically asks the question would all score groups give the same estimate of difficulty for item i ?" (Wright & Mead, p.39, 1977c) The BFS procedure, according to Wright and Mead, "gives a chi-square statistic with G degrees of freedom." (Wright & Mead p. 39, 1977c)

In Research Memorandum # 23.c, Wright and Mead introduce a further refinement to the fit statistic which continues to employ weighting, but in a slightly different form, and the fit statistic is no longer described as a chi-square. The concepts of the "total fit t" and the "between fit t" statistic are introduced in this memorandum.

The fit statistic continues to be based upon the standardized residual, but the current form of the weighting factor appears in the expression:

$$BFS = \sum Z_{Gt}^2 K$$

$$\text{where: } K = \frac{L}{(L - 1)(G - 1)}$$

Interpretation of the fit statistic is not currently described in chi-square terms. The weighted between fit mean square statistic has been run through a transformation procedure which has converted it to a "t-like" statistic, and one which in current literature on the subject is described in t-statistic terms. This transformation is referred to by Wright and Mead as follows:

Finally, this mean square between groups can be expressed in the standardized form

$$t = \frac{av^{1/3}}{Bi} - a + 1.0/a$$

$$\text{where: } a = (4.5(M - 1))^{1/2}$$

This t-statistic tests whether the observed item characteristic curves have a common shape and slope. It has an expected value of about zero and a variance of about one. (Wright & Mead, p.11, 1980)

### The Implications of Weighting Factors

The consequences of weighting the standardized residuals used in computing the between fit mean square are considerable. The standardized residual in this study is not weighted. Using the value  $K = 1$  to represent this fact, the fit statistic produced in this study may be expressed by the expression:

$$BFS = \sum \frac{Z^2}{G} * K$$

To see what happens to BFS when K is interpreted as a weighting factor, consider the example in the next paragraph.

Using a test length of L = 100 items and a score subgroup count of G = 6, the following weighting factor values are produced using the weighting formulas which are presented in Research Memorandum 23 and 23.c, respectively:

$$K_{GX^2} = \frac{L G}{(L - 1) (G - 1)} = \frac{100.6}{(99) (5)} = 1.21$$

$$K_{Gt} = \frac{L}{(L - 1) (G - 1)} = \frac{100}{(99) (5)} = .2$$

Tests in this research range from 95 to 115 questions and 6 score subgroups were always produced by the analysis. These factors mean that the interpretation placed on the fit statistic produced by this analysis would first have to be inflated by approximately 20% for compatibility with the chi-square interpretation (i.e., Research Memorandum 23). Then, the fit statistic would have to be deflated by 80% for presentation to the t-transformation and compatibility with the t-statistic interpretation (i.e., Research Memorandum 23c).

What is the correct interpretation of the fit statistic? Probably there is no clear answer to this question at this time for every evaluation situation. It would be safe to say that Wright and his associates at MESA would favor the current t-statistic interpretation outlined in Research Memorandum 23c.

#### The Decision to Use BICAL

BICAL.3 employs a weighting factor in computing the fit statistic; BICAL does not. This was a decisive difference in the decision on which program to use. While there was ample opportunity to use BICAL.3 in this investigation, there are two reasons that this was not done. First, the unweighted between fit mean square is a more conservative

statistic simply because it has not been modified by a weighting factor. Thus, even if it were to be interpreted as a chi-square, the values computed by the program are seventeen to nineteen percent smaller than Wright has recently thought appropriate. Second, it may be questionable that the current, weighted between fit mean square, transformed to a t-like statistic, as developed by BICAL.3, has produced a more appropriate statistic than the unweighted fit mean square which has been used. This last point was raised during a discussion with Richard Smith in Chicago. Smith, in a general review of fit criteria development, pointed out an important distinction to be made between BICAL's unweighted between fit mean square used in this study and BICAL.3's fit t-statistic which was considered. The current between fit t-statistic is deliberately more robust than the unweighted between fit mean square. Smith indicated that the reason MESA had gotten away from emphasis on the unweighted fit mean square was their experience in evaluating National Board examinations. Some Board tests were up to twelve logits wide, compared to the three or four logit widths typically found in education. When a person who is at the high extreme of ability gets an item at the extreme low end of difficulty wrong in a test that is twelve logits wide, the squared residual is enormous! As Smith said:

. . . the probability of that is absolutely incredible. Can you imagine what that one intervention would do to a person's total fit statistic? He could have 999 other items that fit perfectly and one that does not, and it would send his fit statistic to the moon. (Smith, 1981)

This is exactly what was happening on certain individuals taking National Board tests. MESA was encouraged to develop a fit statistic where this would not happen. The fit t-statistic was the result. While no reference is made in Research Memorandum 23.c to this situation, Wright and Mead do point out the need for a more "robust" statistic. Referring to the total fit means square:

It too could be squared and summed, this time over all persons, to form a total mean square for the evaluation of fit. The resulting mean square, however, is very sensitive to unexpected responses which are far off target. This is unfortunate because when a response is far off target, that is when the item and person are far apart so that the difference between person ability and item difficulty is many logits, then not only is there very little useful information about either person or item in that response, but we hardly expect, nor do we need to expect, the model to hold.

An alternative approach which has similar asymptotic properties, but is more robust against off-target data is to weigh each squared residual by the information  $P_i$ , subscript  $i$  \*  $(1 - P_i)$ , subscript  $i$  it contains and so calculate the information weighted mean square. (Wright & Mead, p. 12, 1977c)

However, Smith felt the weighted fit mean square produced by BICAL.3 could mask item fit discrepancies which might warrant attention. He preferred the opportunity to personally evaluate a "shaky" item fit, rather than having the program make the fit determination to the extent that the current program does. Wright and Smith do not agree on this point. There is no way to know who is correct, but it seems preferable to use the unweighted statistic. This preference is based on lack of comparative knowledge on just how robust the current t-statistic really is. It was unlikely that the MEAP tests could be so wide, in any case, as to require use of an especially robust statistic. These feeling were born out in the analysis. The following table shows the range of item-difficulty for each of the 14 samples.

TABLE C-1

1973-1979 MEAP READING TEST ITEM DIFFICULTY IN LOGITS

4TH GRADE ITEM DIFFICULTY VALUES IN LOGITS			Year	7TH GRADE ITEM DIFFICULTY VALUES IN LOGITS		
LOW	HIGH	DIFFICULTY RANGE		DIFFICULTY RANGE	LOW	HIGH
-2.25	1.80	4.05	1973	3.37	-1.51	4.86
-1.71	1.74	3.45	1974	3.36	-1.43	1.93
-2.12	2.00	4.12	1975	3.61	-1.56	2.05
-2.07	2.12	4.19	1976	3.01	-1.62	1.39
-1.92	2.14	4.06	1977	4.22	-2.36	1.86
-2.12	2.28	4.40	1978	4.60	-2.57	2.03
-2.26	2.35	4.61	1979	4.08	-2.21	1.87
The average range value is 3.94 logits for all 14 samples.						

The range of item-difficulty was least in the 1976 seventh Grade test at 3.01 logits. The greatest range in item-difficulty occurred in the 1979 fourth Grade test at 4.61 logits. The average value is 3.94 logits for the 14 samples. Therefore none of these tests entail the extreme ranges in item-difficulty which is the primary rationale for using a robust fit statistic. Since the condition which the current weighting procedure is designed to correct is not present in these data, there does not appear to be any pressing need in this study to employ the current Rasch analysis computer program, BICAL.3, to cope with extremely large residuals. Instead, the strategy was adopted to choose the more conservative, but straightforward, statistic in this study which is produced by BICAL. It will identify lack of apparent item fit to the Rasch model for these data without weighting the fit statistic. As indicated, weighting introduces a procedural refinement to the analysis which is debated today among MESA staff. While the concept of controlling illogical extremes is something to consider, which weighting is designed to do, the need is not apparent in these data. Test width here is within the limits of common educational experience. The objective nature of the tests considered in this study also suggests that the items have a high content validity built right into them. There is probably less reason to expect extreme residuals in these tests than might be implicit in norm referenced tests. In any event, there is no convincing necessity for weighting the fit statistic in this study. By deciding to use the least refined fit statistic of those which might have been used in this study, the decision comes down squarely on the side which emphasizes interpretation of item fit in probability terms. That is, the decision was made to emphasize a statistical interpretation. Such an interpretation of item fit is appropriate to this analysis. But this approach would be of great concern to Wright. It is not consistent with his present thinking on the point. He expressed his position very effectively in a letter dated May 8, 1981:

Your work with MEAP sounds interesting and I would be very happy to give you whatever counsel on your use of Rasch measurement I can. I am thinking in particular of the fit statistics.

Do not be compulsive about fit decisions based on statistical values. The fit statistics draw your attention to items on which irregular behavior has occurred. Begin with the "worst" items and try to discover why each item did not fit. Examine the test and alternatives of each item carefully. Look at the distribution of incorrect choices. Can you find reason for the items not fitting? If so, then you will be happy to have had your attention drawn to them and will deal with them according to what you decide is wrong with them. As you work in from the "worst" items you will come to a point where either the fit statistics have become ordinary (mean squares nearer to 1.00 than 0.1 Or 0.2, or standardized mean squares nearer to zero than 2 or 3) or you cannot see any reasons for misfit. It is all right to stop at that point. A few marginal items will not damage the measurement you make.



In any case, whenever you use some of these calibrated items to measure a person you will want to check the extent to which that person has used the items according to their calibrations. This quality control of person fit will protect you and your respondents from being misled by meaningless measures from invalid responses." (Wright, 1981)

Wright's position on item fit analysis would appear to be more cautious than the approach which has been taken in this investigation. He counsels, in particular, against being "compulsive about fit decisions based on statistical values." This is very sound advice which can, no doubt, be well supported on both logical and conceptual grounds as a general rule. However, the decision to apply purely statistical criteria in this study to determine item fit is defensible. The corollary issues are discussed more fully in the concluding chapter of this study, but this difference goes to the heart of Rasch analysis applicability. Rasch analysis is introduced in the literature as a statistical tool. Early discussions offered sound arguments in support of the theoretical integrity of the Rasch model in mathematical terms. However, experience in working with the model appears to have revealed some shortcomings in a purely mathematical interpretation which are important to the application of the model. It is very likely that the Rasch model is theoretically sound. Nevertheless, the Rasch model may not be ready for use in an applied psychometric environment. There appears to be a growing body of more subjective concepts involved in the application of the model than was first thought necessary. As the subjective component grows, the practical application of the model may diminish to a point that the Rasch model is no longer a concept which can be seriously considered outside psychometric research laboratories. Such a measurement tool, that is interpretable by a very select few, will be a great disappointment to those seeking objective psychological measurement.

APPENDIX D  
SPSS CONTROL SET EXAMPLE  
FOR DESCRIPTIVE STATISTICS ON  
SAMPLE OF 5,000 DATA

UN NAME            DESCRIPTIVE STATISTICS FOR PILOT (1973) SAMPLE OF 4TH GRADERS  
 COMMENT            DATA FOR THIS RUN IS ON LINE FILE '-PILOT4'

DATA LIST        FIXED (3) /1 RECTYPE 2-3, GRADE 5-6, SEX B.  
                   AGEYRS 10-11, AGEMOS 13-14,  
                   NUMPASS 16-17, NUMTRIED 19-20,  
                   TOTQUEST 22-24,  
                   OBSCOR01 TO OBSCOR23 26-48,  
                   OBSTAT01 TO OBSTAT23 50-72/2  
                   QA1 TO QA5 2-6, QB1 TO QB5 8-12,  
                   QC1 TO QC5 14-18, QD1 TO QD5 20-24,  
                   QE1 TO QE5 26-30, QF1 TO QF5 32-36,  
                   QG1 TO QG5 38-42, QH1 TO QH5 44-48,  
                   QI1 TO QI5 50-54, QJ1 TO QJ5 56-60,  
                   QK1 TO QK5 62-66, QL1 TO QL5 68-72/3  
                   QM1 TO QM5 2-6, QN1 TO QN5 8-12,  
                   QO1 TO QO5 14-18, OP1 TO OP5 20-24,  
                   QQ1 TO QQ5 26-30, QR1 TO QR5 32-36,  
                   QS1 TO QS5 38-42, QT1 TO QT5 44-48,  
                   QU1 TO QU5 50-54, QV1 TO QV5 56-60,  
                   QW1 TO QW5 62-66

COMMENT  
 INPUT MEDIUM    DISK  
 N OF CASES       UNKNOWN  
 ALLOCATE        TRANSPACE=10000  
 COMMENT  
 COMPUTE         FRACTION\*0000  
 COMPUTE         AGE\*0000  
 COMPUTE         FRACTION\*AGEMOS\*.0833  
 COMPUTE         AGE\*AGE\*AGEYRS  
 COMPUTE         AGE\*AGE\*FRACTION  
 COMMENT  
 COND DESCRIPTIVE AGE  
 OPTIONS         3,4  
 STATISTICS      ALL  
 COMMENT  
 FREQUENCIES     INTEGER\*SEX(-9,1)  
                   NUMPASS,NUMTRIED(-9,23)  
                   QA1 TO QW5(-9,1)  
                   OBSCOR01 TO OBSCOR23(-9,5)  
                   OBSTAT01 TO OBSTAT23(-9,1)  
 OPTIONS         3,6,7,8  
 STATISTICS      ALL  
 FINISH

**APPENDIX E**  
**SPSS CONTROL SET EXAMPLE**  
**FOR CREATING STANDARD FORMAT SOURCE FILES ON**  
**SAMPLE OF 5,000 DATA**

UN NAME LINE FILE FOR PILOT (1973) SAMPLE OF 4TH GRADERS  
 COMMFNT  
 FILE NAME PILOT4  
 COMMENT  
 COMMENT DATA FOR THIS RUN IS ON LINE FILE '-FOURTH'  
 COMMENT  
 DATA LIST

FIXED RECTYPE 1-2, GRADE 33-34, SEX 46 (A).  
 AGEYRS 49-50, AGEMOS 51-52,  
 NUMPASS 63-64, NUMTRIED 65-66.  
 QA1 TO QAS 638-642 (A).  
 QB1 TO QBS 643-647 (A).  
 QC1 TO QCS 648-652 (A).  
 QD1 TO QDS 653-657 (A).  
 QE1 TO QES 658-662 (A).  
 QF1 TO QFS 663-667 (A).  
 QG1 TO QGS 668-672 (A).  
 QH1 TO QHS 673-677 (A).  
 QI1 TO QIS 678-682 (A).  
 QJ1 TO QJS 683-687 (A).  
 QK1 TO QKS 688-692 (A).  
 QL1 TO QLS 693-697 (A).  
 QM1 TO QMS 698-702 (A).  
 QN1 TO QNS 703-707 (A).  
 QO1 TO QOS 708-712 (A).  
 QP1 TO QPS 713-717 (A).  
 QQ1 TO QQS 718-722 (A).  
 QR1 TO QRS 723-727 (A).  
 QS1 TO QSS 728-732 (A).  
 QT1 TO QTS 733-737 (A).  
 QU1 TO QUS 738-742 (A).  
 QV1 TO QVS 743-747 (A).  
 QW1 TO QWS 748-752 (A).  
 OBSCOR01 TO OBSCOR23 798-820.  
 OBSTAT01 TO OBSTAT23 866-888 (A)

COMMENT  
 INPUT MEDIUM DISK  
 N OF CASES UNKNOWN  
 ALLOCATE TRANSPACE=10000  
 SAMPLE 0.2577  
 SELECT IF (RECTYPE EQ 17)  
 RECODE GRADE, AGEYRS, AGEMOS, NUMPASS, NUMTRIED,  
 OBSCOR01 TO OBSCOR23 (BLANK=-9)/  
 SEX (BLANK=-9) ('B'=1) ('G'=0) (ELSE=-7)/  
 QA1 TO QWS (BLANK=-9) ('\*' = 1)  
 ('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K' = 0)  
 ('O', 'O' = 0) (ELSE=-7)/  
 OBSTAT01 TO OBSTAT23 (BLANK=-9) ('Y'=1)  
 ('N'=0) ('O', 'O' = 0)/

COMMENT  
 VALUE LABELS GRADE (04) 4TH GRADERS (07) 7TH GRADERS/  
 SEX (1) BOYS (0) GIRLS/  
 OBSCOR01 TO OBSCOR23  
 (0) NO QUEST RIGHT (1) ONE QUEST RIGHT (2) TWO QUEST RIGHT  
 (3) THREE QUEST RIGHT (4) FOUR QUEST RIGHT (5) FIVE QUEST RIGHT/  
 OBSTAT01 TO OBSTAT23  
 (1) OBJECTIVE PASSED (-7) OBJECTIVE NOT PASSED  
 (-9) OBJECTIVE BLANK/

COMMENT  
 MISSING VALUES GRADE, SEX, AGEYRS, AGEMOS, NUMPASS, NUMTRIED.

QA1 TO QM5,OBSCOR01 TO OBSCOR23,  
OBSTAT01 TO OBSTAT23 (-7,-9)

COMMENT  
COUNT TOTQUEST=QA1 TO QM5 (1)  
COMMENT  
COMMENT THE FOLLOWING CARD DESIGNATING LOGICAL I/O UNIT #9 IS OPTIONAL  
RAW OUTPUT UNITS  
COMMENT  
WRITE CASES (1X,F2.0,1X,F2.0,1X,F1.0,1X,F2.0,1X,F2.0,  
1X,F2.0,1X,F2.0,1X,F3.0,1X,23F1.0,1X,23F1.0/  
1X,SF1.0,1X,SF1.0,1X,SF1.0,1X,SF1.0,  
1X,SF1.0,1X,SF1.0,1X,SF1.0,1X,SF1.0,  
1X,SF1.0,1X,SF1.0,1X,SF1.0,1X,SF1.0/  
1X,SF1.0,1X,SF1.0,1X,SF1.0,1X,SF1.0,  
1X,SF1.0,1X,SF1.0,1X,SF1.0,1X,SF1.0,  
1X,SF1.0,1X,SF1.0,1X,SF1.0)  
RECTYPE, GRADE, SEX, AGEYRS, AGEMOS,  
NUMPASS, NUMTRIED, TOTQUEST,  
OBSCOR01 TO OBSCOR23, OBSTAT01 TO OBSTAT23,  
QA1 TO QA5, OB1 TO OB5, OC1 TO OC5, QO1 TO QO5,  
QE1 TO QE5, QF1 TO QF5, QG1 TO QG5, QH1 TO QH5,  
QI1 TO QI5, QJ1 TO QJ5, QK1 TO QK5, QL1 TO QL5,  
QM1 TO QM5, QN1 TO QN5, OO1 TO OO5, OP1 TO OP5,  
OQ1 TO OQ5, OR1 TO OR5, OS1 TO OS5, OT1 TO OT5,  
OU1 TO OU5, OV1 TO OV5, OW1 TO OW5  
READ INPUT DATA  
FINISH

## APPENDIX F

### BICAL - A COMPUTER PROGRAM FOR USE IN CONDUCTING RASCH ANALYSIS

#### CODING BICAL CONTROL CARDS

##### Introduction

Two versions of the Rasch analysis computer program, supplied by the Department of Education at the University of Chicago, were considered in this study: BICAL and BICAL.3. The earlier version, BICAL, produced the Rasch analysis which is the basis for the analysis and conclusions developed in this investigation. The rationale for this choice will be developed in a subsequent part of this section. While the later version of the Rasch analysis program, BICAL.3, was also applied to all fourteen samples considered in this study, the results appeared to be more relevant to considerations for future research than the objectives of this study. Some additional comment was made on this possibility in the concluding section on recommendations for future research in Chapter V of this study.

The documentation which supports the information in this section on coding the control cards necessary to use the earlier version of BICAL is limited. There were two items of significance which accompanied the copy of the BICAL program obtained from Wayne State University: 1) a copy of Research Memorandum Number 18, CALFIT, Sample-Free Item Calibration with a Rasch Measurement Model, published by the Department of Education of the University of Chicago in March of 1975 (Wright & Mead, 1975), and 2) four pages of material entitled "How to Use 'BICAL' with no author or date of publication specified. Memorandum 18 (Wright & Mead, 1975) describes an even earlier version of BICAL than was used here known as CALFIT. Memorandum 18 is a good general reference on Rasch analysis and proved to be accurate in its description of control card format parameters for the BICAL program used in this investigation, as far as it went. However, there is no reference in that memorandum respecting parameters seven to fourteen in the second control card format. Also, the output tables for the example problem in Memorandum 18 (Wright & Mead, 1975) differ in a number of important respects from that actually produced by the program using Sample-of-5,000 data. The four pages which made up the second reference source proved to be complete and very helpful in learning how to code the program control cards but there was no explanation of the output which could be expected from the program. There does not appear to be any single source covering the complete documentation for the BICAL program acquired by Wayne State University. It appears that as a substitute, documentation for CALFIT was probably provided by the University of Chicago supplemented by the four pages of material on control card coding for "BICAL".

The combination of these two references and experience in actually using the program are the basis for most of the material in this appendix. While Memorandum 18 (Wright & Mead, 1975) and the four page supplement were drawn upon heavily in developing this material, neither document is quoted directly here as there is nothing to be gained in

doing so. Any attempt to correlate them directly with this material could be confusing without serving any useful purpose. While the coding procedures outlined in this appendix does have some general application to using other versions of the BICAL program, it is intended to have specific application to the version of that program provided for use in this study. However, comments apply to both versions unless otherwise noted.

### The Control Card Formats

The control cards which are required to run either version of the Rasch analysis computer program considered in this study, BICAL or BICAL.3, are similar in appearance and function. These cards amount to a set of parameter statements which, in affect, tell the program essential facts about the data being processed, and the way it is to be processed. They describe the data and instruct the program where to locate the material being processed in the data file; what to label that material; and which of two alternative processing methods to apply to it. Mistakes in specifying any of the data parameters which the program requires will certainly result in an unsuccessful run, often without any indication that an error condition is present and occasionally at the risk of interminable, and expensive, looping through the program instruction sequence without any indication whatever that there is a problem. The program does not have error indications for a number of problem conditions which experience has shown are easily encountered. Therefore, it is in the user's best interest to be extremely careful in setting up a set of run control cards.

The earliest version of the Rasch analysis computer program used in this study, BICAL, has eight control card formats. The later version, BICAL.3, which was considered but not used, has nine. Each control card format discussed here conveys one, or more, parameters to the program. Usually only one control card is necessary for each format, but there are exceptions. General coding requirements, for both versions of BICAL, as well as the coding used in this study, are described in the following paragraphs.

Format 1 (both BICAL and BICAL.3), the "Title Card": Used to identify each computer run with a title up to 80 characters in length. One card is coded anywhere in columns 1 to 80 with a job title of the user's choice. This card must be included even if the user elects not to title the job. A typical heading for the BICAL runs in this study was: "FITTING 1973 MEAP TEST RESULTS (FOR FOURTH GRADERS) TO THE RASCH MODEL".

Format 2 (both BICAL and BICAL.3), the "Input Description Card": Used to describe up to fourteen data description; computer program processing; or computer system parameters. Again, only one card is used for this format. This is the most complex control card. Only specifically designated columns may be used for each parameter. Five columns have been allocated to each one, though it is highly unlikely every position will ever be used to code any parameter. Parameter codes



go in the rightmost, or low-order, positions. All of the coding done on this card is either numeric or blank. The 14 format parameters and corresponding coding schemes, and the card columns dedicated to them, are as follows:

1. The total number of test items is entered in card columns 1 - 5.

The number of test items must be specified.

2. The minimum number of persons desired in each score group (i.e., persons at a certain ability level), which the program sets up as part of the analysis, is coded in card columns 6 - 10. The program will establish up to six score groups ordered from least to most ability. If the user does not specify score group size, the program will use a group size of 25 persons by default. The default value of 25 persons was coded for minimum score group in this investigation for every computer run.

The minimum score group size is an optional parameter.

3. The minimum score to be considered, or included, in the analysis is coded in card columns 11 - 15. Persons who score below this minimum are automatically eliminated from the analysis by the computer program. Candidates may be expected to get one question in five right by chance since all MEAP reading test items are multiple choice with five alternatives. It was decided in this study that purely chance scores should not be considered in the analysis. Therefore scores equal to 20% of the total possible correct for each test were set as the minimum acceptable scores in this analysis.

The documentation for both BICAL.3 and BICAL indicates that the program will not include any person in the analysis who has either a perfect score or who misses every item.

Subject to these conditions, the minimum score parameter is optional.

4. The maximum score to be considered, or included, in the analysis is coded in columns 16 - 20. Persons who score above this maximum are automatically eliminated from the analysis by the BICAL program. This value was set at 114 for all 14 computer runs employed in this study. The program will not process a perfect score, therefore, whatever number of items in a given test, the maximum score which will be processed by BICAL is one less than the maximum possible score. Consequently, the program was able to discern both the correct number of test items, and one less than that number, throughout this analysis.

The range of the maximum score value must lie between one more than the minimum score specified in paragraph "3" above and one less than the total possible score. If a value is not specified for the most current version of the program (i.e., BICAL.3), a value equal to 90% of the total possible test score will be selected as the default value by the program. The documentation for CALFIT gave no indication of the consequences which follow if a maximum score parameter is not given.

Subject to these limitations, the maximum score parameter is optional.

5. The total number of columns which make up the input record are placed in card columns 21 - 25. Whenever more than one card is used in the input record, the user must specify a number for this parameter which would include, at the very least, up to the last significant column position in the last 80 positions, plus a full 80 positions for each previous "card" used for data on each person taking the test.

