AN APPLICATION OF THE RASCH LOGISTIC MODEL TO THE ASSESSMENT OF CHANGE IN MATHEMATICS ACHIEVEMENT

by

THOMAS JOE O'SHEA

B. Eng., McGill University, 1960
B. Ed., University of Saskatchewan, 1968
M. Ed., University of Manitoba, 1976

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF EDUCATION

in

THE FACULTY OF GRADUATE STUDIES
(Mathematics Education)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 1979

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics Education

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date October 12, 1979

ABSTRACT

Research Supervisor: Dr. D. F. Robitaille

The purpose of this study was to explore the use of the Rasch simple logistic model for the measurement of group change in mathematics achievement. A survey of previous studies revealed no consensus as to the performance of present-day students compared with their counterparts in the past. Few studies attempted to identify changing performance on specific topics within the mathematics curriculum. Groups were compared most commonly on the basis of grade equivalents or raw scores. Neither of these is satisfactory for measuring change; grade equivalents have no natural statistical interpretation, and raw scores yield ordinal rather than interval measures. A possible solution to the problem of scale lay in the use of the Rasch model, since it purports to yield measures of item difficulty and person ability on a common interval scale.

In 1964 and in 1970, all Grade 7 students in British Columbia wrote the Arithmetic Reasoning and Arithmetic Computation tests from the Stanford Achievement Test, Advanced Battery, Form L (1953 Revision). Random samples of 300 test booklets were available from each administration. In 1979, the same tests were administered to a sample of 50 Grade 7

classes, stratified by geographic region and size of school, selected from schools across the province. A random sample of 300 test booklets was drawn from the tests completed in 1979.

The reasoning and computation tests contained 45 and 44 items, respectively. The items were reclassified by the researcher into ten content areas as follows:

1. Whole number concepts and operations (WNC) - 11 items
2. Applications using whole numbers (WNA) - 9 items
3. Common fraction concepts and operations (CFC) - 12 items
4. Applications using common fractions (CFA) - 8 items
5. Decimals (Dec) - 11 items
6. Money (Mon) - 9 items
7. Percent (Pct) - 8 items
8. Elementary algebra (Alg) - 9 items
9. Geometry and graphing (Geo) - 8 items
10. Units of measure (Mea) - 4 items

The computer program BICAL was used to determine estimates of item difficulties and person abilities. A minimum cutoff score was established to eliminate examinees near the guessing level. An item was deemed non-fitting if its fit mean square exceeded unity by four or more standard errors and its discrimination index was less than 0.70.

One of the key requirements of the study was to demonstrate that the two Stanford tests measured the same ability, thereby justifying the regrouping of the items. To this end, for each year a standardized difference score between Rasch ability on the reasoning test and on the computation test was calculated for each person. The

distribution of such scores was compared with the unit normal distribution, using the Kolmogorov-Smirnov statistic. Since no distribution differed from the unit normal at the 0.01 level of significance, the tests were assumed to measure the same ability.

All 89 items were then calibrated as a whole for each year. Items were deleted from the analysis if they showed lack of fit on two of the three administrations. The deletion process was terminated after two recalibrations, with 10 items eliminated.

For each pair of years, two standardized difference scores for the difficulty of each item were calculated: one reflected relative change of difficulty within the curriculum, and the other reflected absolute change of difficulty. For each content area the mean difficulty and standard error of the mean were calculated, and the standardized difference of the mean was determined for each comparison.

The small number of items subsumed under Units of Measure precluded any reliable conclusions on this topic. Of the remaining nine topics only Elementary Algebra showed any relative change of difficulty; from 1964 to 1970 it became easier, both relatively and absolutely, probably due to increased emphasis on this topic within the curriculum. The topic Percent was more difficult in 1970 than in 1964. From 1970 to 1979, Elementary Algebra and both topics dealing with common fractions became more difficult. Overall, from 1964 to 1979, five of the nine topics became more difficult: WNA, CFC, CFA, Dec, and Pct. No topic became less difficult.