The total number of columns which make up the input record is a mandatory parameter.

6. The calibration code is placed in card column 30. The program can employ two methods to estimate test item-difficulty: "PROX" and "UNCON". The user selects PROX by placing a "1" in column 30. The alternate UNCON is specified by a blank or a "2" in column 30.

The calibration code is a mandatory parameter. If the user does not specify either 1 or 2 for PROX or UNCON respectively, and leaves column 30 blank, the program will assume the intention is to use UNCON. In a sense, a blank is a default value which will bring the more generally useful, but more expensive, UNCON method for estimating item difficulty in terms of computer processing time into operation.

7. The scoring code is placed in card column 35. The program scores all test responses dichotomously: correct or incorrect. The BICAL documentation refers to four code values in the scoring code parameter:
  - a. blank Or "0" (zero) directs the program to score items correct or incorrect on the basis of whether or not the item response corresponds exactly to (i.e., is equal to) the scoring key, which is the sixth control card format.
  - b. "1" (one) indicates to BICAL (BICAL only - do not use with BICAL.3) that the data is already scored. By implication, the program employs the response directly as scores (i.e., 1 for correct and 0 for incorrect). Any other values than 1's or 0's would be unacceptable to the program if code 1 is used.

- c. "2" directs the program to score test items correct on the basis that item responses are less than or equal to the key values; incorrect if they are greater than the key values.
- d. "3" directs the program to score the test items correct on the basis that item responses are greater than or equal to the key values; incorrect if they are less than the key values.

Here again, a blank is, in a sense, a default value which has the same effect as code 0. Therefor conscious selection of a scoring code is forced upon the user, making it a mandatory parameter, though it is not so identified in the documentation.

- 8. Logical input unit code is placed in card column 40. This code identifies the logical unit which the computer will use to read the data input file when that file is not incorporated into the control card set. The usual convention is to identify these logical input-output units by number.

While the documentation implies that this is not a mandatory parameter, the user again must make a conscious determination of the appropriate input unit for the test data. A blank will direct BICAL to specific action as effectively as any operable logical unit code.

- 9. The starting column (i.e., the first column) of the data accompanying the test scores which identifies the student who took the test is entered in columns 44 and/or 45.

This parameter was not mentioned in any of the documentation supplied with BICAL. It may not be operable in this version of the program. This description is based entirely on material from Memorandum 23.c (Wright & Mead, 1980). This column indicates the starting point of a data identification field. The end of this field is designated by the next parameter. This field ranges in size from a minimum of one to 20 columns in width. Not only does it provide a means of entering identification information with score data, but this field also serves as identification of output from the program. Should the user elect to employ this optional capability of the program, it is possible to generate a new data file on each individual in the analysis which contains score data, an ability estimate in logits, a score histogram, and an ability/item-difficulty plot.

If a value is not designated, the program assumes column 1 by default.

- 10. The ending column of the data identification field is entered in column 49 and/or 50.

This parameter was not mentioned in any of the formal documentation supplied with Wayne University's copy of BICAL (Wright & Mead, 1975). It may not be operable in this version of the program. This description is based entirely on material in Memorandum Number 23.c (Wright & Mead, 1980).

11. The logical output unit code is placed in card column 55.

This parameter was not mentioned in any of the formal documentation supplied with Wayne University's copy of BICAL (Wright & Mead, 1975). It may not be operable in this version of the program. This description is based entirely on material in Memorandum Number 23.c (Wright & Mead, 1980).

This code identifies the logical unit which the computer will use to output a new data file on each individual in the analysis, should it be desired. This optional file contains score data on ability estimate in logits, a histogram, and an ability/item-difficulty plot. The output file provides for seven individual variables in addition to identification data: raw score, ability in logits, ability standard error, total t fit, mean square standard deviation, weighted mean square, and standardized response residuals. To implement this BICAL.3 option, additional JCL (i.e., job control language) cards would have to be included with the program source code, and the program would have to be recompiled. Detailed specifications for representative JCL cards will be found in the BICAL.3 documentation (Wright & Mead, 1980, pp. 89 - 90).

12. The twelfth parameter, coded in card column 60, was used for a different purpose in BICAL than it is in BICAL.3. In the earlier version of the program, this parameter can be used to place limitations on the histograms and tables which the program produces. In the current version, it is used to set limits for screening persons who do not fit the Rasch model from the analysis which estimates person-ability and/or item-difficulty.

the codes used in each version, and their purpose are presented, for comparison, as follows:

<u>BICAL</u>	<u>BICAL.3</u>
b, 0 : Print all plots.	b, 0 : No one deleted for misfit.
1 : Omit score histogram.	GT 0 : Persons whose total t fit
2 : Omit fit plots.	is greater than CFIT/10
3 : Omit both.	will be deleted for
	misfit.

(Wright & Mead, 1977, p. 105)

(Wright & Mead, 1980, pp. 80 - 90)

13. The simulation mode of the program can be induced by coding any value greater than 0 (zero) in card columns 61 - 65.

This parameter was not mentioned in any of the formal documentation supplied with Wayne University's copy of BICAL (Wright & Mead, 1975). It may not be operable in this version of the program. This description is based entirely on material in Memorandum Number 23.c (Wright & Mead, 1980). However, the simulation mode "could" be operable with this version of the program if access to two random number generator subroutines were provided by the user, or if linkage to the two random number generators which were at one time accessed by this program at the Wayne University Computer Center were again provided. Access to these random number generators, as presently coded in the program, is no longer operable.

While this parameter is no longer operable in the version of BICAL used in this study, BICAL.3 does contain its own random number generators which can be activated by this parameter.

14. The fourteenth, and last, parameter in the second format card is coded in card columns 66 - 70.

This parameter was not mentioned in any of the formal documentation supplied with Wayne University's copy of BICAL (Wright & Mead, 1975). It may not be operable in this version of the program. This description is based entirely on material in Memorandum Number 23.c (Wright & Mead, 1980).

This parameter controls the output of item statistics which BICAL.3 has the capacity to generate. This option was not implemented in this study at any point. BICAL.3 produces elaborate statistical tables and charts without implementing this option which more than adequately serve the purposes of this study. The primary advantage, and purpose, of this option is that it affords a means of generating individual item statistical data in a form which makes it useful as input to further computer processing and analysis. Though not attempted here, this option could be an extremely useful feature of the program for researchers interested in concentrating on item analysis or in creating large test item banks, neither of which pertain in this investigation.

Format 3 (BICAL.3 only), the "Variable Format Card": Used to inform the program of the number of single column fields, which are each read as alpha-numeric data, which comprise the input record. Coding begins in the first card column. A single field is coded "(A1)". the form "(NA1)" is used to code up to 80 (single column) fields, where N represents the number of fields from 1 to 80. When more than 80 columns are required, the form "(80A1/. . ./NA1)" is used. Each slash (/) represents a new input record "card" of, up to a maximum of, 80 columns, and N represents the number of fields in the last card. All specifications prior to the last card identify a full 80 columns.

The total number of columns specified in this parameter card must equal the number specified for the total number of columns which make up the input record for parameter 5 of Format 2.

Format 4 (BICAL.3)/Format 3 (BICAL), the "Item Name Card(s)": Used to provide test item labels up to four characters in length. BICAL.3 documentation indicates that this format is mandatory (Wright & Mead, 1980), where the BICAL documentation does not (Wright & Mead, 1975).

As many of these cards may be used as needed to provide a label for each item in the test. Up to 20 item labels may be specified per card. Coding should begin in the first column of the four column fields set aside for each label, but this is not mandatory so long as a label does not overlap two fields. The first label field is columns 1 through 4 on each card; the second label field is 5 through 8; etc.

Format 5 (BICAL.3)/Format 4 (BICAL), the "Column Select Card(s)": Used to indicate to the program how the data in the input record is to be used. Three uses may be specified for data in the input record: 1) A blank or "0" (zero) indicates that one column is to be skipped; 2) A "1" indicates that the column is to be used; 3) An "&" (ampersand) indicates that the column is to be skipped, as a blank or zero would indicate, but the ampersand identifies a label and is also counted in the total number of items specified in the first parameter of Format 2, though the item is excluded from the analysis. This code makes it easy to delete test items on subsequent analysis that do not fit the Rasch model without changing any of the other format card parameters.

Format 6 (BICAL.3)/Format 5 (BICAL), the "Scoring Key Card(s)": Used to present the scoring key to the program. The same number of cards must be prepared here as are prepared for the preceding format. The entries on these cards reflect all of the correct item responses in the positions corresponding exactly to the item responses in the input record. A blank, or any symbol, may be used in columns identified by the column select card(s), Format 5, to be skipped as they will be, appropriately, ignored.

Format 7 (BICAL.3)/Format 6 (BICAL), the "Options Labels Card": Used to identify up to 20 multiple choice alternatives in BICAL.3 and up to 5 in BICAL. The program keeps track of the number of times each response alternative to an item is chosen. The user should enter whatever symbol, letter, or number is used to identify responses. If five responses were identified by the first letters of the alphabet, for example, this format should be coded "A B C D E".

Format 8 (BICAL.3)/Format 7 (BICAL), the "Data Card(s)": Use to store item score data when that data is presented to the computer as an integral part of the control card set. The control card set may be actually read from punch cards run through a punch card reader, or from disk or tape files. Either way, the test data can be incorporated with the control card set if the user wishes. However, it is often more convenient to present the data to the computer apart from the control

card set. This approach makes it unnecessary to consciously separate data from the control cards on multiple test runs, for one thing. If the data is not read as an integral part of the control card set, this format is not used.

Format 9 (BICAL.3)/Format 7.a (BICAL), the "End of Data Card(s)":  
Use to indicate to the program when the end of the score data card file has been reached. Code an asterisk, "\*", in column 1. BICAL.3 documentation instructs the user to use the same number of these format cards as the number of cards in a single input record (Wright & Mead, 1980). Nothing is said in the BICAL documentation about using more than one of these format cards (Wright & Mead, 1977).

Format 10 (BICAL.3)/Format 8 (BICAL), the "Simulation Header Card":  
Use to direct the program to simulate the data input rather than read it from a card, disk, or tape file. Enter "SIMULATE" in the first eight columns of this card. If this format is used, it must be followed by the next format card.

Format 11 (BICAL.3)/Format 9 (BICAL), the "Simulation Task Description Card": Use to describe four statistical characteristics the user desires the generated sample to have and a seed number for the programs' random number generator.

This format was not mentioned in any of the formal documentation supplied with Wayne University's copy of BICAL (Wright & Mead, 1975). It may not be operable in this version of the program. This description is based entirely on material in Memorandum Number 23.c (Wright & Mead, 1980). When this option is used, this format must be used in conjunction with the preceding format card.

Five parameters are associated with this format. The first four parameters are mandatory in every simulation run. The fifth parameter is mandatory in the first run in a sequential series of simulation runs. Only specifically designated columns may be used for each parameter. Five columns have been allocated for each one. Parameter codes go in the rightmost, or low-order, positions. All coding in this card is numeric or blank. The five parameters and corresponding coding schemes, and the card columns dedicated to describe them, are as follows:

1. The range of item difficulties which the user desires the program to simulate is coded in columns 1 - 5.
2. The total number of persons which the user wants the program to include in the simulation is coded in card columns 6 - 10.
3. The mean ability of the test group that the user desires the program to simulate is coded in card columns 11 - 15.
4. The standard deviation of the test group that the user desires the program to simulate is coded in card columns 16 - 20.

5. The seed number to initiate the program's random number generator is coded in card columns 21 -25 of single simulation runs. If a continual series of runs is being made, this parameter must be coded only in the first control card set of the first run in that series. The parameter is left blank in all other control card sets in that series.

Format 12 (BICAL.3)/Format 10 (BICAL), the "End of Job Card": Use to indicate to the program when the run is terminated. Code four asterisks, "\*\*\*\*", in the first four columns of this format. Only one of these format cards is used in a single run. However, numerous tests, each described by a corresponding set of control card formats, may be included in a single run. This format must be used every run. Otherwise the program will continually loop in search of additional control card sets.

#### Aspects of Control Card Formats Specific to this Study

It is hoped that the preceding section presents sufficient general information on completing BICAL control cards to enable most prospective users to implement the program. The following discussion in this section is concerned with certain impressions developed as a result of applying BICAL to the MEAP Reading Test samples employed in this investigation. Further discussion of these points is raised in this section in an attempt to reduce the risk of confusion to those persons who wish to be able to use BICAL but may never encounter the concerns raised in this study which were associated with certain card formats. Most of the control card formats were so easy to use that they raised no questions. The following paragraphs relate only to issues raised in using control cards for analysis of the data in this study. Each paragraph is identified by the specific format in question and, where applicable, the parameter number associated with the point under discussion is indicated to facilitate reference back to the general material in the preceding section.

Format 2, Parameter 1: The MEAP tests studied here had either 95 items (i.e., the 1974 to 1979 fourth grade tests); 100 items (i.e., the 1974 to 1979 seventh grade tests); or 115 items (i.e., both the fourth grade and seventh grade tests administered in 1973). However, all control sets used in this investigation were coded "115" in the field designating the number of test items to take advantage of a property of the BICAL program which makes it possible to reduce the number of repetitive coding changes otherwise necessary, without the very real risk of making errors in the process. The point to be made here is that it was not necessary in this study to change the item number parameter for each BICAL run. The program will automatically delete test items from the analysis which no-one taking the test gets right. For this study, items not given in the 1974 through 1979 MEAP tests were all coded "0" for every student with the result that these items were ignored in the analysis for those years. This has precisely the same



affect as if they had not been present at all, which is the desired affect, without risking the multitude of potential errors which would have to be avoided in reformatting data files and control cards to actually delete the questions dropped from the 1974 through 1979 tests.

Format 2, Parameter 2: The BICAL program automatically tries to establish at least two score groups (i.e., ability groups), and can create up to as many as six. Should the user not specify score group size, the program will use the default value "25" and attempt to set up ability groups with at least 25 persons. When this happens, there must be at least 50 persons in the sample or the program will not be able to create this default minimum number of groups (i.e., two groups). Actually the user may specify any size score group, but fewer than 10 is not recommended. The number of score groups created by the program in a given run is a function of the minimum group size specified and the total number in the sample. The program first attempts to establish six score groups of approximately equal size, and will do so as long as none is smaller than the specified minimum. If this does occur, the program will repeatedly drop one score group and redistribute the sample across a successively small number of groups until the minimum size requirement specified by the parameter is met or the program finds that at least two groups of minimum size are not possible. Since 1000 persons were in every sample studied here, there was never any chance that there would be fewer than six score groups. Each group has approximately 166 persons in it throughout this analysis.

Format 2, Parameter 3: The minimum score chosen for each of the 14 analysis done in this investigation was chosen to eliminate scores which were possible purely by random guessing. Each question has five alternatives, suggesting that a minimum score equal to approximately 20% of the total score possible would be appropriate for this purpose. Scores exactly equal to 20% of the total possible correct for each test were chosen. This amounted to 23 correct for the 1973 fourth and seventh grade tests having 115 questions each. It was 19 for the fourth grade tests and 20 for the seventh grade tests, respectively, for the years 1974 through 1979. The latter tests had 95 and 100 items respectively. Had the minimum score values chosen for this investigation not been used, the value "1" probably would have been chosen as an appropriate alternative. Nothing is said about guessing in the documentation accompanying BICAL.3 or BICAL. Nor is guessing mentioned as a major concern in the MEAP literature. While guessing may be a major concern under some conditions, it does not appear to be a problem in this study. The adjustment made here for guessing simply provides a more conservative estimate of item-difficulty. This is the primary reason for making any adjustments for guessing at all in this study.

This is an example of an optional parameter. Persons using any computer program should be aware of the consequences of not specifying an optional parameter on how the program will run. It is not enough to know how specifying a parameter will affect the analysis. It is a mistake to assume that no value is required by the computer program simply because none is required for the intended analysis. The user must understand that the computer program will invariably substitute a

default value in such cases, and it may not cause to happen what the user assumes, or desires, will happen. Something should be specified by the user. Conversely the program will stop whenever it attempts to use an undesignated parameter that is mandatory. But the fact that computer programs automatically provide default values for optional parameters, and the consequences, often escapes the notice of a casual user. Good program documentation details the presence, value, and affect of default parameters. Good coding practice suggests that the user consciously code an acceptable value in the analysis even though a parameter is represented as optional. This approach avoids unacceptable surprises at the conclusion of the computer run.

BICAL does not consider a person in the analysis who has a score of zero or a perfect score. However the program will include a person with a score of 1. For the reasons indicated, the decision was made in this study not to accept scores below 20% of the possible total. Scores below this level would not provide the quality of analysis sought. Therefore a minimum score value of 1 would not have been acceptable in this study. However, this probably is the default value for the minimum score parameter. The BICAL.3 (Wright & Mead, 1980) and the CALFIT (Wright & Mead, 1975) documentation differ on this important point. CALFIT documentation says nothing about a default value or the acceptable range for this parameter. The user can not know for certain. He, or she, must guess. This is dangerous. Guessing what a computer program will do can have disastrous results. It may do the unexpected. Whenever an erroneous parameter value could have such results, it is far safer to consider the parameter to be "critical" to the study. Therefore throughout this investigation, the practice of specifying an acceptable value for a program parameter was adopted whenever the program documentation suggests that the parameter, as in this case, is optional. While nothing was said about the default value or the acceptable range of values for the minimum scores parameter in the BICAL documentation (Wright & Mead, 1977c), the BICAL.3 documentation (Wright & Mead, 1980) specifies both. The BICAL.3 documentation specifies a default value equal to 40% of the total test score possible. This value is unnecessarily restrictive for the purposes of this study.

That it is present illustrates the danger in assuming that the default value for a parameter in effect for one version of a program is still in effect for a later version. Here the shift has gone from an apparent value of 1 in CALFIT and BICAL to 40% of the total possible score in BICAL.3. BICAL.3 allows the user to specify a range for the minimum score from 1 to one less than the total possible score. If a user left the parameter blank in the analysis of a 100 item test because a minimum score of 1 is appropriate to the analysis, BICAL would substitute 1 for this parameter, but BICAL.3 would substitute 40!

Format 2, Parameter 4: The nature of MEAP tests is such that a large percentage of the students taking them might score 100%. This could be considered a fortunate outcome in the tests considered in this investigation. In any case, there is no reason to deliberately eliminate high MEAP scores at any level no matter how frequently they may occur. The value "114" was chosen to set the maximum possible score at the highest possible level. This value is one less than the number

of questions (i.e., 115) in the 1973 fourth grade and seventh grade reading tests. These two tests were the largest of the fourteen considered in this analysis. The six other fourth grade tests had 95 items. The six other seventh grade tests had 100 items. As indicated previously, the BICAL program was presented with 115 items in every case, but the program ignored the excess items above the actual number in the test.

When the initial BICAL control set was coded and run, during the first week in April of 1980, only the CALFIT documentation was available at that time (Wright & Mead, 1975). It indicated that this parameter was optional. The decision was made at that time to specify the maximum value which could be given to the score parameter to insure that the highest scores acceptable to the program would be processed. It is essential to this analysis that the maximum score be as large as possible for every sample processed in this study. There is no indication in the BICAL documentation (Wright & Mead, 1977c) that the program has a default value. Documentation for BICAL.3 (Wright & Mead, 1980) refers to a default value of 90% of the total number of test items. Had the value of 114 not been specified as the maximum score, subsequent runs on the newer version of BICAL would have substituted "104" as the maximum score. This would have had a serious negative affect on this analysis. The range of the maximum score value is also specified in the BICAL.3 documentation (Wright & Mead, 1980, p. 88) where nothing is said on this point in connection with the earlier versions.

Any value set by the person using the program must be chosen by criteria which are important in the user's own judgement.

Format 2, Parameter 5: The BICAL program reads data in 80 column increments. Some users will recognize this input format as a "unit record". The important thing to understand about this format is that it must be at least large enough to include all of the input data. It can be larger. If all of the input data can be included in a single 80-column increment, then the value specified for the total number of columns which make up the input record may be set anywhere from the last column data in the analysis is to appear, up to and including column 80. For example, if input data ends at column 57, this parameter may be any value between 57 and 80. However, if more than 80 columns are needed to store the input data, or if the user elects to employ more than 80 columns without placing data in every column from 1 to 80, more than one 80-column increment must be used.

In this investigation, three 80-column increments (i.e., unit records or cards) represented each person-record. The last significant position in the third card was column 66. Therefore, the value 226 (i.e.,  $80 + 80 + 66$ ) could have been specified in this study for this parameter, since the format of the input data would be found in one logical record between column 1 of the first card and column 66 of the third card. This parameter is intended to specify a minimum number of columns in contiguous 80-column increments which would encompass a single person-record. Therefore since 226 columns would do this, the number 226 would have satisfied the requirements of this parameter.

However, the user must carefully note the exact position of the last data character in using this approach. If the number is too small, the last position(s) of significant input data will not be read by the program. If there are blank positions after the data, the user can specify a large enough number for this parameter to make column counting unnecessary and yet insure that the appropriate input data would be considered. For these reasons, the decision was made to specify a value of 240 for this parameter which is equal to three full 80-column increments.

Format 2, Parameter 6: PROX is an abbreviated method for calibrating item-difficulty. It is quite accurate so long as the input sample is symmetrical (i.e., normally distributed) and the test is a "long" test. Both conditions should be present if the user is to consider using PROX. Unfortunately, the CALFIT documentation (Wright & Mead, 1975) does not provide any suggestions on the number of items in a long test. The tests in this investigation are very likely long enough for PROX to be considered since they have 95 to 115 items, but the data is not symmetrical.

MEAP Tests favor high scores. Hence score groups tend to be very negatively skewed. For this reason the decision was made not to use PROX in this study. Since the greatest appeal of this method is computational efficiency, when compared to the alternative UNCON, selecting the most appropriate method of calibrating item-difficulty was a concern in this study. It involves considerable computer processing time and this was expected to be a major expense consideration. However, the inappropriateness of PROX to the data in this investigation left no alternative than to choose UNCON. This method for approximating item-difficulty is appropriate for all test data, but it should be used exclusively whenever that data is generated from a "short" test or when the data is skewed. Again, test length was not a factor here, but the fact that the data tended to be seriously skewed was the determining factor in the decision to use UNCON. All fourteen samples are negatively skewed in this study. Therefore UNCON was chosen as the method for estimating item-difficulty throughout this investigation.

Format 2, Parameter 7: BICAL, the version of the Rasch analysis program used in this investigation, responds to a scoring code of "1" as if all of the item responses presented to the program have already been scored. This was the scoring code used for all of the computer runs employed in this study. MEAP Sample-of-5000 data contains both actual responses and scored responses for each test item. However, only the scored responses (i.e., 1 for correct and 0 for incorrect) were used here. The documentation for the later version of the program (Wright & Mead, 1980, p. 89), BICAL.3, indicates that the code "1" should not be used to indicate scoring code at any time. However, though subsequently corrected and rerun, the first series of runs using BICAL.3 were inadvertently made on all 14 samples using code 1 for this purpose without any indication of difficulty. Apparently the admonition against using code 1 with BICAL.3 to indicate scoring procedure does not adversely affect scoring binary (i.e., 1 for correct and 0 for incorrect) responses where the correct response will always match the key. There was no error indication generated by BICAL.3 when code 1 was

used, and the scoring output was identical to that produced when the samples were rerun. Therefore, it appears that the reruns probably were unnecessary in this instance, but the prospective user is cautioned not to ignore the admonition against using code 1 for the scoring code with the more recent version. Very likely data responses other than 1's and 0's (i.e., binary responses) may not be scored correctly if the user chooses to ignore this admonition. A possible explanation why code 1 worked here despite the warning may be that BICAL.3 defaults to a blank scoring code, or the blank equivalent, when 1 is used, and the program then scores on the basis of responses equivalent to the key. Or, possibly, the program treats responses as being already scored under these conditions. From the speculation that binary coded responses would be treated identically if code 1 were used as it would if either code 0 or a blank were used as the scoring code, while non-binary responses would not, it may be accurate to further speculate that the use of code 1 may now be discouraged merely to simplify coding this parameter. However, when the programmer currently responsible for maintaining BICAL, a Miss Susan Bell, was asked why use of code 1 was discouraged, she said that "the binomial coding method initiated by code 1 was no longer used". This option was instituted after BICAL but before BICAL.3 was released.

Format 2, Parameter 8: Very commonly the logical input unit code used on the Wayne State University computer system may be specified by either coding "3, 5, 8, or 12" for this function, depending on the code used to record the data and the device (i.e., disk, tape drive, etc.) on which it is stored. In this application, a blank or 0 (zero) logical unit code will be interpreted by the BICAL program as logical input unit 5. The program will attempt to read the data from the punch card reader. The logical input unit code 5 is implemented on most computer systems, but the user may specify any logical unit code which has been implemented on the computer system being used to run the analysis. Tape or disk files are the other common source of input data for computer runs. All samples used in this study were transferred from tape onto disk files. The logical unit code 8 designates disk or tape input to the Wayne University computer system, and this code was used to identify the logical unit number in every computer run employed in this study.

Format 2, Parameter 9: Since there was no wish to produce individual files on persons taking these tests in this study, the option which would make it possible to identify such information was not implemented here. The default value "1" in the first column of the identification field was coded throughout the analysis. In fact, this parameter may not even be operable in the version of BICAL which is used in this study.

Format 2, Parameter 10: A code of 0 (zero) was used for the ending column of the data identification field. As indicated in the previous section, this parameter was not mentioned in any of the documentation supplied with the program by Wayne Computing Center personnel (Wright & Mead, 1975). However, it is mentioned in the BICAL documentation (Wright & Mead, 1977c) and also in the BICAL.3 documentation (Wright & Mead, 1980).

The latter material indicates that this individual record identification field may be coded with from one to 20 digits. The number, from 1 to 20 entered in this field indicates to BICAL.3 the size of the identification field. While a larger value than 20 may be entered in this field, the program will default to 20. This is the maximum size of the data identification field permitted by the program. Since there was no intention to use the data identification field in this study, the affect of coding the starting column parameter 1 and the ending column parameter 0 (zero) which caused BICAL to default to a one-column identification field in the first data column, was of no consequence. By coding 1 for this parameter, the same end would have been accomplished and the intent of the program's authors would have been met fully. An end column value is called for in the later documentation and it should fall between 1 and 20. But, experience in using the program with 0 (zero) coded for this parameter has shown that the program will operate without difficulty. There is no indication in the BICAL documentation what the results would be if this parameter were left blank. It should be noted in retrospect at this point that great difficulty was experienced in interpreting the BICAL documentation which was available at the early stages of this investigation. It became necessary to seek out the help of Ernie Bauer who has extensive experience with the program. Dr Bauer is the Director of the Assessment Office of the Oakland (County) School District in Pontiac, Michigan. The Oakland School District had been experimenting with Rasch measurement since 1977 and had successfully implemented the BICAL computer program on a number of occasions.

In the first meeting with Bauer, he indicated that personnel in his office had also found early BICAL documentation difficult to use, but he was now in a position to provide the additional information needed to implement the version of BICAL being used in this investigation. He suggested use of 0 (zero) for this parameter and all of the remaining parameters from 11 on in the second format card. His department had used this approach successfully in runs against test data from his school district. Since there really was no interest in the functions which these parameters, according to the later documentation, would provide, the decision was made to followed Bauer's advice. The program worked at this stage, and so the decision was made to retain zero codes in the last five parameters, including this one, throughout this study on the second format card. Nevertheless, a few added comments respecting the use of the remaining parameter in the second format card may be of interest to those wishing to use BICAL.3, the more recent version of the BICAL program. All of the remaining comments in this section on the remaining parameters in the second format cards are based entirely on the documentation for BICAL.3 (Wright & Mead, 1980).

Format 2, Parameter 11: The numbers "4, 6, 9, and 11" may be used to designate logical output units on the Wayne State University computer system. Usually unit 9 would be considered for this purpose in connection with the Rasch analysis program. The user would code 9 in this instance to output run results on either the card punch or to computer tape or disk files. No file will be output if a blank or zero is coded for this parameter. As previously indicated, this parameter was not used in this study. It is coded 0 (zero) for every computer run.

Format 2, Parameter 12: This parameter, too, was coded 0 (zero) in all of the computer runs in this study. The effect of this parameter varied between BICAL and BICAL.3. In BICAL, it affected the printing of all histograms and plots generated by the program. In BICAL.3, this parameter simply prevented the elimination of persons from the analysis which did not fit the Rasch model. BICAL.3 will automatically eliminate persons who do not fit the Rasch model. This is an extremely valuable feature in some forms of analysis which is not available in BICAL. Both versions of the program eliminate items which do not fit the Rasch model. This feature of BICAL.3 could have been a useful adjunct to this investigation, perhaps, but there are important properties in BICAL, the version of the Rasch analysis program used here, which caused it to be selected in preference to BICAL.3 as the basis for the analysis done in this study.

Format 2, Parameter 13: Both BICAL and BICAL.3 can be induced to read artificial, or randomly generated, data rather than run against actual, or live, data. This simulation mode feature would be of interest to the researcher wishing to study the properties of the Rasch model and/or the computer program using controlled data input. Since there was no present interest in these matters, the value 0 (zero) was coded for this parameter for every computer run, thereby directing the program to process the actual score data presented to it.

Format 2, Parameter 14: BICAL.3 has the capacity to generate statistics in a form suitable as input to further computer analysis by other programs. The primary advantage, and purpose, of this option is that it affords a means of tying the output of the Rasch analysis program to any form of software package which may be available and understood by the user.

Format 3 (BICAL.3 only): Since the input records used in this investigation were comprised of three full 80-column card formats, the variable format card would be coded (80A1/80A1/80A/). This format is mandatory with BICAL.3, but it was not used at all with BICAL. This is the most notable difference between coding the format cards for the two versions of the program as they were employed in this study.

Format 4 (BICAL.3)/Format 3 (BICAL): Item names were used throughout this study. Scored items are grouped by learning objective in the Sample-of-5000 record; five items per objective. Item names used in this study have been tied to 23 objectives in the MEAP program. Each objective is coded "A" through "W". Therefore, the five items in the first objective are labelled "QA1, QA2, QA3, QA4, and QA5". The five items in the second objective are labelled "QB1 through QB5", etc. The last five items, for the 23rd objective, are labelled "QW1 through QW5". Identification of items for objectives dropped in the 1974 through 1979 tests were retained in the input record but not scored or analyzed. Therefore, 115 test items were identified in every computer run, including those items associated with objectives that were dropped after 1973. Each item, in every test analysis, can be identified as to the objective which it is intended to measure.

Format 5 (BICAL.3)/Format 4 (BICAL): Since there are three data input cards to each BICAL and BICAL.3 run executed in this study, three corresponding column select cards were prepared. Because the first data card had only demographic data, none of which was to be input to the computer program, all 80 columns in the first column select card were coded 0 (zero) to cause these entries to be bypassed by the program. Scored test items were coded 1, for correct, and 0 (zero) for incorrect, in groups of five. Beginning in column 2 of the second and third column select card, 1's were coded in groups of five, corresponding to the positions of the scored responses in the data file. Each group of five responses is separated by a blank in the data file, so a zero was coded corresponding to the card columns separating item groups of five each. Eight trailing zeros in the first column select card, and fourteen in the second, correspond to the unused positions in the last two data cards. The "&" code is not used in this study as it was not the purpose of this analysis to investigate the affect items which did not fit the Rasch model have on test analysis. No further analysis was made of test results once non-fitting items were identified. The analysis here focuses on the affect of non-fitting items on the probability a student will pass the objectives set for the test.

Format 6 (BICAL.3)/Format 5 (BICAL): Zeros are used in the first of the three scoring key cards to correspond to the demographic portion of the input record. Every column of the second and third scoring code card was coded "1" in this study. Since 1 is the correct answer and by coding every column of the scoring key cards, this coding scheme guaranteed presentation of the correct answer (i.e., "1" in this study) to the item in the corresponding column select cards. The zero coding in the column select cards simply caused the surplus, unnecessary 1's to be ignored.

This approach saved the unnecessary work of counting scoring key card columns to insure that only those which correspond to the column select and input data are used. Thus considerable chance for making mistakes in coding key cards was avoided in this study. Unfortunately this approach to coding scoring key cards can only be used when correct answers are identified, as here, by the same symbol (i.e., "1"). Otherwise, an accurate key must be carefully prepared to insure that the right code gets into the right column.

Format 7: Since, at first, the BICAL documentation (Wright & Mead, 1977c) on the use of the options label format card was misinterpreted, five alternatives as follows: "1 2 3 4 5". Only two responses, "0 (zero) and 1", should have been used. Because of this coding error, which was not discovered until all of the BICAL runs had been completed successfully, the program tracked five responses: "0 1 2 3 4". However, all responses fell under either 0 or 1, corresponding to incorrect and correct responses respectively, as they should. Therefore, there was no need to recode this parameter and rerun the samples.



Format 8 (BICAL.3)/Format 7 (BICAL): The data was entered from one disk file and the program control cards from another for all of the runs made in this study. The placement of the data in the record, and the data itself, determines how the data format card is to be coded. If the data is not read as an integral part of the control card set, as in this study, this format should not be used.

Format 9 (BICAL.3)/Format 7a (BICAL): Since all of the input data in this study was on computer disk files rather than punch cards, the end of data card format was not used.

## IMPLEMENTING BICAL

### Problems Encountered in Coding BICAL Control Cards

Coding control cards to implement the Rasch analysis computer program was difficult for both versions, BICAL and BICAL.3, used in this study. The attempts to code the cards and use BICAL began in September of 1980. Coding problems encountered at that time were not resolved until March of 1981. Attempts to use BICAL.3 began in late May of 1981. Problems in coding the control cards were again encountered. In addition, the BICAL.3 source code would not compile at first without serious errors. These difficulties were ultimately resolved the seventh of July, 1981.

BICAL: BICAL, the earlier version of the Rasch analysis program, was obtained from the Wayne State University Computer Center in the Spring of 1979. The decision had already been made not to support the program. By the time BICAL could be used, in September of 1980, there was no one at the Computer Center who could provide any assistance.

BICAL Documentation: The documentation received with BICAL was brief and unclear. Its source was not indicated on much of the material received with the computer code and no one at the University Computer Center could track it down.

BICAL Modification: This version of BICAL includes an automatic interrupt which referenced two random number generators that once were available in the Wayne University Computer Center public program files. However the references had been changed and the two subroutines called by the program no longer worked. Reference to one of the routines, under another label in Wayne's revised public file dictionary could be found, but the other never was. Considerable time and effort was spent in tracking this down because there was no way of knowing if the two random number generators were required to run BICAL. It turned out that they were not relevant in any way. The program interrupt may be safely bypassed by entering "IGNORE" at the terminal and the program would proceed to function. It now seems apparent that Wayne University Computer Center personnel had modified the BICAL program to utilize random number generator routines in the program's simulate mode which were more to their liking than the random number generator which is part of the program. That part of the program no longer functions.

After considerable trial an error, the first partially successful BICAL run was accomplished in January of 1981. While all of the expected output was generated, the program had been looping during most of the eight minutes it was allowed to run before being manually terminated. The cost of this first attempt was \$44.03. To learn what had gone wrong, a meeting was set up with Bauer a second time, in early March. Enough was learned at this meeting to correctly prepare the card coding necessary to successfully run BICAL for the first time on March 8, 1981. The final run against the last of the 14 sample in this study was completed April 7, 1981. The next step in this investigation was to determine from these analysis which test items did not fit the Rasch model. At the suggestion of Bauer, a meeting with his assistant, William Veitch, was arranged for the morning of April 17, 1981 to discuss item fit criteria.

It was during this meeting with Veitch that the existence of an even more recent version of BICAL, BICAL.3, became known. Knowledge that a more recent version of BICAL was available forced a considerable delay in this study. It was felt that it should not proceed if the more recent Rasch analysis program could be more appropriate.

The BICAL.3 program that was ordered, and subsequently received on May 23, would not compile despite accompanying assurances that it had been compiled and tested at The University of Chicago. Investigation revealed one major coding error and two minor program design problems which became apparent only with tests involving 100 questions or more. Unfortunately, these problems proved difficult to solve. They were not resolved until the seventh of July, 1981. All fourteen samples in this study were run against BICAL.3 in an early morning terminal session that lasted a few minutes short of four hours on that date.

Two major references which have considerable information on item fit are: 1) Research Memorandum #23 (Wright & Mead, 1977c), and 2) Research Memorandum #18 (Wright & Mead, 1975). The later reference presents a fine discussion of the Rasch evaluation model and CALFIT, a version of the Rasch analysis program which preceded BICAL. Memorandum 18 has only historical value now to persons interested in using BICAL. However in April of 1981, it was one of only two documents which could be located at the time that spoke to the application of the item fit statistic in Rasch analysis with some authority. The authors of both of these monographs, Wright and Mead, had been most active in promoting understanding of the Rasch measurement concept and in developing and disseminating the computer program for its implementation. Dr Benjamin Wright is the Director, and Ronald Mead the Assistant Director, of the Department of Education Measurement and Statistical Laboratory at the University of Chicago. Both men have been extensively involved with development of Rasch measurement in all of its aspects. The development of an item fit statistic has been the subject of much of their attention and both of these monographs address the topic of applying an item fit statistic to Rasch measurement. Unfortunately, Memorandum 18 and Memorandum 23 presented only an abbreviated treatment of this major point, and a number of questions related to item fit could not be resolved through their help alone.

Problems experienced in this study concerned with interpreting the statistics produced by BICAL were discussed with Veitch. In particular, concern was expressed with the concept of an item fit statistic. The literature is unclear on this point and the more effort that was given to resolving this confusion, the more it seemed to take hold. Veitch pointed out that what might be perceived to be contradiction in the literature was more likely the result of changing conviction on the part of the authors. He felt that Memorandum 18 and Memorandum 23 were probably seen by their authors a "true" at the time they were written, but that their perception of a fit statistic has undergone an evolutionary change. He provided a copy of Research Memorandum #23.c (Wright & Mead, 1980) which represents the most recent published statement on using the Rasch analysis computer program: the version referred to as "BICAL.3". Veitch suggested that this study should not be completed without first investigating the most recent application of BICAL. He felt that there might be even more differences between present and early practice which should be considered in this research. He was correct, of course. Subsequent study of Memorandum 23.c revealed that the discussion of an item fit statistic was different from either Memorandum 18 or Memorandum 23.

Veitch pointed out that he, Bauer, and several other personnel in the Oakland Schools Assessment Office had found it necessary to contact both of the authors of these memorandum, Wright and Mead, over the years respecting new developments in Rasch measurement and the applications of the BICAL program. On one occasion, the Oakland Schools District hosted Mead in Pontiac. He came to the District offices to explain the application of BICAL. On that occasion there was considerable discussion respecting the application of an item fit statistic. Veitch then explained his understanding of the application of a item fit statistic. It was the most definitive explanation so far encountered, but presentation of the content and outcome of this part of our discussion must be ignored for the present. They will be brought out in more detail in succeeding sections of this discussion where they are more relevant. Veitch's mention of the more current version of BICAL forced resolution of item fit questions to be postponed. This study could not proceed until more was known about the new program called BICAL.3.

#### CODING PROBLEM SUMMARY

Problems encountered with BICAL.3 because the program would not compile or run properly due to coding errors are inexplicable. This is especially true since the program was represented as "tested" and output data was supplied with the source code that was represented as output from that very source code. These were serious and time consuming errors, yet their occurrence is very unlikely to happen again. But, they might, or at least equally unexplainable problems might occur when any new user attempts to implement an unfamiliar computer program. These events emphasize the reality that computer programs do not always work as intended, even when they have been developed and tested by, as in this case, the most thoroughly competent and reputable source. BICAL.3 worked very well once these few functional coding problems encountered were eliminated.

There were also some problems encountered in coding the control card set for use with BICAL.3 which, for a time, prevented the program from working which were every bit as exasperating as those found in the source code. The same set of control cards that worked well for BICAL caused problems when they were used with BICAL.3. Though the program ran, BICAL.3 would only process the first 60 questions of every test presented to it. The program ignored the remaining questions, ranging in number from 35 to 55, depending on the test. The problem was traced to the Variable Format Card. This card was not part of the BICAL control card format set. It had been placed in the BICAL3 control set between the Data Description Card and the Item Description Card(s), which are the second and third, respectively, BICAL format cards. The Variable Format Card was, at first, coded incorrectly as follows: (240A1). Correct coding, as indicated in the prior discussion of Format 3 (BICAL.3 only), Variable Format Card, is: (80A1/80A1/80A1). Only the first 80 columns were being read by the incorrect format. The BICAL.3 documentation did not indicate how input records which involved more than one unit record should be coded on the Variable Format Card. This problem is representative, in a symptomatic sense, of the basic problem that has pervaded all BICAL and BICAL.3 documentation. It is often too abbreviated to be easily understood by the person interested in using any version of the Rasch analysis computer program.

The documentation leaves a great deal to be desired. The material that is available serves reasonably well as reference for persons who may already know how to use the program or for data processing specialists. The uninitiated user is at a considerable disadvantage because the material may be too abbreviated for his purposes. The most serious problem encountered with the documentation in this investigation seemed to result from the fact that this research involves tests having more than 100 questions and the fact that a relatively large amount of demographic data is part of each record. The documentation appears to be geared to tests having fewer than 100 questions and input records with a minimum amount of identification information in each record. This more limited conception of the tests which will be presented to the program could explain why the problems encountered in this investigation were not anticipated in the descriptive material for setting up run control card parameters. Several control card specifications, when interpreted literally as described, would not work when applied to the data in this study. The program did not work and there was no indication of the cause.

A more extensive example might help. The documentation describes using this program on a comparatively short test given for a few people and very little demographic information accompanies the score data. While the descriptive material adequately describes control card preparation for such a test, it is not adequate for larger test records which include extensive data unrelated to the scores. In any case, more complete documentation would have been extremely helpful toward the implementation of this research.

## PROGRAM EXECUTION

A typical MTS run command used to execute BICAL was: RUN BICAL 5=CONTROL 6=-OUT 8=-SAMPLE. To execute BICAL.3, the run command was: RUN NEW.BICAL 1=-ONE 2=-TWO 5=CONTROL 6=-OUT 8=-SAMPLE. The later run command included the "1=-ONE" and "2=-TWO" elements at the suggestion of the Wayne State University FORTRAN consultant who assisted in efforts to get BICAL.3 to run. Since the consultant did not know how the program functioned exactly, he felt that these elements provided access to "scratch" input and output files should they be needed by the computer system. It later came to be apparent that they were not needed, but these elements were allowed to remain in the BICAL.3 run command. The elements of both run command forms are described as follows:

1. BICAL and BICAL.3 are the names of permanent MTS line files which contain, respectively, the BICAL and BICAL.3 object (i.e., compiled) code.
2. -ONE is the name of a temporary MTS scratch or work file available to the program for containing internally formatted data generated by the computer system.
3. -TWO is the name of another temporary MTS scratch or work file available to the program for containing internally formatted data generated by the computer system.
4. CONTROL is the name of the permanent MTS line file which contains the program format control cards.
5. -OUT is the name of a temporary line file available to the program to contain the program output.
6. -SAMPLE is the name of a temporary MTS line file from which the program reads input data.

The data files were transferred to fourteen temporary line files from computer tape, using \*FS RESTORE commands. Since there were repeated runs requiring clearing of the temporary file -OUT, the decision was made to set up a run file which contained the following MTS instructions:

```

GET -ONE
GET -TWO
GET -OUT
EMPTY -ONE
EMPTY -TWO
EMPTY -OUT
RUN NEW.BICAL 1=-ONE 2=-TWO 5=CONTROL 6=-OUT 8=-SAMPLE

```

To execute a program run, first copy one of the samples to -SAMPLE and then enter: "SOURCE RUN", where "RUN" is the name given to the run file. For example, one instruction sequence might be:

```
COPY -SAMPLE5 TO -SAMPLE
```

## SOURCE RUN

On execution of the SOURCE command, the computer system could read the contents of the run file, executing each command in sequence, as it appears in the run file. A similar approach was used to run BICAL, but without reference to 1=-ONE or 2=-TWO.

### Program Running Time

Since the first successful BICAL run cost over \$44, the prospect that the computer processing in this study might easily exceed the financial resources available for that purpose was of great concern. There is no indication in the documentation about running time for either BICAL or BICAL.3. The program is occasionally referred to in the Rasch literature as an "efficient program", but little else is said about what the user should expect in running it. Probably it is efficient, considering the great number of complex tasks which it performs. BICAL consists of 18 subroutines comprised of 1373 lines of FORTRAN code. BICAL.3 consists of 20 subroutines in 1767 lines of code. Yet either form of the program will read scores for 1000 students who have taken 115 items and generate 40 pages of tables and graphs in less than 90 seconds on the Wayne University AMDAHL/6 computer system.

However, syntax errors in coding the program control cards can lead to endless looping of this program in the run mode, with attendant high CPU charges. Run time guidelines would be helpful to the new user to provide some point of reference when the program is in trouble. There are no error indicators or automatic program interrupts built into the program which will automatically terminate a bad run, nor error messages which will notify the user of problems.

A stopwatch was used on all runs in this study. Most run times ranged between a minute and 15 seconds and a minute and 50 seconds. All runs were done very late at night when traffic on the computer system was at an absolute minimum, so there is probably very little queuing delay in these times. There is no doubt measurable inaccuracy in this timing method, but experience has shown that 1000 cases on 115 items, using UNCON the most elaborate of the two item difficulty estimation procedures, should probably execute in approximately 90 seconds. Knowledge of "typical" run times has proved to be very helpful in this study on those occasions the program did not run correctly, or not at all.

## INTERPRETING BICAL OUTPUT

Introduction

Output generated by the BICAL program employed in this study includes a representation of the control card set; two tables showing results of data editing; two histograms; one table/ogive; two tables; and four plots; all reproduced on 20 to 24 pages. An example of this output, based on the 1974 fourth Grade MEAP test, is presented in the last section of this appendix.

Output Format

Page 1 partially presents the contents of the program control card formats used. The Title Card, Format 1, contains: "FITTING 1973 - 1979 MEAP TEST RESULTS (FOR FOURTH GRADERS) TO THE RASCH MODEL". The Input Description Card, Format 2, is identified in the printout by the label "CONTROL PARAMETERS". There are 19 parameter fields on this card; each is five columns wide. Only 12 are actually functional in the version of BICAL used in this study:

1. Columns 1 - 5 are labelled "NITEM". This field holds the total number of test items considered in the analysis: 115.
2. Columns 6 - 10 are labelled "NGROP". This field holds the minimum number of persons desired in each score group: 25.
3. Columns 11 - 15 are labelled "MINSC". This field holds the minimum score to be considered, or included, in the analysis: 19.
4. Columns 16 - 20 are labelled "MAXSC". This field holds the maximum score to be considered, or included, in the analysis: 114.
5. Columns 21 - 25 are labelled "LREC". This field holds the total number of columns which make up the input record: 240.
6. Columns 26 - 30 are labelled "KCAB". This field holds the calibration code which directs the choice of item calibration methods available in the program: 2.
7. Columns 31 - 35 are labelled "SCORE". This field, the last containing a mnemonic label, holds the scoring code which directs the choice of item scoring methods available in the program: 1.
8. Columns 36 - 40 are labelled "I". This field holds the logical input unit code which directs the choice of logical input unit that the program is to use for reading data: 8.

9. Columns 41 - 45 are labelled "2". This field holds the starting column (i.e., the first column) of the identification data which are part of the card(s) containing the test scores: 0.
10. Columns 45 - 46 are labelled "3". This field holds the ending column of the identification data which are part of the card(s) containing test scores: 0.
11. Columns 51 - 55 are labelled "4". This field holds the logical output unit code which directs the choice of logical output unit that the program is to use for outputting data: 0.
12. Columns 56 - 60 are labelled "5". This field holds the histogram code which directs the choice of producing or not producing the histograms and/or plots available in the program: 0.
13. Columns 61 - 65 are labelled "6". This field holds the simulation code which directs the choice of inducing the simulation mode by using a program, integral random number generator in the program: 0. This parameter is not operable in the version of BICAL used in this study.
14. Columns 66 - 70 are labelled "7". This field holds the code used in later versions of BICAL (i.e., specifically BICAL.3) to control the output of individual item statistics. While a code 0 (zero) appears in this field in the printout, this parameter was not implemented in the version of BICAL used in this study.
15. Columns 71 - 90 are labelled, across five column increments, "8" through "11". No parameter has been implemented in these fields in any version of BICAL to date.

The Item Name Card, Format 3, does not appear in the printout. The Column Select Card, Format 4, is identified in the printout under the label "COLUMNS SELECTED", below the line of asterisks and 0's, at ten column intervals. There are three select cards in this run. The first is all 0's. The next two are comprised of 0's and 1's. The Scoring Key Card, Format 6, is identified in the printout under the label "KEY". There are three key cards in this run. Each one is labelled. The first is all 0's. The next two are all 1's.

None of the remaining six possible format cards are represented in this printout. The balance of page 1 is devoted to a printout of the first complete record encountered in the data file and an indication of the number of items and the number of subjects input to the computer program for analysis. The first record is identified in the printout under the label "FIRST SUBJECT". There are three cards in every person record. Each one is labelled. The first card contains extensive demographic and MEAP test performance data on the individual student.



The next two contain the MEAP reading test scores on that individual. Scores are grouped by test objective; five items to each objective. The number of items presented to the program is 115. This number appears after the label "NUMBER OF ITEMS". The number of items is constant throughout this analysis. The number of subjects presented to the program, in this instance, is 998. While there are 1,000 students in every one of the fourteen samples considered in this analysis, the BICAL program will determine in advance of any further processing whether or not any students in the sample got either all items right or all items wrong. Either way, such persons will be eliminated from further analysis. Apparently there were two individuals who either got all items right, or they got all items wrong, in this sample.

Page 2 is actually three pages long in this printout. Only the first page is labelled "PAGE 2". It presents the number of persons, by item shown in ascending sequence, who selected the different alternatives possible with each question. The table is captioned "ALTERNATIVE RESPONSE FREQUENCIES". Beneath the table caption are nine columns labelled "SEQ NUM, ITEM NAME, 0, 1, 2, 3, 4, UNKN, and KEY", respectively. The numbered columns identify the item alternatives. Since there were only two alternative possibilities in this analysis, incorrect and correct identified by 0 and 1 respectively, it would have been more tidy in this instance to code two options, 0 and 1, rather than five, 0 through four, as was done. The sequence number, as it appears in the test data, appears under the column heading SEQ NUM. The three character label given to each item in this analysis appears under the heading ITEM NAME. The alphabetic portion of the name corresponds to one of 23 learning objectives measured in the MEAP reading tests between 1973 and 1979. The numeric portion of the name identifies the first through fifth item which corresponds to a given learning objective, coded "A" through "W". The number of persons getting each item wrong (i.e., code 0) or right (i.e., code 1) appear under the appropriate column headings. No person appears under alternatives 2 through 4 because, as indicated earlier, these were not legitimate item alternatives in this analysis. All items were coded, so there are no entries under the column heading UNKN. Since the data was already scored, there are no entries under the column headed KEY.