A comparison of decisions using the Rasch model and using the traditional model based on p-values showed the Rasch model to be more conservative. For example, from 1964 to 1979, all nine topics would have been judged more difficult by using the traditional model. It was suggested that the differing decisions were due to the differing behaviours of the standard error of estimate in the two models. In the Rasch model, items of average difficulty are calibrated with the least standard error, while in the traditional model the standard error for items of greatest and least difficulty are estimated with the least standard error. The question of which is preferable was unresolved.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

I wish to express my thanks to my committee chairman, Dr. David Robitaille, for the way in which he supervised the production of this dissertation. He was generally supportive, sometimes demanding, and at all times, he showed confidence in my ability to pursue the topic as I saw fit.

I would also like to thank my committee members: Dr. Merle Ace, for guidance in the use of the Rasch model, Dr. Todd Rodgers, for clarification of statistical issues, Dr. James Sherrill, for detailed textual comments, and Dr. Gail Spitler, for placing the study in the context of general educational issues.

Financial support for the study was provided through a grant from the Educational Research Institute of British Columbia, and their support was much appreciated.

Finally, I wish to acknowledge the contribution of Peg McCartney whose support was a constant source of strength. Her loss changed the pursuit of the doctorate from a joy to a job.

CHAPTER I

STATEMENT OF THE PROBLEM

It is currently fashionable to lament the level of mathematical knowledge possessed by graduates of public educational institutions (e.g., Beltzner, Coleman, & Edwards, 1976). Moreover it is a common belief that schools were more successful in teaching fundamental mathematical skills at some point in the past (e.g., Armbruster, 1977). Objective evidence which might confirm such a belief is difficult to find.

Critics are divided as to the existence and implications of objective measures of mathematical achievement. For example, in a recent book entitled Why Johnny Can't Add: The Failure of the New Math, Morris Kline (1973) attacked the curriculum reforms of the 1960's. In spite of the suggestiveness of the title, Kline failed to document such a failure. In particular, one might have expected his chapter entitled "The Testimony of Tests" to contain data supporting his implied contention that computational proficiency had declined. On the contrary, his main argument in that chapter is that no suitable tests or

testing programs have yet been devised tc carry out the necessary measurements.

The most ambitious attempt to date to assess the state of mathematics education in the United States is the report (1975) of the National Advisory Committee on Mathematical Education (NACOME). The committee reviewed curriculum reforms from 1955 to 1975, identified new curricular emphases in current programs, outlined various patterns of instruction in use, addressed the problem of teacher education, and, finally, tackled the question of evaluation. Their conclusions concerning achievement are overshadowed by the reaction of the committee to the evaluation procedures themselves: "Unfortunately, evaluation in American mathematics education is characterized by use of limited techniques inappropriately matched tc goal assessment tasks" (p. 119). The committee recommended that the use of grade-equivalent scores be abandoned. They argued that testing samples of students would avoid the problem of unjustified over-testing. They suggested that the use of standard norm-referenced tests to assign an overall measure of performance was not appropriate for programs with specific goals. They preferred the development cf suitable collections of test items on carefully constructed, objective-directed test scales. They concluded:

> To make evaluation play a positive and effective role in school mathematics today there is an urgent need to develop a much broader collection of measurement techniques and instruments and to match these evaluation tools more appropriately to the varied purposes of evaluatien. (p. 135)

Background to the Problem

In 1973 the Science Council of Canada agreed to fund a project proposed by six collaborating national mathematics societies. The aim of the project was to take inventory of the mathematical sciences in Canada and to formulate policy in the national interest. The study was completed in 1975 and published a year later as Mathematical Sciences in Canada (Beltzner et al., 1976).

Chapter IV of the study dealt with the teaching of mathematics in Canadian elementary and secondary schools. The authors concluded that "the overall picture in Canada at present contains so much distress, unease and confusion that energetic steps must be initiated immediately to improve the situation" (p. 114). This conclusion appears to be based on two classes of evidence: (1) opinion expressed in briefs and by individuals, and (2) objective data. Regarding the latter:

> The most convincing objective piece of information which was presented to the Mathematics Study consisted of a Report on the Testing of Arithmetic issued by the Department of Education of British Columbia. (pp. 113-114)

This same report also formed the main subject of the 1969-1970 annual report of the Director of the Research and Standards Branch of the British Columbia Department of Education (1971).[1] In his report, the Director summarized some results of two arithmetic testing programs which had been

---

[1] The British Columbia Department of Education was reorganized and renamed the British Columbia Ministry of Education in 1976-77. The terms "Department" and "Ministry" are interchangeable in this study.

carried out in 1964 and in 1970. His summary dealt with the analysis of responses to each test item, and with the performance of the British Columbia students as a whole ccmpared with their United States counterparts. He concluded that there were evident computational difficulties that indicated "a need for a return to that neglected and unpopular procedure called 'repetition' or 'drill'" (p. G62). He pcinted out, however, that the students in 1964 had scored considerably better than the United States standardization group and, although performance had declined, the 1970 British Cclumbia scores corresponded approximately to American test norms.