Page 3 presents output which is the result of further refinement of the input file, prior to actual item calibration. This page amounts to a record of persons and items dropped from the analysis according to the control card parameters and/or program limitations. BICAL will not process either a perfect score or a zero score. This limitation is build into the program. The user may specify even greater limitations on the score range through entries in control card format 2. On taking these restrictions into account, the program will proceed to apply them to the sample being analyzed. The record of persons and items dropped from the analysis as a result appears on this page. Page entries are treated in the following discussion in the order which they appear.

The number of persons dropped from the analysis because they get all the items wrong was 2. This number appears after the label "NUMBER OF ZERO SCORES". The number of persons dropped from the analysis because they got all items right was 0. This number appears after the

label "NUMBER OF PERFECT SCORES". The number of items presented to the program for analysis was 115. This number appears after the label "NUMBER OF ITEMS SELECTED". The number of items in this analysis which were given names by the user was 115. This number appears after the label "NUMBER OF ITEMS NAMED". There is a user set lower limit of 19 items in this analysis which determines the score below which a student will be dropped from the analysis. The number 19 appears after the label "SUBJECTS BELOW" which in turn is followed by the number "10" which is the number of persons dropped from the analysis because they scored below 19. There is also a user set upper limit of 114 items in this analysis which determines the score above which a student will be dropped from the analysis. The number 114 appears after the label "SUBJECTS ABOVE" which in turn is followed by the cipher 0 which is the number of persons dropped from the analysis because they scored above 114. Thus two persons were dropped from the analysis because they got all items wrong and ten were dropped because they fell below a score of 19. Twelve, in total, were dropped from the analysis leaving 988 of the original 1000 records for input to the succeeding phases of analysis performed by the program. The number 988 appears after the label "SUBJECTS IN CALIBRATION". The sum of subjects presented to the program for full analysis, 988, plus the students dropped because of user set scoring limits, 10, equal 998, the total number of subjects considered up to this point in the program. The number appears after the label "TOTAL subjects". Not only does BICAL reject subjects from the analysis who fail to meet certain scoring criteria, but the program will also eliminate items from the analysis which no-one gets correct.

Twenty items were dropped from the analysis because no-one got them right. The balance of page 3 is devoted to a presentation of the items dropped and to a summary of the entire person/item editing process which BICAL has performed up to this point. BICAL will drop any item no-one gets right. It happens, in this instance, that the 20 items that were dropped from the analysis were the twenty items dropped from all fourth grade MEAP reading tests from 1974 on. Since they were not given, no one could get them right of course. The technique was adopted in the analysis of deliberately scoring all items dropped in fourth grade and seventh grade reading tests from 1974 through 1979 as 0. Since no one got them right, the program dropped these 20 items from the analysis without having to make changes to input record formats at the considerable risk of causing input error problems if mistakes were made in the process. Since there were no other items in the analysis which no one got right, only the twenty items dropped from the '74 through '79 tests appeared at this point in the printout. The dropped items appear in a table captioned "REJECTED ITEMS". Beneath the table caption are three columns labelled "ITEM NUMBER, ITEM NAME, and ANSWERED CORRECTLY". The sequence number of the dropped item, as it appears in the test data, appears under the column heading "ITEM NUMBER". The three character label given to each item in the analysis appears under the heading "ITEM NAME". The number of persons getting each of these items right appear under the heading "ANSWERED CORRECTLY". The program does not print sequence number over 99 in this table. Therefore, the sequence numbers for items QW1 through QW5 (i.e., items 111 through 115) print as asterisks (\*), indicating that the number is too large for the space provided by the program to print it. At this point, BICAL will re-tally

the number of persons which should be retained in the analysis. When items are dropped, it becomes possible that a person whose score includes one or more of these items no longer meet the score parameters set for the run. This happened here.

The number of subjects deleted (i.e., 2) from the first pass because they missed all items is repeated after the label "SUBJECTS DELETED". The fact that an additional subject has been deleted after dropping the 20 items is reflected in the number 987 which appears after the label "SUBJECTS REMAINING". This number is one less than the number of persons retained on the first pass, 988, which appears after the label "SUBJECTS IN CALIB" above.

Deletion of 20 items from the original 115 items presented to the program leaves a possible score of 95. The total number of items deleted, 20, and the resultant total score, 95, appear after the labels "ITEMS DELETED" and "POSSIBLE SCORE" respectively. This is appropriate for the 1974 fourth grade reading test as there were only 95 items actually in the test. The minimum score of 19, set by the user, appears after the label "MINIMUM SCORE". The program sets the maximum possible score at a level one less than the maximum number of acceptable items which, in this case, amounts to a score of 94. This number appears after the label "MAXIMUM SCORE". At this point, BICAL has completed editing person and item acceptability according to score parameters which are either set by the program and/or by the user. The balance of the program output is based upon items and persons which meet or exceed these parameters. In this example, that means that the remaining analysis sample of 1000 1974 fourth graders taking the MEAP reading test proceeds on the basis of 95 items and 987 students.

Page 4 is actually two pages long in the printout. Only the first page is labelled "Page 4". It presents the first of two histograms produced by the program. Under the heading "DISTRIBUTION OF ABILITY", this histogram shows the "COUNT" and "PROPORTION" of students taking the test who score at every possible scoring level in this test from 1 to 95 items correct. This chart is a graphic representation of test group ability in terms of the proportion of that group at each score level.

Page 5 is actually two pages long in the printout. Only the first page is labelled "Page 5". It presents the second of two histograms produced by the program. Under the heading "DISTRIBUTION OF EASINESS", this histogram shows the "COUNT" and "PROPORTION" of students taking the test who gets each item correct. This chart is a graphic representation of item-difficulty in terms of the proportion of that group that succeeds in answering each item correctly.

Page 6 is actually two pages long in the printout. Only the first page is labelled "PAGE 6". It presents the results of the item-difficulty estimation processing in tabular and ogive form. There are two procedure alternatives: "PROX" and "UNCON". The procedure chosen for this example BICAL run is indicated by the phrase "PROCEDURE USED UNCON" in the upper left corner of the page above the table portion of the printout. UNCON is an item-difficulty estimation procedure which often requires more than one iteration. That is, it is a procedure

which may repeat, or cycle, two or more times in an attempt to reach an optimum result. UNCON cycles until further repetitions would afford little or no improvement in the item-difficulty estimate. There were three iterations in this run, which is indicated by the phrase "NUMBER OF ITERATIONS = 3" which also appears in the upper left corner of the page, just below the procedure identification. The tabular portion of this page has two parts.

The first part of the table is comprised of columns headed "SEQUENCE NUMBER, ITEM NAME, ITEM DIFFICULTY, STANDARD ERROR, and LAST DIFF CHANGE", shows the estimates of individual item-difficulty developed by the UNCON procedure. Items appear in the same sequential order which they have in the input data, accompanied by the user assigned names. The mean item-difficulty is set at zero. Items which actually have fewer correct answers than the Rasch model has predicted for the size and ability level of the sample produce negative difficulty values. On the other hand, when the observed score exceeds the scored predicted by the model, item-difficulty is positive. Predicted performance on any one item is based upon performance of the sample on the other items. For example, if 80% of the group got all the other items in a test right, 80% of the group would be expected to get any given item right. This 80% ability level of the group, then, is set as the mean ability level of 0 (zero). Items exactly equal in difficulty to that ability level, therefore, have a mean difficulty level of 0. But, if a smaller proportion of the sample actually get an item right than their ability level would suggest should happen, the item-difficulty is negative. When more persons taking the test than expected get an item right, the item-difficulty is positive. The standard error is presented to the right of each item-difficulty estimate. The last column shows how much adjustment in an item difficulty estimate occurred between the last and the next to last iteration of the estimation procedure. This column provides an indicator of the stability which has been accumulated in the difficulty estimates. Small numbers in this column, suggest little difference between the last two estimates. When these differences become small enough, on aggregate, the estimation procedure is terminated.

The second part of the table is comprised of columns headed "RAW SCORE, SCALE ABILITY, and STANDARD ERROR" shows the estimates for person ability at each of the score levels possible in the test. Ability estimates complement item difficulty estimates and are also developed independently of item-difficulty by the UNCON procedure. All possible raw score levels are shown in descending order. Corresponding to each score level is an estimated ability level in logits, accompanied by the standard error of that estimate. The mean and standard deviation of group ability is shown at the bottom of this table: "MEAN ABILITY = 1.34" and "SD OF ABILITY = 1.50".

The third part of this table is an ogive labelled "TEST CHARACTERISTIC CURVE". This is a graphic representation of the ability estimates from -1.77 to 4.83 logits. It portrays the range of ability over scores ranging from 16 to 94 on this test.

Pages 7 and 8 comprise one table with three parts: "ITEM CHARACTERISTIC CURVE, DEPARTURE FROM EXPECTED ICC, and FIT Z-SQUARED". Items, by sequence number and name, are shown in the two leftmost columns of the first part. The entries correspond to the items, in all three parts, under the column headings 1ST GROUP, 2ND GROUP, 3RD GROUP, FOURTH GROUP, 5TH GROUP, and 6TH GROUP".

BICAL will estimate from two to six score level (i.e., ability) groups. Each group will have approximately the same number of persons, but there may be considerable variation in score range between groups. In this example run, the program created six groups ranging in size from 150 persons to 184 persons. The range of scores in the first group is 22 points; the second group 25 points; the third group 10 points; the fourth group 7 points; the fifth group 4 points; and the sixth group 5 points. These ability groups provide a means of comparing different aspects of item performance between different levels of ability. The three parts of this table show the development of three facets of the "item fit statistic" which will be used to determine whether or not an item fits the Rasch model and should be retained as a legitimate part of the test or thrown out because it does not fit the model.

The first part of the table (i.e., the ITEM CHARACTERISTIC CURVE) presents the proportion, within each ability group of the students who are in that ability group, which actually get each item right. For example, 30% of the students in the first group got item QA1 right while 99% of the students in the sixth group got it right. These values, from each ability group, should correspond approximately to the item characteristic curve at the respective ability levels represented by each group.

The second part of the table (i.e., the DEPARTURE FROM EXPECTED ICC) presents the result of subtracting the proportion of students in an ability group which the model predicts will get an item right from the proportion actually observed getting it right. Negative values indicate that the predicted value was larger than the observed value. These differences are "item residuals". Positive residuals result when observed proportions are larger than predicted proportions. These residuals, from each ability group, represent the extent of departure from the item characteristic curve at the respective ability level represented by each group. These item residual values constitute the basis for the fit statistic desired. At this point, they are comparable to deviation from mean values in that they constitute measures of variance and have both positive and negative values.

The third part of the table (i.e., the FIT Z-SQUARED) presents these item residual values in standardized form. This final step is accomplished by squaring each residual and dividing the result by the standard deviation of all of the squared residuals across every item within the group. This statistic has a number of the characteristics of the Z-score or Z-statistic. The similarity in method used to compute the statistic is quite evident for one thing. Therefore this statistic is labelled in the printout "FIT Z-SQUARED". Despite evident similarities to the Z-statistic, however, this statistic should not be interpreted like a Z-score. In fact, no effort is made in this study to

interpret this statistic by any standard. Application of any item fit statistic is subject to considerable discussion. It is the one aspect of Rasch analysis in greatest need of definition at this time. It is not the objective of this research to add to that definition, but to apply Rasch analysis in its present state of development to a specific situation. The statistic to be applied here is the average FIT Z-SQUARED computed, as one might expect, by summing the individual FIT Z-SQUARED values across the six ability groups and dividing that sum by the number (i.e., 6) of ability groups. This statistic is found under the last column of part three of this table, captioned "FIT MN. SQ."

Pages 9, 10, and 11 comprise one table with three parts: "SERIAL ORDER, DIFFICULTY ORDER, and FIT ORDER". Items, by sequence number and name, are shown in the two leftmost columns of the first part. Items, identified by sequence number and name, are shown in descending item difficulty order in part two. Items, identified by sequence number and name, are shown in descending fit mean square order in part three. The entries, in all three parts, correspond to the items under the column heading "ITEM DIFF, DISC INDEX, and FIT MN SQ".

BICAL presents items in three ways in this table to facilitate item analysis. Each of the three parts of this table includes the fit mean square statistic. The third part is the most useful table in this analysis. The fit mean square in this table lends itself directly to application of the fit statistic because it is arranged in ascending order. It is a simple matter, once the critical value for the fit statistic has been determined, to isolate those fit mean square values which are larger.

Pages 12 through 15 each present a plot. They are, respectively, "ITEM Z SQUARE (Y) VERSUS PROB (RIGHT) (X), FIT MEAN SQUARE (Y) VERSUS DIFFICULTY (X), FIT MEAN SQUARE VERSUS DISCRIMINATION (X), and DISCRIMINATION (Y) VERSUS DIFFICULTY (X)". The plotting symbol is the item sequence number on all four plots. Since the program apparently was not designed to plot sequence numbers larger than 99, items with sequence numbers from 100 to 115 do not appear intelligibly on these plots. Various combinations of 0 (zero) and special characters, or a blank, represent sequence numbers in this range. The reason for the symbols results from the fact that the computer is misinterpreting these three digit numbers, because insufficient space has been allotted to print them properly, and printing what symbol it "thinks" apply. This is one of the more graphic examples, and consequences, of the assumption implicit in the program's design which is that tests presented to it would be no larger than 99 items. These plots tend to be rather "busy" in the runs done in this study. For example, the plot of ITEM Z SCORE (Y) VERSUS PROB (RIGHT) (X) on page 12 attempts to present 690 sequence numbers (i.e., 115 items by 6 ability groups), while at the same time the program is unable to properly represent sequence numbers over 99! Digits are run together frequently so that it is difficult, at best, to determine if the numbers represent one, two, or three digit sequence number. Consequently, this plot is impossible to interpret. The next three plots attempt to represent only one series of sequence numbers from 1 to 115 each. All fail to represent sequence numbers over 99 properly. However they are interpretable. Nevertheless, none of these

plots contributed any significant insight to the objectives of this investigation. They are intended as a visual representation of various relationships which might be elements in the final determination of whether or not to keep an item which does not fit the Rasch model. The fact that an item may not fit the model should be tempered by the possibility that outside factors, such as guessing, insufficient time, and the like, may cause an item not to fit the model. However the intelligent use of these plots is very difficult to imagine in the absence of a thorough grounding in their use and interpretation. The limited amount of documentation that is available on the subject describes their use in connection with a limited example which lacks any of the sample size, score distribution, or test design considerations relevant to the data used in this investigation.

There is no intent in this investigation to do further analysis of item fit beyond the point of determining that the item's fit mean square has exceeded the critical value. This fact is taken in this investigation as sufficient reason for rejecting the item. That is, item fit to the Rasch model has been determined entirely on the strength of the fit statistic used in this investigation. This study is not concerned with examining possibilities which may have caused an item not to fit. Consequently, these plots have little more than passing interest here. They have not been used in any way to evaluate item fit, or for any other purpose, in this investigation.

#### THE FIT STATISTIC

APPENDIX C: THE ITEM FIT STATISTIC - EVOLUTIONARY CHANGES OF INTERPRETATION, focuses on the item fit statistic. Rasch measurement seems to promise truly objective measurement, but problems of interpretation associated with the decision that an item does, or does not, fit the Rasch model could withhold that promise. The issues encountered during attempts to define and use the "item fit statistic" in this research raised some important questions about the objectivity of Rasch measurement which are fundamental to the practical application of this tool in test measurement. While Rasch measurement theory portrays an objective test measurement tool, Rasch measurement application may entail too many subjective elements to make this possible in a practical sense. Interpretation of an item fit statistic has changed over the years. This has made it a bit difficult to comprehend for the purposes of this investigation. Conviction as to the appropriateness of a statistic in determining item fit to the Rasch model, independent of subjective considerations, seems to have softened over the years as well. Objectivity in test measurement may yet be possible, but the tendency seems to be growing to tack subjective elements of interpretation to the use of an item fit statistic, or at least increase the complexity of using the statistic, so that the result is an impractical measure. Before proceeding with this study, considerable time and effort was devoted to gaining a working understanding of an item fit statistic. Like the application of the BICAL computer program at the early stages of this study, the more understanding was sought, the more elusive the concept of an item fit statistic became. The concept is undergoing change. That is why, most probably, it seemed so hard to pin down at first. APPENDIX C attempts

to give some perspective to the evolutionary change of the item fit statistic concept. Once understood, users will have to decide for themselves what emphasis should be placed on a purely statistical interpretation of item fit compared to some combination of statistical and subjective considerations.





FITTING 1973-1978 MEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL

ALTERNATIVE RESPONSE FREQUENCIES

SEQ ITEM NUM NAME	0	1	2	3	4	UNKN	KEY
1 OA1	1	170	830	0	0	0	1
2 OA2	1	237	763	0	0	0	1
3 OA3	1	370	630	0	0	0	1
4 OA4	1	245	755	0	0	0	1
5 OA5	1	301	699	0	0	0	1
6 OA6	1	256	744	0	0	0	1
7 OA7	1	278	722	0	0	0	1
8 OA8	1	184	806	0	0	0	1
9 OA9	1	278	722	0	0	0	1
10 OA10	1	252	748	0	0	0	1
11 OC1	1	1000	0	0	0	0	1
12 OC2	1	1000	0	0	0	0	1
13 OC3	1	1000	0	0	0	0	1
14 OC4	1	1000	0	0	0	0	1
15 OC5	1	1000	0	0	0	0	1
16 OC6	1	188	812	0	0	0	1
17 OC7	1	234	766	0	0	0	1
18 OC8	1	203	797	0	0	0	1
19 OC9	1	286	704	0	0	0	1
20 OC10	1	266	734	0	0	0	1
21 OE1	1	96	904	0	0	0	1
22 OE2	1	214	786	0	0	0	1
23 OE3	1	184	816	0	0	0	1
24 OE4	1	121	869	0	0	0	1
25 OE5	1	128	871	0	0	0	1
26 OF1	1	228	776	0	0	0	1
27 OF2	1	367	633	0	0	0	1
28 OF3	1	406	594	0	0	0	1
29 OF4	1	370	680	0	0	0	1
30 OF5	1	280	720	0	0	0	1
31 OH1	1	182	808	0	0	0	1
32 OH2	1	180	820	0	0	0	1
33 OH3	1	139	865	0	0	0	1
34 OH4	1	133	867	0	0	0	1
35 OH5	1	194	806	0	0	0	1
36 OH6	1	195	804	0	0	0	1
37 OH7	1	183	817	0	0	0	1
38 OH8	1	228	772	0	0	0	1
39 OH9	1	166	834	0	0	0	1
40 OH10	1	284	716	0	0	0	1
41 OI1	1	232	768	0	0	0	1
42 OI2	1	215	785	0	0	0	1
43 OI3	1	248	752	0	0	0	1
44 OI4	1	283	617	0	0	0	1
45 OI5	1	188	812	0	0	0	1
46 OI6	1	244	756	0	0	0	1
47 OI7	1	237	763	0	0	0	1
48 OI8	1	558	441	0	0	0	1
49 OI9	1	395	605	0	0	0	1
50 OI10	1	321	679	0	0	0	1
51 OK1	1	362	618	0	0	0	1
52 OK2	1	362	618	0	0	0	1





FITTING 1973-1979 NEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL

PAGE 3

NUMBER OF ZERO SCORES 2  
 NUMBER OF PERFECT SCORES 0

NUMBER OF ITEMS SELECTED 118  
 NUMBER OF ITEMS NAMED 118

-----  
 SUBJECTS BELOW 19 10  
 SUBJECTS ABOVE 114 0  
 SUBJECTS IN CALIB. 898  
 -----  
 TOTAL SUBJECTS 898

REJECTED ITEMS

ITEM NUMBER	ITEM NAME	ANSWERED CORRECTLY	
11	QC1	0	LOW SCORE
12	QC2	0	LOW SCORE
13	QC3	0	LOW SCORE
14	QC4	0	LOW SCORE
15	QC5	0	LOW SCORE
66	QN1	0	LOW SCORE
67	QN2	0	LOW SCORE
68	QN3	0	LOW SCORE
69	QN4	0	LOW SCORE
70	QN5	0	LOW SCORE
71	QO1	0	LOW SCORE
72	QO2	0	LOW SCORE
73	QO3	0	LOW SCORE
74	QO4	0	LOW SCORE
75	QO5	0	LOW SCORE
**	QW1	0	LOW SCORE
**	QW2	0	LOW SCORE
**	QW3	0	LOW SCORE
**	QW4	0	LOW SCORE
**	QW5	0	LOW SCORE

SUBJECTS DELETED = 2  
 SUBJECTS REMAINING = 897

ITEMS DELETED = 20  
 POSSIBLE SCORE = 88

MINIMUM SCORE = 19  
 MAXIMUM SCORE = 94

FITTING 1973-1978 NEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL  
 SCORE DISTRIBUTION OF ABILITY

SCORE	COUNT	PROPORTION	2	4	5	6	10
1	0	0.00	1				
2	0	0.00	1				
3	0	0.00	1				
4	0	0.00	1				
5	0	0.00	1				
6	0	0.00	1				
7	1	0.00	1				
8	0	0.00	1				
9	0	0.00	1				
10	0	0.00	1				
11	0	0.00	1				
12	0	0.00	1				
13	0	0.00	1				
14	1	0.00	1				
15	1	0.00	1				
16	0	0.00	1				
17	0	0.00	1				
18	0	0.00	1				
19	3	0.00	1				
20	4	0.00	1				
21	4	0.01	1				
22	3	0.00	1				
23	10	0.01	1				
24	12	0.01	1				
25	3	0.00	1				
26	11	0.01	1				
27	10	0.01	1				
28	10	0.01	1				
29	15	0.02	1				
30	7	0.01	1				
31	13	0.01	1				
32	6	0.01	1				
33	11	0.01	1				
34	8	0.01	1				
35	8	0.01	1				
36	8	0.01	1				
37	10	0.01	1				
38	6	0.01	1				
39	9	0.01	1				
40	8	0.01	1				
41	8	0.01	1				
42	3	0.00	1				
43	3	0.00	1				
44	7	0.01	1				
45	6	0.01	1				
46	7	0.00	1				
47	7	0.01	1				
48	6	0.01	1				
49	8	0.01	1				
50	5	0.01	1				
51	6	0.01	1				
52	7	0.01	1				
53	5	0.01	1				
54	4	0.00	1				

55	10	0.01	XXXXXXXXXX	1
56	7	0.01	XXXXXXX	1
57	7	0.01	XXXXXXXXXX	1
58	6	0.01	XXXXXXX	1
59	3	0.00	XXX	1
60	5	0.01	XXXXX	1
61	6	0.01	XXXXXXX	1
62	6	0.01	XXXXXXX	1
63	9	0.01	XXXXXXXXXX	1
64	9	0.01	XXXXXXXXXX	1
65	12	0.01	XXXXXXXXXXXX	1
66	16	0.02	XXXXXXXXXXXXXXXX	1
67	13	0.01	XXXXXXXXXXXX	1
68	15	0.02	XXXXXXXXXXXXXXXX	1
69	7	0.01	XXXXXXX	1
70	15	0.02	XXXXXXXXXXXXXXXX	1
71	14	0.01	XXXXXXXXXXXX	1
72	12	0.01	XXXXXXXXXXXX	1
73	15	0.02	XXXXXXXXXXXX	1
74	19	0.02	XXXXXXXXXXXXXXXX	1
75	22	0.02	XXXXXXXXXXXXXXXX	1
76	11	0.01	XXXXXXXXXX	1
77	16	0.02	XXXXXXXXXXXX	1
78	25	0.03	XXXXXXXXXXXXXXXX	1
79	15	0.02	XXXXXXXXXXXX	1
80	24	0.02	XXXXXXXXXXXXXXXX	1
81	23	0.02	XXXXXXXXXXXX	1
82	28	0.03	XXXXXXXXXXXXXXXX	1
83	42	0.04	XXXXXXXXXXXXXXXX	1
84	31	0.03	XXXXXXXXXXXXXXXX	1
85	26	0.03	XXXXXXXXXXXX	1
86	34	0.03	XXXXXXXXXXXXXXXX	1
87	41	0.04	XXXXXXXXXXXXXXXX	1
88	34	0.03	XXXXXXXXXXXXXXXX	1
89	31	0.03	XXXXXXXXXXXX	1
90	32	0.03	XXXXXXXXXXXX	1
91	34	0.03	XXXXXXXXXXXXXXXX	1
92	28	0.03	XXXXXXXXXXXX	1
93	15	0.02	XXXXXXXXXXXX	1
94	10	0.01	XXXXXXX	1
95	1	0.00	X	1

-----  
FULL SCALE = 0.04  
-----

FITTING 1973-1978 NEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL  
 DISTRIBUTION OF EASINESS

ITEM	COUNT	PROPORTION	1	2	4	6	8	10
1	828	0.84	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
2	762	0.77	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
3	627	0.64	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
4	753	0.76	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
5	698	0.71	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
6	744	0.75	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
7	721	0.73	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
8	804	0.81	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
9	721	0.73	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
10	748	0.76	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
16	808	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
17	764	0.77	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
18	796	0.81	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
19	703	0.71	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
20	731	0.74	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
21	898	0.91	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
22	782	0.79	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
23	814	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
24	864	0.88	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
25	869	0.88	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
26	772	0.78	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
27	632	0.64	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
28	591	0.60	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
29	678	0.69	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
30	717	0.73	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
31	805	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
32	817	0.83	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
33	876	0.89	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
34	861	0.87	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
35	865	0.88	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
36	805	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
37	804	0.81	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
38	816	0.83	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
39	771	0.78	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
40	834	0.84	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
41	715	0.72	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
42	766	0.78	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
43	784	0.79	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
44	752	0.76	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
45	617	0.63	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
46	809	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
47	756	0.77	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
48	763	0.77	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
49	441	0.45	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
50	603	0.61	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
51	676	0.68	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
52	815	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
53	710	0.72	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
54	689	0.70	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
55	612	0.62	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
56	695	0.70	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
57	460	0.47	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
58	634	0.64	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
59	537	0.54	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					



60	485	0.49	XXXXXXXXXXXXXXXXXXXX	
61	745	0.75	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
62	767	0.78	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
63	549	0.56	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
64	625	0.63	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
65	464	0.47	XXXXXXXXXXXXXXXXXXXX	
76	825	0.84	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
77	809	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
78	743	0.75	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
79	637	0.65	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
80	647	0.66	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
81	603	0.61	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
82	616	0.62	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
83	609	0.62	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
84	634	0.64	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
85	695	0.70	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
86	738	0.75	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
87	825	0.84	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
88	727	0.74	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
89	661	0.67	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
90	698	0.71	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
91	606	0.61	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
92	709	0.72	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
93	567	0.57	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
94	608	0.62	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
95	694	0.70	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
96	539	0.55	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
97	672	0.68	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
98	820	0.83	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
99	436	0.44	XXXXXXXXXXXXXXXXXXXX	
100	802	0.81	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
101	733	0.74	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
102	826	0.84	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
103	552	0.56	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
104	735	0.74	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
105	814	0.82	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
106	735	0.74	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
107	799	0.81	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
108	652	0.66	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
109	582	0.59	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
110	594	0.60	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	

-----  
FULL SCALE = 1.00  
-----

FITTING 1973-1978 NEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL

PROCEDURE USED UCOM  
 NUMBER OF ITERATIONS = 3

SEQUENCE NUMBER	ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR	LAST DIFF CHANGE	RAW SCORE	SCALE ABILITY	STANDARD ERROR	TEST CHARACTERISTIC CURVE
1	QA1	-0.902	0.088	-0.004	94	4.83	1.01	
2	QA2	-0.331	0.089	-0.002	93	4.12	0.72	
3	QA3	0.610	0.080	0.003	92	3.69	0.59	
4	QA4	-0.261	0.088	-0.001	91	3.39	0.52	
5	QA5	0.141	0.084	0.001	90	3.15	0.47	
6	QB1	-0.192	0.087	-0.001	90	3.15	0.47	
7	QB2	-0.022	0.085	-0.000	89	2.94	0.43	
8	QB3	-0.681	0.094	-0.004	88	2.77	0.40	
9	QB4	-0.022	0.085	-0.000	87	2.62	0.38	
10	QB5	-0.223	0.086	-0.001	86	2.48	0.36	
16	QO1	-0.726	0.095	-0.004	85	2.36	0.34	
17	QO2	-0.347	0.089	-0.002	85	2.36	0.34	
18	QO3	-0.611	0.093	-0.003	84	2.25	0.33	
19	QO4	0.106	0.084	0.000	83	2.14	0.32	
20	QO5	-0.095	0.086	-0.001	82	2.04	0.31	
21	QE1	-1.711	0.120	-0.007	81	1.95	0.30	
22	QE2	-0.493	0.091	-0.003	80	1.86	0.29	
23	QE3	-0.771	0.096	-0.004	80	1.86	0.29	
24	QE4	-1.278	0.107	-0.005	79	1.78	0.29	
25	QE5	-1.336	0.108	-0.006	78	1.70	0.28	
26	QF1	-0.411	0.090	-0.002	77	1.62	0.27	
27	QF2	0.578	0.080	0.003	76	1.55	0.27	
28	QF3	0.833	0.079	0.004	75	1.48	0.26	
29	QF4	0.277	0.082	0.001	75	1.48	0.26	
30	QF5	0.006	0.085	-0.000	74	1.41	0.26	
31	QG1	-0.690	0.094	-0.004	73	1.34	0.26	
32	QG2	-0.799	0.096	-0.004	72	1.28	0.25	
33	QG3	-1.420	0.111	-0.006	71	1.21	0.25	
34	QG4	-1.244	0.106	-0.005	71	1.21	0.25	
35	QG5	-1.289	0.107	-0.006	70	1.15	0.25	
36	QH1	-0.690	0.094	-0.004	69	1.09	0.24	
37	QH2	-0.681	0.094	-0.004	68	1.04	0.24	
38	QH3	-0.789	0.096	-0.004	67	0.98	0.24	
39	QH4	-0.403	0.090	-0.002	66	0.92	0.24	
40	QH5	-0.961	0.100	-0.004	66	0.92	0.24	
41	QI1	0.021	0.085	0.000	65	0.87	0.23	
42	QI2	-0.363	0.090	-0.002	64	0.81	0.23	
43	QI3	-0.509	0.092	-0.003	63	0.76	0.23	
44	QI4	-0.253	0.088	-0.001	62	0.71	0.23	
45	QI5	0.673	0.079	0.003	61	0.65	0.23	
46	QJ1	-0.726	0.095	-0.004	61	0.65	0.23	
47	QJ2	-0.284	0.088	-0.002	60	0.60	0.23	
48	QJ3	-0.339	0.089	-0.002	58	0.55	0.22	
49	QJ4	1.713	0.076	0.008	58	0.50	0.22	
50	QJ5	0.759	0.079	0.004	57	0.45	0.22	
51	QK1	0.291	0.082	0.001	56	0.40	0.22	
52	QK2	0.685	0.079	0.003	56	0.40	0.22	
53	QK3	0.056	0.085	0.000	55	0.35	0.22	
54	QK4	0.203	0.083	0.001	54	0.31	0.22	
55	QK5	0.704	0.079	0.003	53	0.26	0.22	
56	QL1	0.161	0.084	0.001	52	0.21	0.22	







FITTING 1973-1978 MEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL

SERIAL ORDER      DIFFICULTY ORDER      FIT ORDER

SEQ ITEM    ITEM    DISC    INDX    FIT    SEQ ITEM    ITEM    DIFF    INDX    DISC    FIT    POINT  
 NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    NUM NAME    BISER

1 Q41	1.31	8.44	1	21 Q21	-1.71	1.04	2	Q2	11	Q21	0.0	0.0
2 Q42	-0.33	1.82	1	33 Q23	-1.42	1.19	2	25	13	Q23	0.0	0.0
3 Q43	0.61	0.79	1	25 Q25	5.88	1.34	1	33	13	Q25	0.0	0.0
4 Q44	-0.26	1.31	1	35 Q25	8.01	1.29	1	35	13	Q25	0.0	0.0
5 Q45	0.14	1.28	1	24 Q24	3.84	1.28	1	24	13	Q24	0.0	0.0
6 Q46	-0.18	1.30	1	34 Q24	8.90	1.24	1	34	13	Q24	0.0	0.0
7 Q47	-0.02	1.17	1	40 Q45	3.13	1.18	1	40	13	Q45	0.0	0.0
8 Q48	-0.68	1.36	1	12 Q20	12.20	1.20	1	12	13	Q20	0.0	0.0
9 Q49	-0.68	1.24	1	102 Q42	5.61	1.20	1	102	13	Q42	0.0	0.0
10 Q495	-0.02	1.24	1	102 Q42	5.61	1.20	1	102	13	Q42	0.0	0.0
11 Q50	0.0	0.0	1	76 Q41	0.0	1.20	1	76	13	Q41	0.0	0.0
12 Q50	0.0	0.0	1	88 Q43	0.0	1.08	1	88	13	Q43	0.0	0.0
13 Q50	0.0	0.0	1	32 Q22	0.0	1.08	1	32	13	Q22	0.0	0.0
14 Q50	0.0	0.0	1	38 Q43	0.0	1.25	1	38	13	Q43	0.0	0.0
15 Q50	0.0	0.0	1	23 Q23	0.0	1.19	1	23	13	Q23	0.0	0.0
16 Q50	0.0	0.0	1	105 Q45	4.53	1.20	1	105	13	Q45	0.0	0.0
17 Q50	-0.35	1.21	1	16 Q01	4.71	1.18	1	16	13	Q01	0.0	0.0
18 Q50	-0.61	1.20	1	77 Q42	3.25	1.21	1	77	13	Q42	0.0	0.0
19 Q50	0.11	1.11	1	46 Q41	1.65	1.09	1	46	13	Q41	0.0	0.0
20 Q50	-0.10	1.11	1	31 Q41	1.22	1.04	1	31	13	Q41	0.0	0.0
21 Q50	-1.71	1.04	1	36 Q41	2.03	1.22	1	36	13	Q41	0.0	0.0
22 Q50	-0.48	0.81	1	37 Q42	5.74	1.22	1	37	13	Q42	0.0	0.0
23 Q50	-0.77	1.19	1	8 Q43	4.20	1.22	1	8	13	Q43	0.0	0.0
24 Q50	-1.34	1.14	1	107 Q42	2.25	1.16	1	107	13	Q42	0.0	0.0
25 Q50	-1.04	0.90	1	43 Q43	22.82	1.14	1	43	13	Q43	0.0	0.0
26 Q50	-0.41	0.76	1	18 Q03	10.43	1.20	1	18	13	Q03	0.0	0.0
27 Q50	0.58	0.50	1	43 Q43	22.82	1.14	1	43	13	Q43	0.0	0.0
28 Q50	0.83	0.50	1	22 Q22	18.82	0.81	1	22	13	Q22	0.0	0.0
29 Q50	0.28	0.71	1	26 Q41	9.08	0.76	1	26	13	Q41	0.0	0.0
30 Q50	0.01	0.72	1	39 Q44	8.33	1.26	1	39	13	Q44	0.0	0.0
31 Q50	-0.69	1.04	1	62 Q42	1.30	1.09	1	62	13	Q42	0.0	0.0
32 Q50	-0.80	1.08	1	42 Q42	2.34	1.11	1	42	13	Q42	0.0	0.0
33 Q50	-1.24	1.08	1	48 Q43	3.31	1.19	1	48	13	Q43	0.0	0.0
34 Q50	-1.42	1.19	1	17 Q02	7.35	1.21	1	17	13	Q02	0.0	0.0
35 Q50	-1.24	1.08	1	48 Q43	3.31	1.19	1	48	13	Q43	0.0	0.0
36 Q50	-1.29	1.22	1	47 Q42	6.13	1.22	1	47	13	Q42	0.0	0.0
37 Q50	-0.68	1.29	1	4 Q44	7.87	1.31	1	4	13	Q44	0.0	0.0
38 Q50	-0.75	1.25	1	44 Q44	9.74	1.30	1	44	13	Q44	0.0	0.0
39 Q50	-0.40	1.26	1	10 Q45	5.94	1.19	1	10	13	Q45	0.0	0.0
40 Q50	-0.88	1.18	1	61 Q41	3.57	1.15	1	61	13	Q41	0.0	0.0
41 Q50	0.02	1.17	1	3 Q24	3.24	1.30	1	3	13	Q24	0.0	0.0
42 Q50	-0.36	1.24	1	78 Q43	4.51	1.13	1	78	13	Q43	0.0	0.0
43 Q50	-0.25	1.14	1	86 Q41	2.76	1.15	1	86	13	Q41	0.0	0.0
44 Q50	-0.51	1.30	1	106 Q41	8.07	0.80	1	106	13	Q41	0.0	0.0
45 Q50	0.67	1.08	1	101 Q41	1.56	1.15	1	101	13	Q41	0.0	0.0
46 Q50	-0.73	1.08	1	101 Q41	1.56	1.15	1	101	13	Q41	0.0	0.0
47 Q50	-0.28	0.89	1	20 Q05	2.17	1.11	1	20	13	Q05	0.0	0.0
48 Q50	-0.34	1.18	1	88 Q43	3.65	1.19	1	88	13	Q43	0.0	0.0
49 Q50	-0.07	1.11	1	1.52	1.11	1.11	1	1.52	13	1.52	0.0	0.0
50 Q50	0.76	0.88	1	9 Q44	5.61	1.24	1	9	13	Q44	0.0	0.0

TABLE CONTINUED

SERIAL ORDER				DIFFICULTY ORDER				FIT ORDER							
SEQ ITEM	ITEM	DIFF	DISC	FIT	SEQ ITEM	ITEM	DIFF	DISC	FIT	SEQ ITEM	ITEM	DIFF	DISC	FIT	POINT
51 QK1	0.29	0.54	6.55	1	74 Q04	0.00	0.00	0.00	0.00	18 Q03	-0.61	1.20	1.08	3.25	1
52 QK2	0.59	0.88	20.31	1	75 Q05	0.00	0.00	0.00	0.00	19 Q04	-1.24	1.08	1.08	3.31	1
53 QK3	0.06	0.88	3.76	1	11 Q01	0.00	0.00	0.00	0.00	10 Q05	0.00	1.19	0.88	3.33	1
54 QK4	0.20	0.63	17.11	1	12 Q02	0.00	0.00	0.00	0.00	11 Q02	0.00	0.88	0.88	3.53	1
55 QK5	0.70	0.74	5.63	1	13 Q03	0.00	0.00	0.00	0.00	12 Q03	-0.96	1.18	0.88	3.57	1
56 QL1	0.16	0.90	1.20	1	14 Q04	0.00	0.00	0.00	0.00	13 Q04	0.67	1.11	0.88	3.60	1
57 QL2	1.60	0.62	12.65	1	15 Q05	0.00	0.00	0.00	0.00	14 Q05	-0.34	1.18	0.88	3.65	1
58 QL3	0.57	0.78	5.96	1	30 QF5	0.01	0.72	9.33	1	48 QJ3	0.00	1.11	0.88	3.76	1
59 QL4	1.18	1.07	5.88	1	66 QN1	0.00	0.00	0.00	0.00	45 QI5	0.67	1.11	0.88	3.76	1
60 QL5	1.46	0.72	16.58	1	67 QN2	0.00	0.00	0.00	0.00	5 QAS	0.14	1.28	0.88	3.84	1
61 QM1	-0.20	1.15	1.78	1	68 QN3	0.00	0.00	0.00	0.00	79 QP4	0.55	1.28	0.88	3.95	1
62 QM2	-0.37	1.08	1.28	1	69 QM4	0.00	0.00	0.00	0.00	24 QE4	-1.28	1.16	0.88	4.07	1
63 QM3	1.08	0.67	8.81	1	70 QM5	0.00	0.00	0.00	0.00	23 QE3	-0.77	1.19	0.88	4.20	1
64 QM4	0.62	1.02	1.62	1	71 QM1	0.00	0.00	0.00	0.00	16 QP1	-0.87	1.20	0.88	4.28	1
65 QM5	1.58	0.75	18.54	1	72 QM2	0.00	0.00	0.00	0.00	18 QP3	-0.18	1.13	0.88	4.51	1
66 QN1	0.00	0.00	0.00	1	73 QM3	0.00	0.00	0.00	0.00	16 QO1	-0.73	1.18	0.88	4.53	1
67 QN2	0.00	0.00	0.00	1	111 QW1	0.00	0.00	0.00	0.00	105 QU5	-0.77	1.20	0.88	4.59	1
68 QN3	0.00	0.00	0.00	1	112 QW2	0.00	0.00	0.00	0.00	17 QO2	-0.35	1.21	0.88	4.71	1
69 QN4	0.00	0.00	0.00	1	113 QW3	0.00	0.00	0.00	0.00	102 QU2	-0.88	1.20	0.88	4.73	1
70 QN5	0.00	0.00	0.00	1	114 QW4	0.00	0.00	0.00	0.00	100 QI5	-0.66	1.22	0.88	4.77	1
71 QO1	0.00	0.00	0.00	1	115 QW5	0.00	0.00	0.00	0.00	81 QO1	0.76	0.76	0.88	5.17	1
72 QO2	0.00	0.00	0.00	1	41 QI1	0.02	1.17	3.24	1	50 QJ5	0.76	1.15	0.88	5.20	1
73 QO3	0.00	0.00	0.00	1	53 QK3	0.06	0.97	2.62	1	8 Q04	-0.02	1.24	0.88	5.30	1
74 QO4	0.00	0.00	0.00	1	82 Q52	0.06	0.87	2.62	1	8 Q04	-0.02	1.24	0.88	5.30	1
75 QO5	0.00	0.00	0.00	1	5 QAS	0.14	1.28	3.84	1	22 QE2	-0.48	0.81	0.88	5.74	1
76 QP1	-0.87	1.20	4.28	1	5 QAS	0.14	1.28	3.84	1	22 QE2	-0.48	0.81	0.88	5.74	1
77 QP2	-0.73	1.21	3.76	1	80 QN5	0.14	1.14	1.88	1	38 QH3	-0.78	1.25	0.88	5.74	1
78 QP3	-0.18	1.13	4.51	1	98 QL1	0.18	0.90	1.20	1	59 QL4	1.16	1.07	0.88	5.88	1
79 QP4	0.55	1.28	3.95	1	95 Q55	0.17	1.31	1.10	1	3 QAS	0.61	1.07	0.88	5.88	1
80 QP5	-0.07	1.11	9.14	1	80 QP5	0.48	0.85	8.58	1	51 QK1	0.29	0.75	0.88	6.35	1
81 QO1	0.76	0.76	5.17	1	54 QM4	0.20	0.63	17.11	1	39 QM4	-0.40	1.26	0.88	6.44	1
82 QO2	0.68	0.87	1.26	1	29 QF4	0.28	0.71	9.08	1	58 QL3	0.57	0.78	0.88	6.44	1
83 QO3	0.72	1.20	7.18	1	51 QK1	0.28	0.71	9.08	1	58 QL3	0.57	0.78	0.88	6.44	1
84 QO4	0.57	0.78	5.99	1	84 QO4	0.55	0.75	6.55	1	84 QO4	0.57	0.78	0.88	6.44	1
85 QO5	0.57	0.78	5.99	1	9.08	0.71	9.08	0.71	9.08	0.57	0.78	0.88	6.44	1	
86 QR1	-0.15	1.15	1.72	1	108 QV3	0.45	0.81	1.54	1	56 QI1	1.14	0.72	0.88	6.44	1
87 QR2	-0.87	1.27	9.14	1	80 QP5	0.48	0.85	8.58	1	51 QK1	0.29	0.75	0.88	6.35	1
88 QR3	-0.07	1.11	1.52	1	79 QP4	0.55	1.28	3.95	1	85 Q55	0.17	1.31	0.88	6.35	1
89 QR4	0.39	1.28	6.11	1	84 QO4	0.57	0.78	5.99	1	83 QO3	0.72	1.20	0.88	6.35	1
90 QR5	0.14	1.14	1.88	1	58 QL3	0.55	0.75	6.55	1	37 QG3	-1.42	1.19	0.88	6.35	1
91 QS1	0.74	0.65	10.56	1	27 QF2	0.58	0.50	22.82	1	37 QG3	-1.42	1.19	0.88	6.35	1
92 QS2	0.06	0.87	2.62	1	3 QAS	0.61	0.81	5.88	1	44 QI4	-0.25	1.30	0.88	6.07	1
93 QS3	0.88	0.77	13.35	1	64 QM4	0.62	1.02	1.62	1	80 QP5	0.48	0.86	0.88	6.58	1
94 QS4	0.72	5.30	5.30	1	45 QI5	0.67	1.11	3.60	1	63 QM3	1.09	0.67	0.88	6.81	1
95 QS5	0.17	1.31	7.10	1	82 QO2	0.68	1.26	1.26	1	4 QAS	-0.26	1.31	0.88	6.81	1
96 QT1	1.14	0.72	6.44	1	52 QK2	0.69	0.54	20.31	1	29 QF4	0.28	0.71	0.88	6.08	1
97 QT2	0.32	3.14	3.14	1	55 QK5	0.70	0.74	5.63	1	87 QR2	-0.87	1.27	0.88	6.14	1
98 QT3	-0.83	1.32	11.08	1	83 QO3	0.72	1.20	7.18	1	30 QF5	0.01	0.71	0.88	6.23	1
99 QT4	1.74	0.88	19.17	1	84 Q54	0.73	1.17	5.30	1	1 QAS	-0.90	1.31	0.88	6.44	1
100 QT5	-0.66	1.22	4.77	1	91 Q51	0.74	0.74	10.56	1	109 QV4	-0.89	0.69	0.88	6.69	1
101 QU1	-0.11	1.15	1.93	1	50 QJ5	0.76	1.15	5.20	1	110 QV5	0.81	1.07	0.88	6.70	1

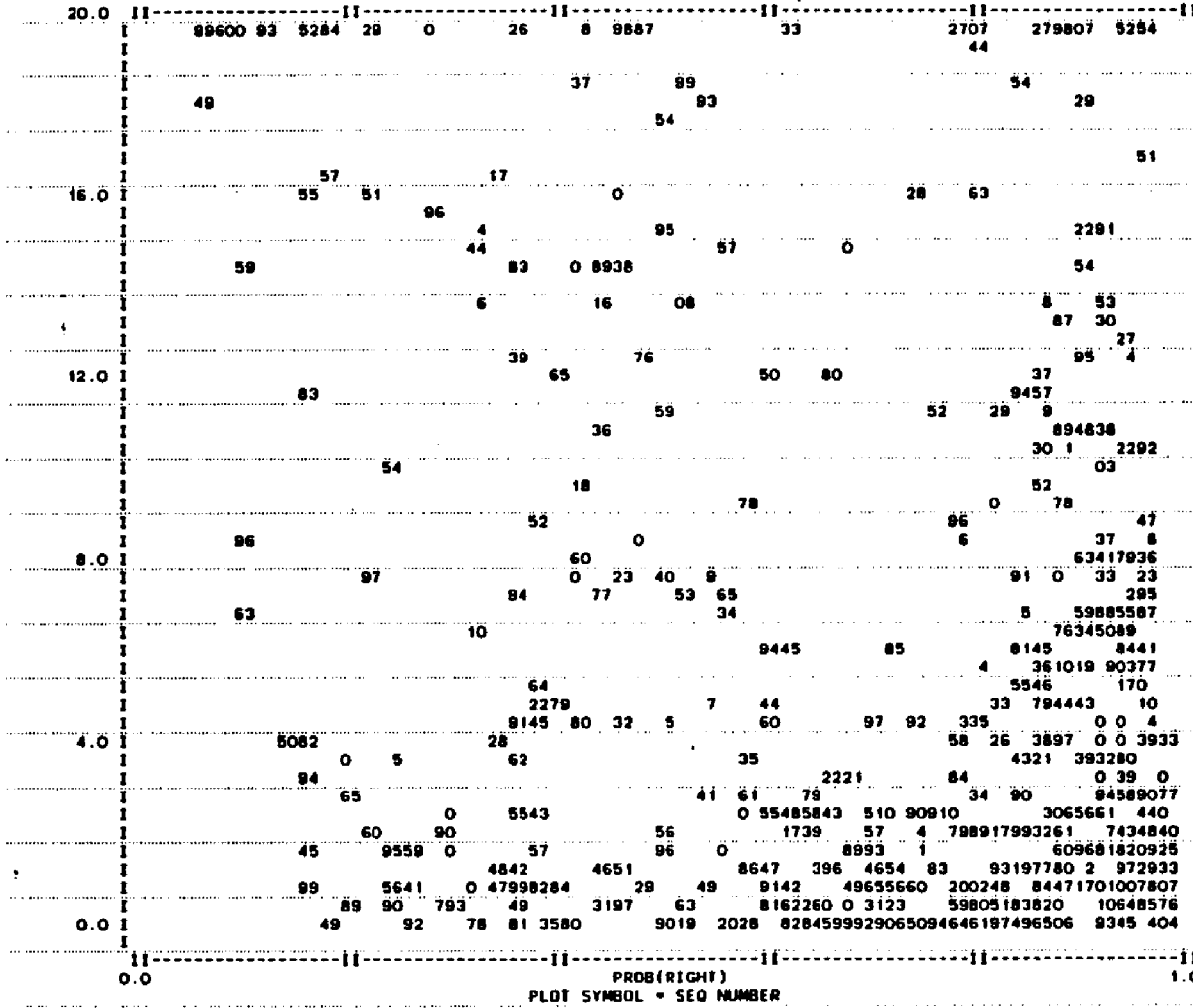
TABLE CONTINUED



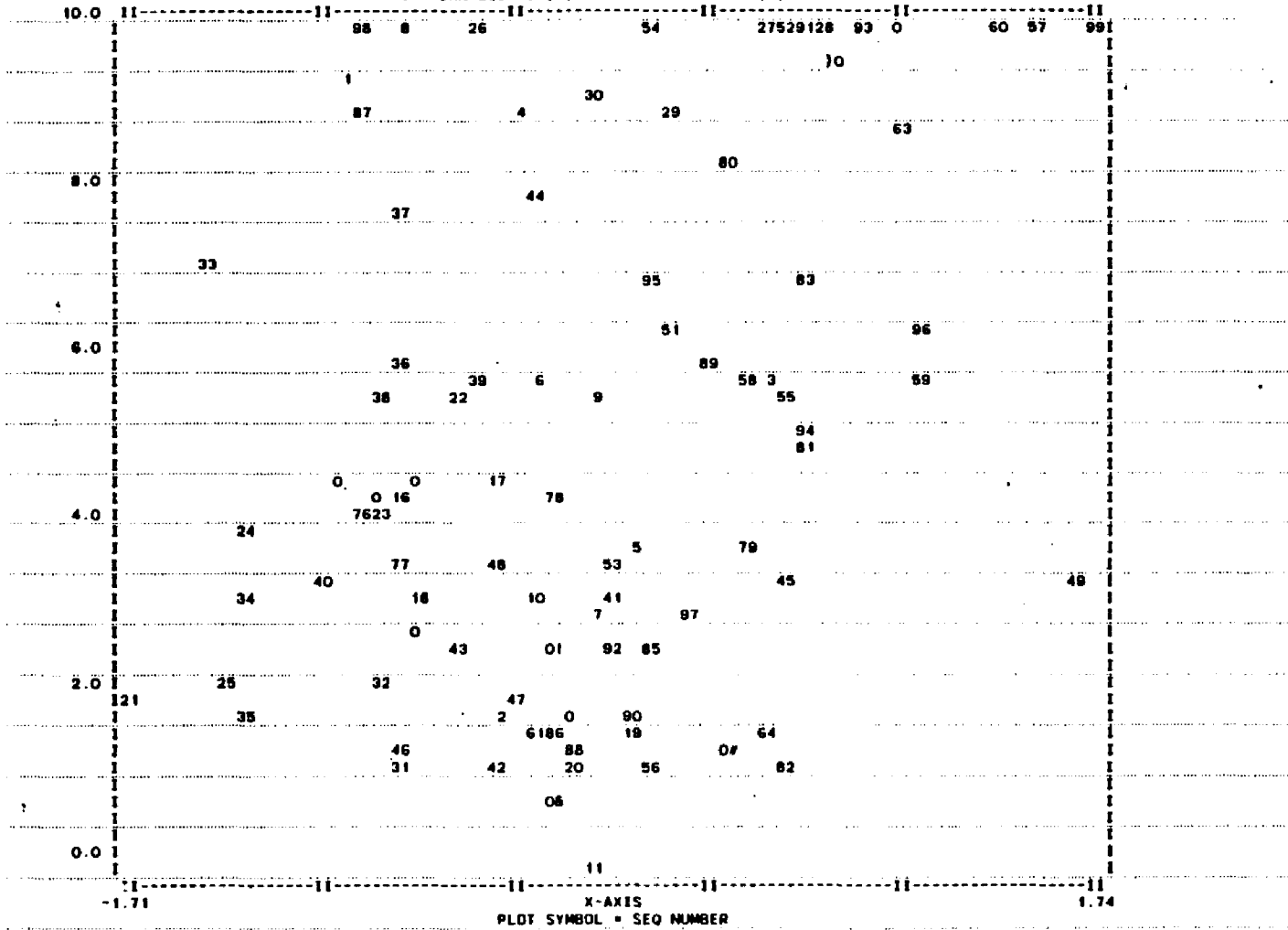


FITTING 1973-1979 MEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL  
 [ITEM Z SQUARE (Y) VERSUS PRPB(RIGHT)](X)