The study carried out by the British Columbia Department of Education has been influential both at the provincial level where it was used to bolster arguments for a return to "drill", and at the national level where it was cited as objective evidence of a currently unsatisfactory state of affairs in mathematics education. To warrant such influence it is reascnable to assume that the study was well-documented and founded on a solid inferential base. This may not be the case as the following discussion will show.

Reported Change in British Cclumbia

The 1970 Report on the Testing of Arithmetic issued by the British Columbia Department of Education constitutes Appendix A of this study. It should be referred tc for details. What follows here is a general description of that

study and a critique of its procedures.

In March 1964 the Department of Education administered the Stanford Achievement Test, Advanced Battery, Partial, Form L, to the population of Grade 7 students of British Columbia. Completed forms were returned by 29 204 students out of an estimated enrolment of 29 533. The Stanford Achievement Test has a long history with many revisions dating back to 1923. The edition used for the British Columbia study was the 1953 revision which was standardized in the spring of 1952 on a norm sample representative by geographic region and by size of school system in the U.S.A., excluding pupils in segregated Negro systems (Kelley, Madden, Gardner, Terman, & Ruch, 1953).

The content of the Stanford battery was based on the curriculum of American schools of the late 1940's. The battery consisted of six tests: Reading was measured by two tests--Paragraph Meaning and Word Meaning; Language and Spelling were each measured by a single test; Arithmetic was measured by two tests--Arithmetic Reasoning and Arithmetic Computation. The battery was administered in four sittings, each of approximately forty minutes duration, over four days. The two arithmetic tests required one sitting each. They are contained in Appendix B.

The 1964 testing program was undertaken to assist in establishing reasonably consistent standards across the province. Each classroom teacher received a pupil report and a class listing. Summaries were prepared for each classroom, each school, and each school district. Computer programs were

used to determine distributions of scores and to calculate percentiles (Conway, 1964).

A random sample of three hundred completed test batteries was drawn with one hundred from each of the upper third, middle third, and lower third as defined by total score on the battery. The one hundred papers for each of the upper third and lower third were used to calculate item difficulties and validities for all the items on the six tests. The difficulty value was defined to be the percentage of the two hundred respondents who either failed to respond to the item or who gave an incorrect response. The validity figure for each item was determined by subtracting the percentage of respondents in the lower third sample who responded correctly from the percentage of those in the upper third sample who responded correctly. This is abbreviated as U-L% in the report. These two calculations appear to have been customary with the Research and Standards branch in all its testing programs. The three hundred test papers were filed along with data sheets showing details of the tabulations and calculations.

In May 1970 the Department readministered the two arithmetic tests from the same battery to all Grade 7 students in British Columbia. From the estimated 40 252 students enrolled, 38 377 completed forms were returned. "The purpose of the second administration was to determine the changes that had occurred in achievement in the ordinary arithmetic type of item" (British Columbia Department of Education, 1970, p. 1).

A procedure similar to that of 1964 was followed in

analyzing the data in 1970. It is not clear, however, whether an analysis was made for each classroom, school, and district as in the previous administration. Modal-age grade equivalents based on the 1952 U.S. norms were again determined for the population. The excess over U.S. modal-age grade equivalents was found to have dropped by 0.9 years on the reasoning test and by 1.1 years on the computation test.

Validity figures for each item were determined as in 1964. Again three hundred papers were drawn as a stratified random sample. This time, however, the one hundred papers for the middle third were used as well as the one hundred for each of the upper and lower thirds for determining the item difficulties. The tables on pages 5 and 6 of Appendix A set out the values obtained on the two administrations.