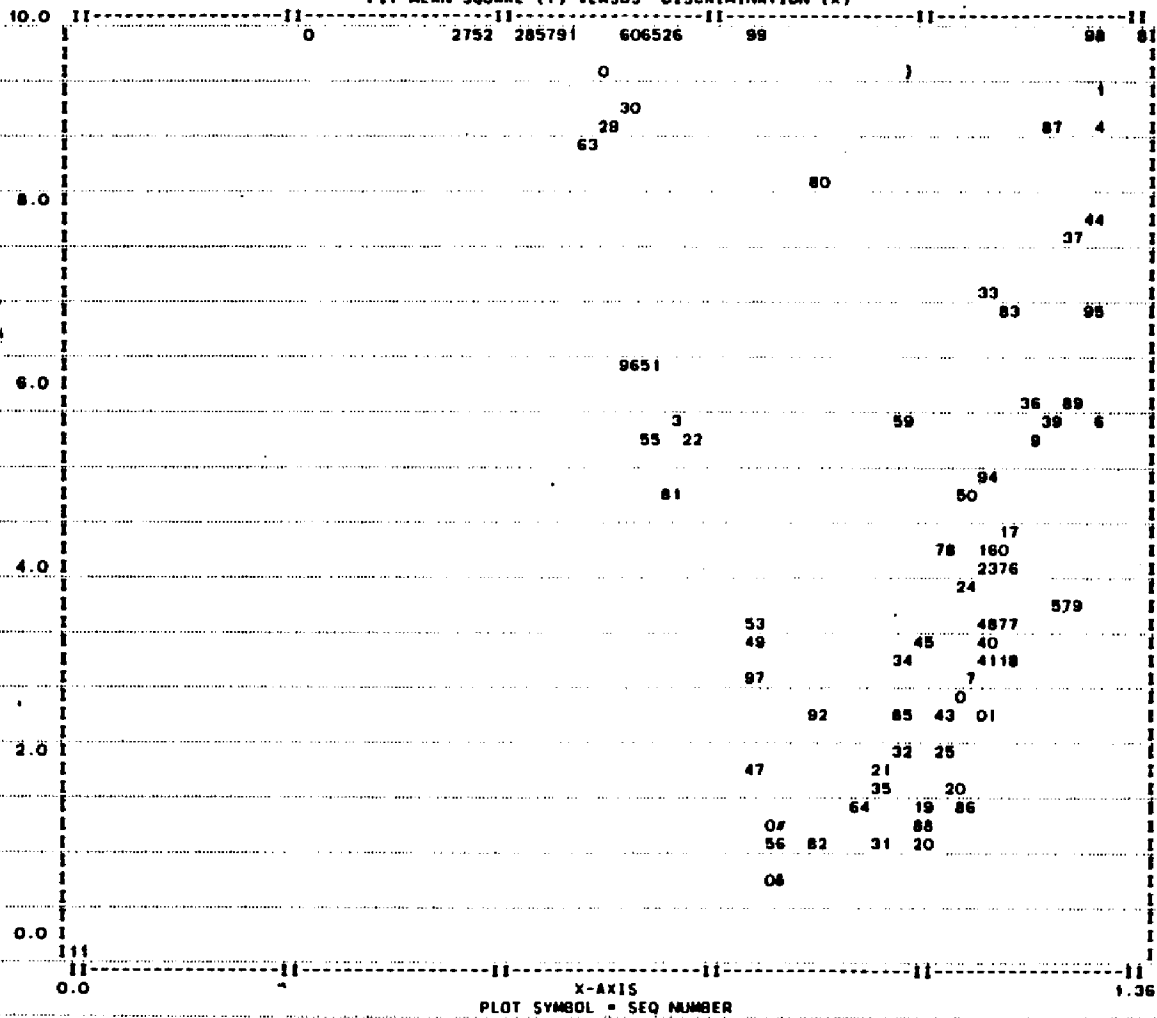
PAGE 12



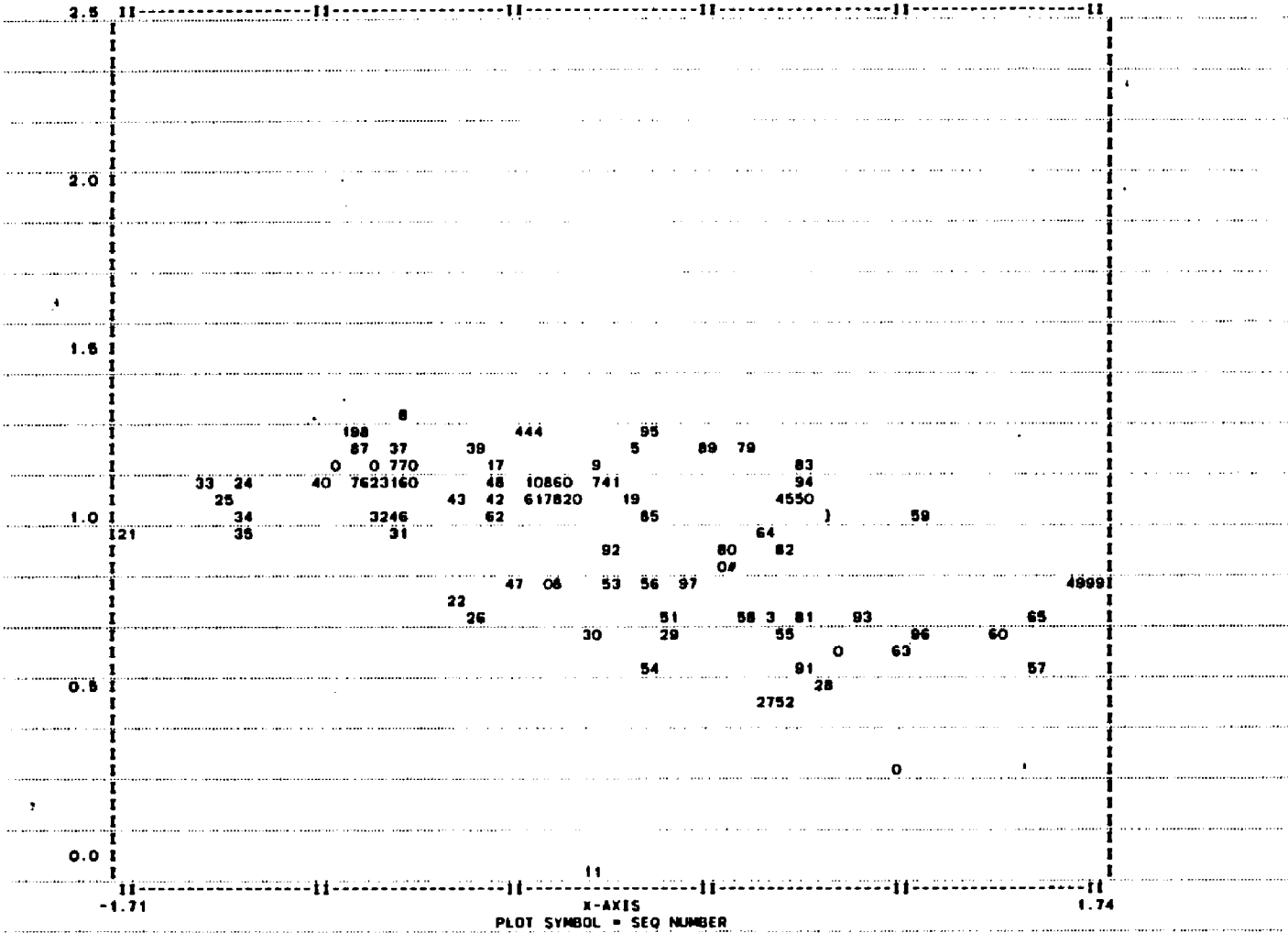
FITTING 1973-1979 MEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL  
 FIT MEAN SQUARE (Y) VERSUS DIFFICULTY (X)



FITTING 1973-1978 MEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL  
 FIT MEAN SQUARE (Y) VERSUS DISCRIMINATION (X)



FITTING 1973-1979 MEAP TEST RESULTS (FOR 4TH GRADERS) TO THE RASCH MODEL  
 DISCRIMINATION (Y) VS DIFFICULTY (X)



APPENDIX G

SPSS CONTROL SET EXAMPLE FOR TRANSFORMING SCORES ON MEAP OBJECTIVES IN

SAMPLE OF 5,000 DATA

WHEN ITEMS THAT DO NOT FIT THE RASCH MODEL ARE RESCORED

SPSS FOR OS/360, VERSION H, RELEASE 8.1, AUGUST 15, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM  
 ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT) ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS  
 SPSS PRIMER (BRIEF INTRO TO SPSS) SPSS POCKET GUIDE, RELEASE 8  
 SPSS UPDATE (USE W/SPSS, 2ND FOR REL. 7 & 8) KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION.. ALLOWS FOR.. 102 TRANSFORMATIONS  
 WORKSPACE 71680 BYTES 409 RECODE VALUES + LAG VARIABLES  
 TRANSSPACE 10240 BYTES 1641 IF/COMPUTE OPERATIONS

```

1 EDIT
2 RUN NAME RESCORING 1973 THROUGH 1979 4TH GRADE 'NO-FIT ITEMS'
3 COMMENT
4 FILE NAME RESCORE
5 COMMENT
6 COMMENT
7 COMMENT DATA FOR THESE RUNS (I.E., 14 RUNS): RSAMPLE1-RSAMPLE14
8 COMMENT
9 VARIABLE LIST RECTYPE, GRADE, SEX, AGEYRS, AGEMOS,
10 NUMPASS, NUMYR1ED, TOTQUEST,
11 OBSCOR01 TO OBSCOR23, OBSTAT01 TO OBSTAT23,
12 QA1 TO QA5, QB1 TO QB5, QC1 TO QC5, QD1 TO QD5,
13 QE1 TO QE5, QF1 TO QF5, QG1 TO QG5, QH1 TO QH5,
14 QI1 TO QI5, QJ1 TO QJ5, QK1 TO QK5, QL1 TO QL5,
15 QM1 TO QM5, QN1 TO QN5, OO1 TO OOS, OP1 TO OP5,
16 QQ1 TO QQ5, QR1 TO QR5, QS1 TO QS5, QT1 TO QT5,
17 QU1 TO QU5, QV1 TO QV5, QW1 TO QW5
18 COMMENT
19 INPUT MEDIUM DISK
20 N OF CASES UNKNOWN
21 COMMENT
22 INPUT FORMAT FIXED (1X,F2.0,1X,F2.0,1X,F1.0,1X,F2.0,1X,F2.0,
23 1X,F2.0,1X,F2.0,1X,F3.0,1X,23F1.0,1X,23F1.0/
24 1X,5F1.0,1X,5F1.0,1X,5F1.0,1X,5F1.0,
25 1X,5F1.0,1X,5F1.0,1X,5F1.0,1X,5F1.0,
26 1X,5F1.0,1X,5F1.0,1X,5F1.0,1X,5F1.0/
27 1X,5F1.0,1X,5F1.0,1X,5F1.0,1X,5F1.0,
28 1X,5F1.0,1X,5F1.0,1X,5F1.0,1X,5F1.0,
29 1X,5F1.0,1X,5F1.0,1X,5F1.0)
30 COMMENT
  
```

ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

VARIABLE	FORMAT	RECORD	COLUMNS
RECTYPE	F 2. 0	1	2- 3

## ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

VARIABLE	FORMAT	RECORD	COLUMNS
GRADE	F 2. 0	1	5- 6
SEX	F 1. 0	1	8- 8
AGEYRS	F 2. 0	1	10- 11
AGE MOS	F 2. 0	1	13- 14
NUMPASS	F 2. 0	1	16- 17
NUMTRIED	F 2. 0	1	19- 20
TOTQUEST	F 3. 0	1	22- 24
OBSCOR01	F 1. 0	1	26- 26
OBSCOR02	F 1. 0	1	27- 27
OBSCOR03	F 1. 0	1	28- 28
OBSCOR04	F 1. 0	1	29- 29
OBSCOR05	F 1. 0	1	30- 30
OBSCOR06	F 1. 0	1	31- 31
OBSCOR07	F 1. 0	1	32- 32
OBSCOR08	F 1. 0	1	33- 33
OBSCOR09	F 1. 0	1	34- 34
OBSCOR10	F 1. 0	1	35- 35
OBSCOR11	F 1. 0	1	36- 36
OBSCOR12	F 1. 0	1	37- 37
OBSCOR13	F 1. 0	1	38- 38
OBSCOR14	F 1. 0	1	39- 39
OBSCOR15	F 1. 0	1	40- 40
OBSCOR16	F 1. 0	1	41- 41
OBSCOR17	F 1. 0	1	42- 42
OBSCOR18	F 1. 0	1	43- 43
OBSCOR19	F 1. 0	1	44- 44
OBSCOR20	F 1. 0	1	45- 45
OBSCOR21	F 1. 0	1	46- 46
OBSCOR22	F 1. 0	1	47- 47
OBSCOR23	F 1. 0	1	48- 48
OBSTAT01	F 1. 0	1	50- 50
OBSTAT02	F 1. 0	1	51- 51
OBSTAT03	F 1. 0	1	52- 52
OBSTAT04	F 1. 0	1	53- 53
OBSTAT05	F 1. 0	1	54- 54
OBSTAT06	F 1. 0	1	55- 55
OBSTAT07	F 1. 0	1	56- 56
OBSTAT08	F 1. 0	1	57- 57
OBSTAT09	F 1. 0	1	58- 58
OBSTAT10	F 1. 0	1	59- 59
OBSTAT11	F 1. 0	1	60- 60

ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

VARIABLE	FORMAT	RECORD	COLUMNS
OBSTAT12	F 1. 0	1	61- 61
OBSTAT13	F 1. 0	1	62- 62
OBSTAT14	F 1. 0	1	63- 63
OBSTAT15	F 1. 0	1	64- 64
OBSTAT16	F 1. 0	1	65- 65
OBSTAT17	F 1. 0	1	66- 66
OBSTAT18	F 1. 0	1	67- 67
OBSTAT19	F 1. 0	1	68- 68
OBSTAT20	F 1. 0	1	69- 69
OBSTAT21	F 1. 0	1	70- 70
OBSTAT22	F 1. 0	1	71- 71
OBSTAT23	F 1. 0	1	72- 72
QA1	F 1. 0	2	2- 2
QA2	F 1. 0	2	3- 3
QA3	F 1. 0	2	4- 4
QA4	F 1. 0	2	5- 5
QA5	F 1. 0	2	6- 6
QB1	F 1. 0	2	8- 8
QB2	F 1. 0	2	9- 9
QB3	F 1. 0	2	10- 10
QB4	F 1. 0	2	11- 11
QB5	F 1. 0	2	12- 12
QC1	F 1. 0	2	14- 14
QC2	F 1. 0	2	15- 15
QC3	F 1. 0	2	16- 16
QC4	F 1. 0	2	17- 17
QC5	F 1. 0	2	18- 18
QD1	F 1. 0	2	20- 20
QD2	F 1. 0	2	21- 21
QD3	F 1. 0	2	22- 22
QD4	F 1. 0	2	23- 23
QD5	F 1. 0	2	24- 24
QE1	F 1. 0	2	26- 26
QE2	F 1. 0	2	27- 27
QE3	F 1. 0	2	28- 28
QE4	F 1. 0	2	29- 29
QE5	F 1. 0	2	30- 30
QF1	F 1. 0	2	32- 32
QF2	F 1. 0	2	33- 33
QF3	F 1. 0	2	34- 34
QF4	F 1. 0	2	35- 35



ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

VARIABLE	FORMAT	RECDRD	COLUMNS
QF3	F 1. 0	2	36- 36
QG1	F 1. 0	2	38- 38
QG2	F 1. 0	2	39- 39
QG3	F 1. 0	2	40- 40
QG4	F 1. 0	2	41- 41
QG5	F 1. 0	2	42- 42
QH1	F 1. 0	2	44- 44
QH2	F 1. 0	2	45- 45
QH3	F 1. 0	2	46- 46
QH4	F 1. 0	2	47- 47
QH5	F 1. 0	2	48- 48
QI1	F 1. 0	2	50- 50
QI2	F 1. 0	2	51- 51
QI3	F 1. 0	2	52- 52
QI4	F 1. 0	2	53- 53
QI5	F 1. 0	2	54- 54
QJ1	F 1. 0	2	56- 56
QJ2	F 1. 0	2	57- 57
QJ3	F 1. 0	2	58- 58
QJ4	F 1. 0	2	59- 59
QJ5	F 1. 0	2	60- 60
QK1	F 1. 0	2	62- 62
QK2	F 1. 0	2	63- 63
QK3	F 1. 0	2	64- 64
QK4	F 1. 0	2	65- 65
QK5	F 1. 0	2	66- 66
QL1	F 1. 0	2	68- 68
QL2	F 1. 0	2	69- 69
QL3	F 1. 0	2	70- 70
QL4	F 1. 0	2	71- 71
QL5	F 1. 0	2	72- 72
QM1	F 1. 0	3	2- 2
QM2	F 1. 0	3	3- 3
QM3	F 1. 0	3	4- 4
QM4	F 1. 0	3	5- 5
QM5	F 1. 0	3	6- 6
QN1	F 1. 0	3	8- 8
QN2	F 1. 0	3	9- 9
QN3	F 1. 0	3	10- 10
QN4	F 1. 0	3	11- 11
QN5	F 1. 0	3	12- 12

ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

VARIABLE	FORMAT	RECORD	COLUMNS
Q01	F 1. 0	3	14- 14
Q02	F 1. 0	3	15- 15
Q03	F 1. 0	3	16- 16
Q04	F 1. 0	3	17- 17
Q05	F 1. 0	3	18- 18
Q01	F 1. 0	3	20- 20
Q02	F 1. 0	3	21- 21
Q03	F 1. 0	3	22- 22
Q04	F 1. 0	3	23- 23
Q05	F 1. 0	3	24- 24
Q01	F 1. 0	3	26- 26
Q02	F 1. 0	3	27- 27
Q03	F 1. 0	3	28- 28
Q04	F 1. 0	3	29- 29
Q05	F 1. 0	3	30- 30
QR1	F 1. 0	3	32- 32
QR2	F 1. 0	3	33- 33
QR3	F 1. 0	3	34- 34
QR4	F 1. 0	3	35- 35
QR5	F 1. 0	3	36- 36
Q51	F 1. 0	3	38- 38
Q52	F 1. 0	3	39- 39
Q53	F 1. 0	3	40- 40
Q54	F 1. 0	3	41- 41
Q55	F 1. 0	3	42- 42
Q11	F 1. 0	3	44- 44
Q12	F 1. 0	3	45- 45
Q13	F 1. 0	3	46- 46
Q14	F 1. 0	3	47- 47
Q15	F 1. 0	3	48- 48
QU1	F 1. 0	3	50- 50
QU2	F 1. 0	3	51- 51
QU3	F 1. 0	3	52- 52
QU4	F 1. 0	3	53- 53
QU5	F 1. 0	3	54- 54
QV1	F 1. 0	3	56- 56
QV2	F 1. 0	3	57- 57
QV3	F 1. 0	3	58- 58
QV4	F 1. 0	3	59- 59
QV5	F 1. 0	3	60- 60
QW1	F 1. 0	3	62- 62

ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

VARIABLE	FORMAT	RECORD	COLUMNS
QW2	F 1. 0	3	63- 63
QW3	F 1. 0	3	64- 64
QW4	F 1. 0	3	65- 65
QW5	F 1. 0	3	66- 66

THE INPUT FORMAT PROVIDES FOR 169 VARIABLES. 169 WILL BE READ  
IT PROVIDES FOR 3 RECORDS ('CARDS') PER CASE. A MAXIMUM OF 72 'COLUMNS' ARE USED ON A RECORD.

31 ALLOCATE TRANSPACE=30000

SPECIFIED SPACE ALLOCATION. . . . . ALLOWS FOR . . . . . 300 TRANSFORMATIONS  
 WORKSPACE 51920 BYTES . . . . . 1200 RECODE VALUES + LAG VARIABLES  
 TRANSPACE 30000 BYTES . . . . . 4800 IF/COMPUTE OPERATIONS

- 32 COMPUTE COUNT=0
- 33 COMPUTE CA1=0
- 34 COMPUTE CA4=0
- 35 COMPUTE CB2=0
- 36 COMPUTE CB3=0
- 37 COMPUTE CB4=0
- 38 COMPUTE CE2=0
- 39 COMPUTE CF1=0
- 40 COMPUTE CF4=0
- 41 COMPUTE CFS=0
- 42 COMPUTE CG3=0
- 43 COMPUTE CH2=0
- 44 COMPUTE CH4=0
- 45 COMPUTE CI4=0
- 46 COMPUTE CK4=0
- 47 COMPUTE CP1=0
- 48 COMPUTE CQ1=0
- 49 COMPUTE CR2=0
- 50 COMPUTE CT1=0
- 51 COMPUTE CT3=0
- 52 COMPUTE CW3=0
- 53 COMPUTE CW4=0
- 54 COMPUTE SCOREQA=0
- 55 COMPUTE SCOREQB=0
- 56 COMPUTE SCOREQE=0
- 57 COMPUTE SCOREQF=0
- 58 COMPUTE SCOREQG=0
- 59 COMPUTE SCOREQH=0
- 60 COMPUTE SCOREQI=0
- 61 COMPUTE SCOREQK=0
- 62 COMPUTE SCOREQP=0
- 63 COMPUTE SCOREQQ=0
- 64 COMPUTE SCOREQR=0

65 COMPUTE	SCORED=0
66 COMPUTE	SCORED=0
67 COMPUTE	OBSTAT=0
68 COMPUTE	OBSTAT=0
69 COMPUTE	OBSTAT=0
70 COMPUTE	OBSTAT=0
71 COMPUTE	OBSTAT=0
72 COMPUTE	OBSTAT=0
73 COMPUTE	OBSTAT=0
74 COMPUTE	OBSTAT=0
75 COMPUTE	OBSTAT=0
76 COMPUTE	OBSTAT=0
77 COMPUTE	OBSTAT=0
78 COMPUTE	OBSTAT=0
79 COMPUTE	OBSTAT=0
80 COMPUTE	NEPASS=0
81 IF	(OAI EO 0) CA1+CA1+1
82 IF	(OAI EO 0) COUNT-COUNT+1
83 IF	(OAI EO 0) QA1+1
84 IF	(OAI EO 0) CA4+CA4+1
85 IF	(OAI EO 0) COUNT-COUNT+1
86 IF	(OAI EO 0) QA4+1
87 IF	(OAI EO 0) CB2+CB2+1
88 IF	(OAI EO 0) COUNT-COUNT+1
89 IF	(OAI EO 0) CB2+1
90 IF	(OAI EO 0) CB3+CB3+1
91 IF	(OAI EO 0) COUNT-COUNT+1
92 IF	(OAI EO 0) CB3+1
93 IF	(OAI EO 0) CB4+CB4+1
94 IF	(OAI EO 0) COUNT-COUNT+1
95 IF	(OAI EO 0) CB4+1
96 IF	(OAI EO 0) CE2+CE2+1
97 IF	(OAI EO 0) COUNT-COUNT+1
98 IF	(OAI EO 0) CE2+1
99 IF	(OAI EO 0) CF1+CF1+1
100 IF	(OAI EO 0) COUNT-COUNT+1
101 IF	(OAI EO 0) CF1+1
102 IF	(OAI EO 0) CF4+CF4+1
103 IF	(OAI EO 0) COUNT-COUNT+1
104 IF	(OAI EO 0) CF4+1
105 IF	(OAI EO 0) CF5+CF5+1
106 IF	(OAI EO 0) COUNT-COUNT+1
107 IF	(OAI EO 0) CF5+1
108 IF	(OAI EO 0) CG3+CG3+1
109 IF	(OAI EO 0) COUNT-COUNT+1
110 IF	(OAI EO 0) CG3+1
111 IF	(OAI EO 0) CH2+CH2+1
112 IF	(OAI EO 0) COUNT-COUNT+1
113 IF	(OAI EO 0) CH2+1
114 IF	(OAI EO 0) CH4+CH4+1
115 IF	(OAI EO 0) COUNT-COUNT+1
116 IF	(OAI EO 0) CH4+1
117 IF	(OAI EO 0) C14+C14+1

```

118 IF      (Q14 EQ 0) COUNT=COUNT+1
119 IF      (Q14 EQ 0) Q14=1
120 IF      (QK4 EQ 0) CK4=CK4+1
121 IF      (QK4 EQ 0) COUNT=COUNT+1
122 IF      (QK4 EQ 0) QK4=1
123 IF      (OP1 EQ 0) CP1=CP1+1
124 IF      (OP1 EQ 0) COUNT=COUNT+1
125 IF      (OP1 EQ 0) OP1=1
126 IF      (QO1 EQ 0) CO1=CO1+1
127 IF      (QO1 EQ 0) COUNT=COUNT+1
128 IF      (QO1 EQ 0) QO1=1
129 IF      (QR2 EQ 0) CR2=CR2+1
130 IF      (QR2 EQ 0) COUNT=COUNT+1
131 IF      (QR2 EQ 0) QR2=1
132 IF      (QT1 EQ 0) CT1=CT1+1
133 IF      (QT1 EQ 0) COUNT=COUNT+1
134 IF      (QT1 EQ 0) QT1=1
135 IF      (QT3 EQ 0) CT3=CT3+1
136 IF      (QT3 EQ 0) COUNT=COUNT+1
137 IF      (QT3 EQ 0) QT3=1
138 IF      (QW3 EQ 0) CW3=CW3+1
139 IF      (QW3 EQ 0) COUNT=COUNT+1
140 IF      (QW3 EQ 0) QW3=1
141 IF      (QW4 EQ 0) CW4=CW4+1
142 IF      (QW4 EQ 0) COUNT=COUNT+1
143 IF      (QW4 EQ 0) QW4=1
144 COMPUTE SCOREQA=SCOREQA+QA1
145 COMPUTE SCOREQA=SCOREQA+QA2
146 COMPUTE SCOREQA=SCOREQA+QA3
147 COMPUTE SCOREQA=SCOREQA+QA4
148 COMPUTE SCOREQA=SCOREQA+QA5
149 IF      (SCOREQA GE 4) OBSTATA=1
150 COMPUTE SCOREQB=SCOREQB+QB1
151 COMPUTE SCOREQB=SCOREQB+QB2
152 COMPUTE SCOREQB=SCOREQB+QB3
153 COMPUTE SCOREQB=SCOREQB+QB4
154 COMPUTE SCOREQB=SCOREQB+QB5
155 IF      (SCOREQB GE 4) OBSTATB=1
156 COMPUTE SCOREQE=SCOREQE+QE1
157 COMPUTE SCOREQE=SCOREQE+QE2
158 COMPUTE SCOREQE=SCOREQE+QE3
159 COMPUTE SCOREQE=SCOREQE+QE4
160 COMPUTE SCOREQE=SCOREQE+QE5
161 IF      (SCOREQE GE 4) OBSTATE=1
162 COMPUTE SCOREQF=SCOREQF+QF1
163 COMPUTE SCOREQF=SCOREQF+QF2
164 COMPUTE SCOREQF=SCOREQF+QF3
165 COMPUTE SCOREQF=SCOREQF+QF4
166 COMPUTE SCOREQF=SCOREQF+QF5
167 IF      (SCOREQF GE 4) OBSTATF=1
168 COMPUTE SCOREQG=SCOREQG+QG1
169 COMPUTE SCOREQG=SCOREQG+QG2
170 COMPUTE SCOREQG=SCOREQG+QG3
    
```

171 COMPUTE	SCOREQG=SCOREQG+QG4
172 COMPUTE	SCOREQG=SCOREQG+QG5
173 IF	(SCOREQG GE 4) OBSTATG=1
174 COMPUTE	SCOREQH=SCOREQH+QH1
175 COMPUTE	SCOREQH=SCOREQH+QH2
176 COMPUTE	SCOREQH=SCOREQH+QH3
177 COMPUTE	SCOREQH=SCOREQH+QH4
178 COMPUTE	SCOREQH=SCOREQH+QH5
179 IF	(SCOREQH GE 4) OBSTATH=1
180 COMPUTE	SCOREQI=SCOREQI+QI1
181 COMPUTE	SCOREQI=SCOREQI+QI2
182 COMPUTE	SCOREQI=SCOREQI+QI3
183 COMPUTE	SCOREQI=SCOREQI+QI4
184 COMPUTE	SCOREQI=SCOREQI+QI5
185 IF	(SCOREQI GE 4) OBSTATI=1
186 COMPUTE	SCOREQK=SCOREQK+QK1
187 COMPUTE	SCOREQK=SCOREQK+QK2
188 COMPUTE	SCOREQK=SCOREQK+QK3
189 COMPUTE	SCOREQK=SCOREQK+QK4
190 COMPUTE	SCOREQK=SCOREQK+QK5
191 IF	(SCOREQK GE 4) OBSTATK=1
192 COMPUTE	SCOREQP=SCOREQP+QP1
193 COMPUTE	SCOREQP=SCOREQP+QP2
194 COMPUTE	SCOREQP=SCOREQP+QP3
195 COMPUTE	SCOREQP=SCOREQP+QP4
196 COMPUTE	SCOREQP=SCOREQP+QP5
197 IF	(SCOREQP GE 4) OBSTATP=1
198 COMPUTE	SCOREQQ=SCOREQQ+QQ1
199 COMPUTE	SCOREQQ=SCOREQQ+QQ2
200 COMPUTE	SCOREQQ=SCOREQQ+QQ3
201 COMPUTE	SCOREQQ=SCOREQQ+QQ4
202 COMPUTE	SCOREQQ=SCOREQQ+QQ5
203 IF	(SCOREQQ GE 4) OBSTATQ=1
204 COMPUTE	SCOREQR=SCOREQR+QR1
205 COMPUTE	SCOREQR=SCOREQR+QR2
206 COMPUTE	SCOREQR=SCOREQR+QR3
207 COMPUTE	SCOREQR=SCOREQR+QR4
208 COMPUTE	SCOREQR=SCOREQR+QR5
209 IF	(SCOREQR GE 4) OBSTATR=1
210 COMPUTE	SCOREQT=SCOREQT+QT1
211 COMPUTE	SCOREQT=SCOREQT+QT2
212 COMPUTE	SCOREQT=SCOREQT+QT3
213 COMPUTE	SCOREQT=SCOREQT+QT4
214 COMPUTE	SCOREQT=SCOREQT+QT5
215 IF	(SCOREQT GE 4) OBSTATT=1
216 COMPUTE	SCOREQW=SCOREQW+QW1
217 COMPUTE	SCOREQW=SCOREQW+QW2
218 COMPUTE	SCOREQW=SCOREQW+QW3
219 COMPUTE	SCOREQW=SCOREQW+QW4
220 COMPUTE	SCOREQW=SCOREQW+QW5
221 IF	(SCOREQW GE 4) OBSTATW=1
222 COMPUTE	NEWPASS=NEWPASS+OBSTATA
223 COMPUTE	NEWPASS=NEWPASS+OBSTATB

```

224 COMPUTE NEWPASS=NEWPASS*OBSTAT03
225 COMPUTE NEWPASS=NEWPASS*OBSTAT04
226 COMPUTE NEWPASS=NEWPASS*OBSTATE
227 COMPUTE NEWPASS=NEWPASS*OBSTATF
228 COMPUTE NEWPASS=NEWPASS*OBSTATG
229 COMPUTE NEWPASS=NEWPASS*OBSTATH
230 COMPUTE NEWPASS=NEWPASS*OBSTATI
231 COMPUTE NEWPASS=NEWPASS*OBSTAT10
232 COMPUTE NEWPASS=NEWPASS*OBSTATK
233 COMPUTE NEWPASS=NEWPASS*OBSTAT12
234 COMPUTE NEWPASS=NEWPASS*OBSTAT13
235 COMPUTE NEWPASS=NEWPASS*OBSTAT14
236 COMPUTE NEWPASS=NEWPASS*OBSTAT15
237 COMPUTE NEWPASS=NEWPASS*OBSTATP
238 COMPUTE NEWPASS=NEWPASS*OBSTATQ
239 COMPUTE NEWPASS=NEWPASS*OBSTATR
240 COMPUTE NEWPASS=NEWPASS*OBSTAT19
241 COMPUTE NEWPASS=NEWPASS*OBSTAT17
242 COMPUTE NEWPASS=NEWPASS*OBSTAT21
243 COMPUTE NEWPASS=NEWPASS*OBSTAT22
244 COMPUTE NEWPASS=NEWPASS*OBSTATW
245 COMMENT
246 WRITE CASES {IX,F2.0,IX,F2.0,IX,F1.0,IX,F2.0,IX,F2.0,
247 IX,F2.0,IX,F2.0,IX,F3.0,IX,23F1.0,IX,23F1.0/
248 IX,5F1.0,IX,5F1.0,IX,5F1.0,IX,5F1.0,
249 IX,5F1.0,IX,5F1.0,IX,5F1.0,IX,5F1.0,
250 IX,5F1.0,IX,5F1.0,IX,5F1.0,IX,5F1.0/
251 IX,5F1.0,IX,5F1.0,IX,5F1.0,IX,5F1.0,
252 IX,5F1.0,IX,5F1.0,IX,5F1.0,IX,5F1.0,
253 IX,5F1.0,IX,5F1.0,IX,5F1.0,IX,F4.0/
254 21F3.0,IX,F2.0}
255 RECTYPE, GRADE, SEX, AGEYRS, AGEMOS,
256 NMPASS, NUMTRIED, TOTQUEST,
257 OBSCOR01 TO OBSCOR23, OBSTAT01 TO OBSTAT23,
258 QA1 TO QAS, OB1 TO OB5, OC1 TO QCS, QO1 TO QO5,
259 QE1 TO QES, QF1 TO QF5, QG1 TO QGS, QH1 TO QH5,
260 QI1 TO QIS, QJ1 TO QJ5, QK1 TO QKS, QL1 TO QLS,
261 QM1 TO QMS, QN1 TO QNS, QO1 TO QOS, QP1 TO QP5,
262 QQ1 TO QQS, QR1 TO QRS, QS1 TO QSS, QT1 TO QTS,
263 QU1 TO QUS, QV1 TO QVS, QW1 TO QWS
264 COUNTY, CA1, CA4, CB2, CB3, CB4, CE2, CF1, CF4, CF5, CG3,
265 CH2, CH4, CI4, CK4, CP1, CQ1, CR2, CT1, CT3, CW3, CW4, NEWPASS
266 COMMENT

```

TRANSSPACE REQUIRED... 21300 BYTES

213 TRANSFORMATIONS  
0 RECODE VALUES + LAG VARIABLES  
784 IF/COMPUTE OPERATIONS

```

267 T-TEST PAIRS=NEWPASS WITH NMPASS
268 OPTIONS 2
269 COMMENT

```

\*\*\*\*\* T-TEST PROBLEM REQUIRES 56 BYTES OF WORKSPACE \*\*\*\*\*

RESCORING 1973 THROUGH 1979 4TH GRADE 'NO-FIT ITEMS'

08-31-81

PAGE 11

270 READ INPUT DATA  
271 FINISH

NORMAL END OF JOB.  
271 CONTROL CARDS WERE PROCESSED.  
0 ERRORS WERE DETECTED.



APPENDIX H  
LOG OF COMPUTER RUNS USING BICAL TO PROCESS  
FOURTEEN RANDOM SAMPLES TAKEN FROM  
SAMPLE OF 5,000 DATA

LOG ON AT 20:05:10 ON 08-31-81

#FILES

FOURTH73 MODEL7TH RUNSAM SAMPLE SEVEN73 SEVEN74 SEVEN75 SEVEN76 SEVEN77 SEVEN78 SEVEN79

#LIST RUN SAM

>"RUN" DOES NOT EXIST.

>Enter Replacement or "CANCEL".

?CANCEL

#LIST RUNSAM

```
> 1 GET -OUT
> 2 GET -NEWSAM
> 3 EMPTY -OUT
> 4 EMPTY -NEWSAM
> 5 RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
> 6 LIST -OUT(LAST-5)
> 7 RUN *FS 0=*T*
```

# END OF FILE

#EDIT RUNSAM

:MOVE 3 3 1

: 1,5 EMPTY -OUT

:RENUMBER

:P 1 \*L

```
: 1 GET -OUT
: 2 EMPTY -OUT
: 3 GET -NEWSAM
: 4 EMPTY -NEWSAM
: 5 RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
: 6 LIST -OUT(LAST-5)
: 7 RUN *FS 0=*T*
```

:A 5 :CONTROL:-CONTROL:

: 5 RUN \*SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM

:i 5

?EMPTY -CONTROL

?EMPTY -SAMPLE

?

:RENUMBER

:P 1 \*L

```
: 1 GET -OUT
: 2 EMPTY -OUT
: 3 GET -NEWSAM
: 4 EMPTY -NEWSAM
: 5 RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
: 6 EMPTY -CONTROL
: 7 EMPTY -SAMPLE
: 8 LIST -OUT(LAST-5)
: 9 RUN *FS 0=*T*
```

:STOP

#MOUNT 0620 9TP \*T\* VOL=000620 RING-IN 'MEAP7374'

# 0620 9TP \*T\* VOL=000620 RING-IN 'MEAP7374'

# \*T\* (0620): Mounted on T4C7

#RUN \*FS 0=\*T\*

#EXECUTION BEGINS 20:10:38

-RESTORE RSAMPLE1 -SAM1

= FILE 23 "RSAMPLE1(1)" HAS BEEN QUEUED FOR RESTORATION

-RESTORE RSAMPLE3 -SAM3

= FILE 24 "RSAMPLE3(1)" HAS BEEN QUEUED FOR RESTORATION

-RESTORE RSAMPLE5 -SAM5

= FILE 25 "RSAMPLE5(1)" HAS BEEN QUEUED FOR RESTORATION

-RESTORE RSAMPLE7 -SAM7

= FILE 26 "RSAMPLE7(1)" HAS BEEN QUEUED FOR RESTORATION

-RESTORE RSAMPLE9 -SAM9

```

* FILE 27 "RSAMPLE9(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE11 -SAM11
* FILE 28 "RSAMPLE11(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE13 -SAM13
* FILE 29 "RSAMPLE13(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE2 -SAM2
* FILE 30 "RSAMPLE2(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE4 -SAM4
* FILE 31 "RSAMPLE4(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE6 -SAM6
* FILE 32 "RSAMPLE6(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE8 -SAM8
* FILE 33 "RSAMPLE8(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE10 -SAM10
* FILE 34 "RSAMPLE10(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE12 -SAM12
* FILE 35 "RSAMPLE12(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE RSAMPLE14 -SAM14
* FILE 36 "RSAMPLE14(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE (68) -FOURTH73
* FILE 68 "RECODE.73.4TH.GRADE.NO.FIT.ITEMS(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE (69) -FOURTH74
* FILE 69 "RECODE.74.4TH.GRADE.NO.FIT.ITEMS(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE (70) -FOURTH75
* FILE 70 "RECODE.75.4TH.GRADE.NO.FIT.ITEMS(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE (71) -FOURTH76
* FILE 71 "RECODE.76.4TH.GRADE.NO.FIT.ITEMS(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE (72) -FOURTH77
* FILE 72 "RECODE.77.4TH.GRADE.NO.FIT.ITEMS(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE (73) -FOURTH78
* FILE 73 "RECODE.78.4TH.GRADE.NO.FIT.ITEMS(1)" HAS BEEN QUEUED FOR RESTORATION
*RESTORE (74) -FOURTH79
* FILE 74 "RECODE.79.4TH.GRADE.NO.FIT.ITEMS(1)" HAS BEEN QUEUED FOR RESTORATION
*STOP
* FILE 23 "RSAMPLE1(1)" ... HAS BEEN RESTORED TO -SAM1
* FILE 24 "RSAMPLE3(1)" ... HAS BEEN RESTORED TO -SAM3
* FILE 25 "RSAMPLE5(1)" ... HAS BEEN RESTORED TO -SAM5
* FILE 26 "RSAMPLE7(1)" ... HAS BEEN RESTORED TO -SAM7
* FILE 27 "RSAMPLE9(1)" ... HAS BEEN RESTORED TO -SAM9
* FILE 28 "RSAMPLE11(1)" ... HAS BEEN RESTORED TO -SAM11
* FILE 29 "RSAMPLE13(1)" ... HAS BEEN RESTORED TO -SAM13
* FILE 30 "RSAMPLE2(1)" ... HAS BEEN RESTORED TO -SAM2
* FILE 31 "RSAMPLE4(1)" ... HAS BEEN RESTORED TO -SAM4
* FILE 32 "RSAMPLE6(1)" ... HAS BEEN RESTORED TO -SAM6
* FILE 33 "RSAMPLE8(1)" ... HAS BEEN RESTORED TO -SAM8
* FILE 34 "RSAMPLE10(1)" ... HAS BEEN RESTORED TO -SAM10
* FILE 35 "RSAMPLE12(1)" ... HAS BEEN RESTORED TO -SAM12
* FILE 36 "RSAMPLE14(1)" ... HAS BEEN RESTORED TO -SAM14
* FILE 68 "RECODE.73.4TH.GRADE.NO.FIT.ITEMS(1)" ... HAS BEEN RESTORED TO -FOURTH73
* FILE 69 "RECODE.74.4TH.GRADE.NO.FIT.ITEMS(1)" ... HAS BEEN RESTORED TO -FOURTH74
* FILE 70 "RECODE.75.4TH.GRADE.NO.FIT.ITEMS(1)" ... HAS BEEN RESTORED TO -FOURTH75
* FILE 71 "RECODE.76.4TH.GRADE.NO.FIT.ITEMS(1)" ... HAS BEEN RESTORED TO -FOURTH76
* FILE 72 "RECODE.77.4TH.GRADE.NO.FIT.ITEMS(1)" ... HAS BEEN RESTORED TO -FOURTH77
* FILE 73 "RECODE.78.4TH.GRADE.NO.FIT.ITEMS(1)" ... HAS BEEN RESTORED TO -FOURTH78
* FILE 74 "RECODE.79.4TH.GRADE.NO.FIT.ITEMS(1)" ... HAS BEEN RESTORED TO -FOURTH79
#EXECUTION TERMINATED T=17.605 $2.35
#GET -CONTROL
#READY.
#GET -SAMPLE
#READY.

```

```

#COPY -FOURTH73 TO -CONTROL
#COPY -SAM1 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 20:26:44
#EXECUTION TERMINATED T=18.378 $3.04
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 532
> 533
> 534 252 FINISH
> 535 O NORMAL END OF JOB.
> 536 252 CONTROL CARDS WERE PROCESSED.
> 537 O ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS 0--T*
#EXECUTION BEGINS 20:28:20
- DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
7Y
*SAVE -OUT T.STATS.FOR.73.4TH.GRADE
* FILE "T.STATS.FOR.73.4TH.GRADE(2)" ... HAS BEEN SAVED
*SAVE -NEWSAM RECODED.73.4TH.GRADE.SAMPLE
* FILE "RECODED.73.4TH.GRADE.SAMPLE(2)" ... HAS BEEN SAVED
*STOP
#EXECUTION TERMINATED T=2.569 $ .37
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)+-OUT
#EXECUTION BEGINS 20:32:09
*PRINT* ASSIGNED RECEIPT NUMBER 617291
*PRINT* 617291 HELD
*PRINT* 617291 RELEASED TO CSCO(SP), 15 PAGES.
#EXECUTION TERMINATED T=0.368 $ .60
#COPY -FOURTH74 TO -CONTROL
#COPY -SAM3 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 20:33:27
#EXECUTION TERMINATED T=15.85 $2.61
# EMPTY -CONTROL
#DONE.

```

```

# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 473
> 474
> 475          195 FINISH
> 476      O   NORMAL END OF JOB.
> 477          195 CONTROL CARDS WERE PROCESSED.
> 478          0 ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O**T*
#EXECUTION BEGINS  20:34:50
# DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
#SAVE -OUT T.STATS.FOR.74.4TH.GRADE
# FILE "T.STATS.FOR.74.4TH.GRADE(1)" ... HAS BEEN SAVED
#SAVE -NEWSAM RECODED.74.4TH.GRADE.SAMPLE
# FILE "RECODED.74.4TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
#STOP
#EXECUTION TERMINATED  T=2.535      $ .36
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS**PGF(20)**-OUT
#EXECUTION BEGINS  20:38:16
#PRINT* ASSIGNED RECEIPT NUMBER 617295
#PRINT* 617295 HELD
#PRINT* 617295 RELEASED TO CSCO(SP), 14 PAGES.
#EXECUTION TERMINATED  T=0.326      $ .56
#COPY -FOURTH75 TO -CONTROL
#COPY -SAMS TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS  20:40:03
#EXECUTION TERMINATED  T=15.208     $2.51
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 455
> 456
> 457          177 FINISH
> 458      O   NORMAL END OF JOB.
> 459          177 CONTROL CARDS WERE PROCESSED.
> 460          0 ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O**T*
#EXECUTION BEGINS  20:41:35
# DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
#SAVE -OUT T.STATS.FOR.75.4TH.GRADE
# FILE "T.STATS.FOR.75.4TH.GRADE(1)" ... HAS BEEN SAVED

```

```

=SAVE -NEWSAM RECODED.75.4TH.GRADE.SAMPLE
= FILE "RECODED.75.4TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
=STOP
#EXECUTION TERMINATED T=2.501      $.38
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)+-OUT
#EXECUTION BEGINS 20:46:07
*PRINT* ASSIGNED RECEIPT NUMBER 617297
*PRINT* 617297 HELD
*PRINT* 617297 RELEASED TO CSCO(SP), 14 PAGES.
#EXECUTION TERMINATED T=0.312      $.56
#COPY -FOURTH76 TO -CONTROL
#COPY -SAM7 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 20:47:12
#EXECUTION TERMINATED T=15.917     $2.63
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 473
> 474
> 475          195 FINISH
> 476      0   NORMAL END OF JOB.
> 477          195 CONTROL CARDS WERE PROCESSED.
> 478          0 ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O=*T*
#EXECUTION BEGINS 20:48:26
= DD YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
=SAVE -OUT T.STATS.FOR.76.4TH.GRADE
= FILE "T.STATS.FOR.76.4TH.GRADE(1)" ... HAS BEEN SAVED
=SAVE -NEWSAM RECODED.76.4TH.GRADE.SAMPLE
= FILE "RECODED.76.4TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
=STOP
#EXECUTION TERMINATED T=2.452      $.35
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)+-OUT
#EXECUTION BEGINS 20:51:39
*PRINT* ASSIGNED RECEIPT NUMBER 617299
*PRINT* 617299 HELD
*PRINT* 617299 RELEASED TO CSCO(SP), 14 PAGES.
#EXECUTION TERMINATED T=0.298      $.56
#COPY -FOURTH77 TO -CONTROL
#COPY -SAM9 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT

```

```

#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 20:52:39
#EXECUTION TERMINATED T=14.602 $2.41
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 435
> 436
> 437 159 FINISH
> 438 0 NORMAL END OF JOB.
> 439 159 CONTROL CARDS WERE PROCESSED.
> 440 0 ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS Q=*T*
#EXECUTION BEGINS 20:53:43
= DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
#SAVE -OUT T.STATS.FOR.77.4TH.GRADE
# FILE "T.STATS.FOR.77.4TH.GRADE(1)" ... HAS BEEN SAVED
#SAVE -NEWSAM RECODED.77.4TH.GRADE.SAMPLE
# FILE "RECODED.77.4TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
#STOP
#EXECUTION TERMINATED T=2.577 $.37
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)*-OUT
#EXECUTION BEGINS 20:56:58
*PRINT* ASSIGNED RECEIPT NUMBER 617305
Attn ... *PAGEPR Cancelled.
*PRINT* 617305 HELD
*PRINT* 617305 CANCELLED.
#EXECUTION TERMINATED T=0.076 $.05
#RUN *PAGEPR SCARDS=*PGF(20)*-OUT
#EXECUTION BEGINS 20:59:37
*PRINT* ASSIGNED RECEIPT NUMBER 617306
*PRINT* 617306 HELD
Attn ... *PAGEPR Cancelled.
*PRINT* 617306 CANCELLED.
#EXECUTION TERMINATED T=0.281 $.08
#COPY -OUT -HLD784TH
#LIST -HLD784TH
> 1 ISPSS BATCH SYSTEM
> 2
>ATTN!
#LIST -HLD784TH(LAST)
> 440 0 ERRORS WERE DETECTED.
# END OF FILE
#
#ATTN!
#
#ATTN!

```

08-31-81

PAGE 1

```

#RENAME -HLD784TH -HLD774TH
#DONE.
#COPY -FOURTH8 TO -CONTROL
#COPY -SAM11 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 21:05:56
#EXECUTION TERMINATED T=14.309 $2.36
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 432
> 433
> 434 156 FINISH
> 435 O NORMAL END OF JOB.
> 436 156 CONTROL CARDS WERE PROCESSED.
> 437 O ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS 0+*1*
#EXECUTION BEGINS 21:07:05
= DO YOU WISH TO CONTINUE?...ENTER *Y* OR *N*
?Y
=SAVE -OUT T. STATS.FOR.78.4TH.GRADE
= FILE "T.STATS.FOR.78.4TH.GRADE(1)" ... HAS BEEN SAVED
=SAVE -NEWSAMRECODED.78.4TH.GRADE.SAMPLE
= *** ERROR ** FILE "-NEWSAMRECODED.78.4TH.GRADE.SAMP" ...IS EMPTY???
= CANCEL THE SAVE?... ENTER Y OR N
?
= CANCEL THE SAVE?... ENTER Y OR N
?Y
= SAVE CANCELLED...ENTER NEXT COMMAND
= *** ATTN
=SAVE -NEWSAM RECODED.78.4TH.GRADE.SAMPLE
= FILE "RECODED.78.4TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
=STOP
#EXECUTION TERMINATED T=2.654 RC=4 $1.40
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)+-HLD774TH
#EXECUTION BEGINS 21:11:37
*PRINT* ASSIGNED RECEIPT NUMBER 617315
*PRINT* 617315 HELD
*PRINT* 617315 RELEASED TO CSC0(SP), 13 PAGES.
#EXECUTION TERMINATED T=0.27 $1.53
#RUN *PAGEPR SCARDS=*PGF(20)+-OUT
#EXECUTION BEGINS 21:13:30
*PRINT* ASSIGNED RECEIPT NUMBER 617316
*PRINT* 617316 HELD
*PRINT* 617316 RELEASED TO CSC0(SP), 13 PAGES.
#EXECUTION TERMINATED T=0.304 $1.53

```



```

#COPY -FOURTH9 TO -CONTROL
#COPY -SAM13 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 21:14:52
#EXECUTION TERMINATED T=14.254 $2.36
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 426
> 427
> 428 150 FINISH
> 429 O NORMAL END OF JOB.
> 430 150 CONTROL CARDS WERE PROCESSED.
> 431 O ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS 0**T*
#EXECUTION BEGINS 21:16:05
# DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
#SAVE -OUT T.STATS.FOR.79.4TH.GRADE
# FILE "T.STATS.FOR.79.4TH.GRADE(1)" ... HAS BEEN SAVED
#SAVE -NEWSAM RECODED.79.4TH.GRADE.SAMPLE
# FILE "RECODED.79.4TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
#STOP
#EXECUTION TERMINATED T=2.456 $ .36
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)+-OUT
#EXECUTION BEGINS 21:19:24
*PRINT* ASSIGNED RECEIPT NUMBER 617318
*PRINT* 617318 HELD
*PRINT* 617318 RELEASED TO CSCO(SP), 13 PAGES.
#EXECUTION TERMINATED T=0.279 $ .53
#RUN *PAGEPR SCARDS=*PGF(20)+-LOG
#EXECUTION BEGINS 21:20:03
*PRINT* ASSIGNED RECEIPT NUMBER 617319
*PRINT* 617319 HELD

```

```

DONE.
#COPY SEVENT3 TO -CONTROL
#COPY -SAM2 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 21:21:43
#EXECUTION TERMINATED T=15.411 $2.55
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 462
> 463
> 464 184 FINISH
> 465 O NORMAL END OF JOB.
> 466 184 CONTROL CARDS WERE PROCESSED.
> 467 O ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O**T*
#EXECUTION BEGINS 21:22:56
# DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
#SAVE -OUT T.STATS.FOR.73.7TH.GRADE
# FILE "T.STATS.FOR.73.7TH.GRADE(2)" ... HAS BEEN SAVED
#SAVE -NEWSAM RECODED.73.7TH.GRADE.SAMPLE
# FILE "RECODED.73.7TH.GRADE.SAMPLE(2)" ... HAS BEEN SAVED
#STOP
#EXECUTION TERMINATED T=2.54 $0.37
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS*PGF(20)*-OUT
#EXECUTION BEGINS 21:26:20
*PRINT* ASSIGNED RECEIPT NUMBER 617321
*PRINT* 617321 HELD
*PRINT* 617321 RELEASED TO CSC0(5P), 14 PAGES.
#EXECUTION TERMINATED T=0.32 $0.56
#COPY SEVENT4 TO -CONTROL
#COPY -SAM4 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 21:27:20
#EXECUTION TERMINATED T=15.67 $2.60
# EMPTY -CONTROL

```

```

#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 480
> 481
> 482          202 FINISH
> 483      O   NORMAL END OF JOB.
> 484          202 CONTROL CARDS WERE PROCESSED.
> 485          O   ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O*T*
#EXECUTION BEGINS 21:28:31
* DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
*SAVE -OUT T.STATS.FOR.74.7TH.GRADE
* FILE "T.STATS.FOR.74.7TH.GRADE(1)" ... HAS BEEN SAVED
*SAVE -NEWSAM RECODED.74.7TH.GRADE.SAMPLE
* FILE "RECODED.74.7TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
*STOP
#EXECUTION TERMINATED T=2.508      $.36
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)*-OUT
#EXECUTION BEGINS 21:31:50
*PRINT* ASSIGNED RECEIPT NUMBER 617326
*PRINT* 617326 HELD
*PRINT* 617326 RELEASED TO CSCO(SP), 14 PAGES.
#EXECUTION TERMINATED T=0.301      $.56
#COPY SEVENT5 TO -CONTROL
#COPY -SAM6 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 21:33:07
#EXECUTION TERMINATED T=15.713      $2.60
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 480
> 481
> 482          202 FINISH
> 483      O   NORMAL END OF JOB.
> 484          202 CONTROL CARDS WERE PROCESSED.
> 485          O   ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O*T*
#EXECUTION BEGINS 21:34:20
* DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
*SAVE -OUT T.STATS.FOR.75.7TH.GRADE

```

```

* FILE "T.STATS.FOR.75.7TH.GRADE(1)" ... HAS BEEN SAVED
*SAVE -NEWSAM RECODED.75.7TH.GRADE.SAMPLE
* FILE "RECODED.75.7TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
*STOP
#EXECUTION TERMINATED T=2.55      $.37
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS*PGF(20)+-OUT
#EXECUTION BEGINS 21:37:32
*PRINT* ASSIGNED RECEIPT NUMBER 617327
*PRINT* 617327 HELD
*PRINT* 617327 RELEASED TO CSCO(SP), 14 PAGES.
#EXECUTION TERMINATED T=0.339      $.57
#COPY SEVEN76 TO -CONTROL
#COPY -SAM8 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 21:38:32
#EXECUTION TERMINATED T=14.62      $2.42
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 433
> 434
> 435          157 FINISH
> 436          0   NORMAL END OF JOB.
> 437          157 CONTROL CARDS WERE PROCESSED.
> 438          0   ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS 0--T*
#EXECUTION BEGINS 21:39:38
= DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
?Y
*SAVE -OUT T.STATS.FOR.76.7TH.GRADE
* FILE "T.STATS.FOR.76.7TH.GRADE(1)" ... HAS BEEN SAVED
*SAVE -NEWSAM RECODED.76.7TH.GRADE.SAMPLE
* FILE "RECODED.76.7TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
*STOP
#EXECUTION TERMINATED T=2.54      $.37
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS*PGF(20)+-OUT
#EXECUTION BEGINS 21:43:01
*PRINT* ASSIGNED RECEIPT NUMBER 617330
*PRINT* 617330 HELD
*PRINT* 617330 RELEASED TO CSCO(SP), 13 PAGES.
#EXECUTION TERMINATED T=0.28      $.53
#COPY SEVEN77 TO -CONTROL
#COPY -SAM10 TO -SAMPLE
#SOURCE RUNSAM

```

```

# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5*-CONTROL 6*-OUT 8*-SAMPLE 9*-NEWSAM
#EXECUTION BEGINS 21:44:00
#EXECUTION TERMINATED T=14.659 $2.43
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 433
> 434
> 435 157 FINISH
> 436 O NORMAL END OF JOB.
> 437 157 CONTROL CARDS WERE PROCESSED.
> 438 O ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS 0*+T*
#EXECUTION BEGINS 21:45:05
# DO YOU WISH TO CONTINUE?...ENTER "Y" OR "N"
7Y
#SAVE -OUT T.STATS.FOR.77.7TH.GRADE
# FILE "T.STATS.FOR.77.7TH.GRADE(1)" ... HAS BEEN SAVED
#SAVE -NEWSAM RECCDED.77.7TH.GRADE.SAMPLE
# FILE "RECCDED.77.7TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
#STOP
#EXECUTION TERMINATED T=2.533 $ .37
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS**PGF(20)*-OUT
#EXECUTION BEGINS 21:48:19
*PRINT* ASSIGNED RECEIPT NUMBER 617333
*PRINT* 617333 HELD
*PRINT* 617333 RELEASED TO CSC0(SP). 13 PAGES.
#EXECUTION TERMINATED T=0.282 $ .53
#COPY SEVEN78 TO -CONTROL
#COPY -SAM12 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5*-CONTROL 6*-OUT 8*-SAMPLE 9*-NEWSAM
#EXECUTION BEGINS 21:49:15
#EXECUTION TERMINATED T=14.111 $2.34
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)

```

```

> 415
> 416
> 417          139 FINISH
> 418  O  NORMAL END OF JOB.
> 419          139 CONTROL CARDS WERE PROCESSED.
> 420          0 ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O*T*
# EXECUTION BEGINS 21:50:20
# DO YOU WISH TO CONTINUE?... ENTER "Y" OR "N"
?Y
#SAVE -OUT T.STATS.FOR.78.7TH.GRADE
# FILE "T.STATS.FOR.78.7TH.GRADE(1)" ... HAS BEEN SAVED
#SAVE -NEWSAM RECODED.78.7TH.GRADE.SAMPLE
# FILE "RECODED.78.7TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
#STOP
#EXECUTION TERMINATED T=2.669 $ .39
#EMPTY -NEWSAM
#DONE.
#RUN *PAGEPR SCARDS=*PGF(20)*-OUT
#EXECUTION BEGINS 21:54:13
#PRINT* ASSIGNED RECEIPT NUMBER 617336
#PRINT* 617336 HELD
#PRINT* 617336 RELEASED TO CSC0(SP), 13 PAGES.
#EXECUTION TERMINATED T=0.276 $ .53
#COPY SEVENT9 TO -CONTROL
#COPY -SAM14 TO -SAMPLE
#SOURCE RUNSAM
# GET -OUT
#READY.
# EMPTY -OUT
#DONE.
# GET -NEWSAM
#READY.
# EMPTY -NEWSAM
#DONE.
# RUN *SPSS 5--CONTROL 6--OUT 8--SAMPLE 9--NEWSAM
#EXECUTION BEGINS 21:55:16
#EXECUTION TERMINATED T=13.421 $ 2.22
# EMPTY -CONTROL
#DONE.
# EMPTY -SAMPLE
#DONE.
# LIST -OUT(LAST-5)
> 397
> 398
> 399          121 FINISH
> 400  O  NORMAL END OF JOB.
> 401          121 CONTROL CARDS WERE PROCESSED.
> 402          0 ERRORS WERE DETECTED.
# END OF FILE
# RUN *FS O*T*
# EXECUTION BEGINS 21:56:19
# DO YOU WISH TO CONTINUE?... ENTER "Y" OR "N"
?Y
#SAVE -OUT T.STATS.FOR.79.7TH.GRADE
# FILE "T.STATS.FOR.79.7TH.GRADE(1)" ... HAS BEEN SAVED
#SAVE -NEWSAM RECODED.79.7TH.GRADE.SAMPLE
# FILE "RECODED.79.7TH.GRADE.SAMPLE(1)" ... HAS BEEN SAVED
#STOP

```

#EXECUTION TERMINATED T=2.581 8.37

#EMPTY -NEWSAM

# Assuming "EMPTY" for "EMPTY". OK?

70K

#DONE.

#RUN \*PAGEPR SCARDS=\*PGF(20)\*-OUT

#EXECUTION BEGINS 21:59:50

\*PRINT\* ASSIGNED RECEIPT NUMBER 617338

\*PRINT\* 617338 HELD

\*PRINT\* 617338 RELEASED TO CSCO(SP), 13 PAGES.

#EXECUTION TERMINATED T=0.27 8.53

#RUN \*PAGEPR SCARDS=\*PGF(20)\*-LOG

#EXECUTION BEGINS 22:00:22

\*PRINT\* ASSIGNED RECEIPT NUMBER 617339

\*PRINT\* 617339 HELD

## BIBLIOGRAPHY

- Anderson, E. B. The numerical solution of a set of conditional estimation equations. Journal of the Royal Statistical Society B, 1972, 34, 42-54.
- Anderson, B., Cooley, W., Holliday, A., Mosley, W. and Turnbull, A. B. Report of the Michigan Educational Assessment Program's External Advisory Panel on Evaluation. Lansing: Michigan Department of Education and National Assessment of Educational Progress, 1977.
- Anderson, J., Kearney, G. E. and Everett, A. V. An evaluation of Rasch's statistical model for test items. The British Journal of Mathematics and Statistical Psychology, 1968, 21 231-238.
- Baker, C. C. T. Directory of Mathematics. New York: Hart publishing Company, Inc. 1966.
- Brink, N. E. The effect of item discrimination and range of item easiness on the standard error of ability estimate using the Rasch model. Dissertation Abstracts. 3947-A, 1970.
- Brink, N. E. Rasch's logistic model vs. the Guttman model. Journal of Psychological Measurement, 1972, 32, 921-927.
- Cypress, B. K. The effects of diverse test score distribution characteristics on the estimation of ability parameter of the Rasch measurement model. Dissertation Abstracts, 2761-A, 1972.
- Curtledge, C. M. A comparison of equipercentile and Rasch equating methodologies. Dissertation Abstracts, 5141-A, 1977.
- Dayton, C. M. The Design of Educational Experiments. New York: McGraw-Hill Book Company 1970.
- Dinero, T. E. A computer simulation investigating the applicability of the Rasch model with varying item discriminations. Proceedings Annual Meeting of the National Council on Measurement in Education, San Francisco, 1976.
- Douglas, G. A. Test design for the Rasch psychological model. Dissertation Abstracts, 4427-A, 1975.
- Draba, R. E. The Rasch model and legal criteria of a "reasonable" classification. Dissertation Abstracts, 245-A, 1979.
- Education Laboratory. How to Use "BICAL", Chicago: University of Chicago, 1976.



- Epstein, K. I. An empirical investigation of criterion-referenced testing model. 17th Annual Conference of the Military Testing Association, Ft. Benjamin Harrison: United States Army, 1975.
- Good, C. V. (ed.) Dictionary of Education. 3 ed. New York: McGraw-Hill Book Company, 1973.
- Grunlund, N. E. Measurement and Evaluation in Teaching. 3 ed. New York: Macmillan publishing Co., Inc. 1976.
- James, R. C. (ed.) Mathematics Dictionary. 3 ed. Princeton: D. Van Nostrand Company, Inc. 1968.
- Laska, S. A. J. Influence of time of calibration on Rasch model item difficulties. Dissertation Abstracts, 3247-A, 1979.
- Magnussen, D. Test Theory. Reading: Addison-Wesley Publishing Company, 1967.
- Michigan Department of Education Staff. A Staff Response to the Report: An assessment of the Michigan Accountability System. Lansing: Michigan Department of Education, 1974.
- Michigan Educational Assessment Program Staff. Michigan Educational Assessment Program Grades 4 and 7 Item and Objective Handbook. Lansing: Michigan Department of Education, no date.
- Michigan Educational Assessment Program Staff. Student performance Expectations. Lansing: Michigan Department of Education, no date.
- Michigan Educational Assessment Program Staff. Questions and Answers About the Michigan Educational Assessment Program. Lansing: Michigan Department of Education, no date.
- Michigan Educational Assessment Program Staff. First Report, Objectives and Procedures 1974-75. Lansing: Michigan Department of Education, 1974.
- Michigan Educational Assessment Program Staff. Technical Report, Michigan Educational Assessment Program. Lansing: Michigan Department of Education, 1977.
- Michigan Educational Assessment Program Staff. Interpretive Manual 1978-79. Lansing: Michigan Department of Education, 1979.
- Michigan Educational Assessment Program Staff. The Status of Basic Skills Attainment in Michigan Public Schools. Lansing: Michigan Department of Education, 1979.

- Michigan Educational Assessment Staff. DRAFT: Communications Skills Objectives -- Reading -- Speaking/Listening -- Writing. Lansing: Michigan Department of Education, 1979.
- Minium, E. W. Statistical Reasoning in Psychology and Education. 2 ed. San Jose: John Wiley & Sons, 1978.
- O'Reilly, R. P., Schuder, R. T., Kidder, R. S., Salter, Hayford, P. D. The Validation and Refinement of Measures of Literal Comprehension in Reading for Use in Policy Research and Classroom Management. Albany: The University of the State of New York, The State Department of Education, Division of Research, 1976.
- Passmore, D. L. An application of the Rasch one parameter logistic measurement model to the national league for nursing achievement test in normal nutrition. Dissertation Abstracts. 963-A, 1974.
- Porter, J. W. The virtues of a state assessment program. Phi Delta Kappan, 1968, 57-10, 667-668.
- Porter, J. W. The accountability story in Michigan. Phi Delta Kappan, 1972, 54-2, 98-99.
- Porter, J. W. Spotlight on Michigan: what are we getting for our tax dollar? Compact, 1973, 7-5 19-21.
- Porter, J. W. Task force '74: recommendations for better schools. National Association of Secondary School Principals Bulletin, 1975a, 59-391, 19-24.
- Porter, J. W. If I were a school board member. Colorado Journal of Education Research, 1975b, 14-3, 2-7.
- Porter, J. W. Education: the challenging frontier. Colorado Journal of Education Research, 1976, 15-3, 18-22.
- Porter, J. W. Michigan's edu-checkup. Social Policy, 1977, 8-2 41-44.
- Porter, J. W. The limits of school power. Phi Delta Kappan, 1978, 59-5, 319-320.
- Public Law 95-561. Title I - Amendment to Title I of the Elementary and Secondary Education Act of 1965. 1978.
- Rasch, G. Probability Models for Some Intelligence and Attainment Tests. Copenhagen: Denmark's Paedagogiske Institute, 1960.
- Rasch, G. On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkly Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1961, 4, 321-333.

- Rasch, G. An individualistic approach to item analysis. P. F. Luzarsfeld and N. W. Henry (eds.), Reading in Mathematical and Social Science. Chicago: Science Research Associates, 1966a.
- Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966b, 19-1, 49-57.
- Rentz, R. R. and Basheur, W. L. Equating Reading Tests with the Rasch Model. Athens: Kesen Laboratory, College of Education, University of Georgia, 1976.
- Roeber, E. Personal interview, Lansing, Michigan, May 2, 1980.
- Ryan, J. P. and Hamm, D. W. Practical procedures for increasing the reliability of classroom tests by using the Rasch model. Proceedings, National Council of Measurement in Education, San Francisco, 1976.
- Wells, R. A. The probabilistic interpretation of test scores calibrated by the Rasch model. Dissertation Abstracts, 4012-A, 1973.
- Whitely, S. E. and Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11-12, 163-178.
- Whitely, S. E. Models, meanings and misunderstandings: some issues in applying Rasch's theory. Journal of Educational Measurement, 1977, 14-3, 227-235.
- Willmott, A. S. and Fowles, D. E. The Objective Interpretation of Test Performance. Bootle: NFER Publishing Company Ltd., 1974.
- Wright, B. D., Panchapakesan, N. A. A procedure for sample-free item analysis. Educational and Psychological Analysis, 1969, 29, 23-48.
- Wright, B. D., Mead, R. J. CALFIT: Sample-Free Item Calibration with a Rasch Measurement Model, Research Memorandum Number 18. Chicago: Mesa Press, 1975.
- Wright, B. D. Misunderstanding the Rasch model. Journal of Educational Measurement, 1977a, 14-3, 219-225.
- Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977b, 14-2, 97-116.
- Wright, B. D., Mead, R. J. BICAL: Calibrating Items and Scales with the Rasch Model, Research Memorandum Number 23. Chicago: Mesa Press, 1977c.

Wright, B. D., Stone, M. H. Best Test Design: Rasch Measurement. Chicago: Mesa Press, 1979.

Wright, B. D., Mead, R. J. BICAL: Calibrating Items with the Rasch Model, Research Memorandum Number 23c. Chicago: Mesa Press, 1980.

ABSTRACT

USING THE RASCH MODEL TO EVALUATE TEST ITEMS FOR  
GRADE 4 AND GRADE 7 MICHIGAN EDUCATIONAL  
ASSESSMENT PROGRAM CRITERION-REFERENCED READING  
TESTS ADMINISTERED 1973 THROUGH 1979

by

DONALD JOHN MCPHERSON

April, 1983

Advisor: Donald Marcotte, Ph.D.

Major: Educational Evaluation and Research

Degree: Doctor of Philosophy

This study investigates whether or not Rasch measurement is appropriate in connection with tests given under the Michigan Educational Assessment Program (MEAP)? These are criterion referenced tests which measure reading and mathematics achievement of students at the fourth and seventh grade levels. Limitations of score variability is a desirable outcome and scores tend to be strongly skewed on the high side. Therefore, traditional test evaluation methods are likely to be less effective when used in the analysis of MEAP scores than might be expected if they had been designed as norm referenced tests. The Rasch model appears to offer a better standard of measurement in this type of situation. The model is named for a Danish mathematician who originally posed the concept in 1960: Georg Rasch. Rasch devised a simple two-parameter model (i.e., item-difficulty and person-ability) that employs raw test scores directly as measures of achievement. BICAL, a computer program developed by the Measurement and Statistical Laboratory (MESA) of the Department of Education at The University of Chicago to perform

Rasch analysis, is applied in this study to scores from fourteen MEAP reading tests taken by fourth grade and seventh grade students from 1973 through 1979. A "fit statistic" is generated by the program which is used to determine the fit of MEAP reading test items to the Rasch model. The statistic is interpreted as an F-statistic with one and five degrees of freedom at an alpha level of 0.05. Items that did not fit the Rasch model were found in all but one of the fourteen tests considered in this investigation. The results of the analysis that was done indicate that the use of Rasch measurement may be appropriate in connection with MEAP reading tests in the development of test items and as a means of measuring improved or declining achievement over time. However, while Rasch measurement seems to promise truly objective measurement, there is a real possibility that the Rasch measurement model may not, in practice, be easy to use.

## AUTOBIOGRAPHICAL STATEMENT

Donald J. McPherson was born in Detroit, Michigan on January 1, 1931. He received his college preparatory high-school education at Culver Military Academy, Culver, Indiana where he graduated in June, 1949. He attended Michigan State University, Lansing, Michigan, one year then completed his Bachelor of Science in Business Administration, with a major in Marketing, June, 1954 at Wayne State University, Detroit, Michigan. He completed his Master of Education degree at Wayne State University in June, 1980, and is presently pursuing the degree of Doctor of Philosophy in Evaluation and Research at Wayne State University. He is also employed as Project Manager in the Data Processing Department of Comprehensive Health Services of Detroit, a large health maintenance organization (HMO). Mr. McPherson has an extensive background in data processing, systems development, and education. He has worked as Systems Analyst for the J. L. Hudson Company, Detroit, Michigan; Instructor and Program Director in Data Processing at Ferris State College, Big Rapids, Michigan; Director of Education for the Association for Systems Management, Cleveland, Ohio; Education Director for the Data Processing Management Association, Park Ridge, Illinois; Project Manager for the American Nuclear Society, Hinsdale, Illinois; and Managing Director of the Hearing Instruments Institute, Livonia, Michigan. His primary interests lie in the variety of forms information systems take, and their implementation, in an ambulatory health care setting. He is a member of the American Statistical Association, the American Public Health Association, the Detroit Business Micro User's Association, and Phi Delta Kappa.