With respect to the item analysis all that has been discussed to this point is the computational procedure. Now the essence of the problem may be delineated. The purpose of the 1970 program was to assess change in performance since 1964. This was done in two ways. The first was to compare the grade equivalent means. The second was to compare the difficulty of each item as determined in 1964 and in 1970. Conclusions were drawn with respect to items which had changed in difficulty, with respect to areas of the curriculum which had become more difficult, and with respect to the reasoning processes of students based on patterns of item difficulty. The question arises as to what criteria were used to decide that a change had indeed taken place in the difficulty of an item.

It appears from the tables on pages 5 and 6 in Appendix A that a change of more than 2% on the reasoning test and of more than 3% on the computation test was considered to reflect a true change in item difficulty. The exceptions to this rule are items 2 and 4 on the reasoning test and item 5 on the computation test. A summary of the numbers of changes is given in Table 1.1.

Table 1.1

Summary of Reported Changes in Difficulty Value

Test	No. of Items Decreasing in Difficulty	No. of Items Increasing in Difficulty
Reasoning	12	17
Computation	4	26

Of critical importance is the fact that nowhere in the report itself was there any mention that the item difficulty values were determined on the basis of <u>samples</u>. That information was obtained only upon examination of the files which contained detailed summary sheets of the sample responses. The 1964 figures were based on a sample of just two hundred papers even though a further one hundred were available. In 1970 the Department decided to use the additional one hundred papers from the middle third to obtain more representative item difficulties on the 1970

administration. Since the figures were sample based there should have been an attempt to take that fact into account when deciding which items had significantly changed in difficulty. A need to reanalyze the data is clearly indicated by this oversight.

Tables 1.2 and 1.3 summarize the results of the reanalysis of the data carried out for the present study. The 1964 difficulty figures were determined using all three hundred available papers. Item responses were tabulated directly from the original test papers and a comparison was made with the Department's summary sheets. Some discrepancies were found. An examination of individual papers showed some scoring errors. In most of these cases the item had been designated correct although more than one response had been marked by the student. In re-marking the papers, where the evidence was strong that the student favoured the correct alternative, the item was marked correct. Where it was not clear which alternative the student wished the marker to accept, the item was marked incorrect. As a result of this procedure a total of 35 changes were made. No item differed from the original tally by more than 2 out of 200. The same procedure was followed in recording the information from the 1970 sample test papers. Six scoring changes were made. The difficulty figures for each year were extended to one decimal place. In the Ch* column of the tables the key is as follows:

+ : more difficult in 1970

- : less difficult in 1970

0 : no change in difficulty.

Table 1.2

Item Difficulties on the Reasoning Test

Item No.	Original Analysis %-Diff 1964	Original Analysis %-Diff 1970	Ch*	Reanalysis 1964 %-Diff	Reanalysis 1964 SE	Reanalysis 1970 %-Diff	Reanalysis 1970 SE	Ch*
1	7	5	0	4.7	1.2	4.7	1.2	0
2	3	6	0	3.7	1.1	5.7	1.3	0
3	4	5	0	2.7	0.9	5.0	1.3	0
4	15	13	-	14.0	2.0	13.3	2.0	0
5	12	7	-	10.0	1.7	7.0	1.5	0
6	10	7	-	8.3	1.6	7.3	1.5	0
7	20	20	0	19.7	2.3	20.3	2.3	0
8	11	17	+	12.3	1.9	17.0	2.2	0
9	17	22	+	15.3	2.1	21.7	2.4	+
10	18	14	-	19.0	2.3	13.7	2.0	0
11	12	14	0	9.7	1.7	14.0	2.0	0
12	18	17	0	18.7	2.3	17.3	2.2	0
13	18	17	0	14.7	2.0	16.7	2.2	0
14	39	38	0	36.7	2.8	37.7	2.8	0
15	22	25	+	22.3	2.4	24.7	2.5	0
16	25	30	+	21.3	2.4	30.0	2.7	+
17	20	21	0	19.0	2.3	21.0	2.4	0
18	23	26	+	22.0	2.4	26.3	2.6	0
19	35	34	0	31.0	2.7	33.7	2.7	0
20	34	36	0	31.7	2.7	36.3	2.8	0
21	27	25	0	23.7	2.5	25.0	2.5	0
22	20	37	+	18.3	2.2	36.7	2.8	+
23	21	24	+	20.3	2.3	23.7	2.5	0
24	58	58	0	60.0	2.8	57.7	2.9	0
25	50	54	+	50.0	2.9	54.3	2.9	0
26	39	44	+	41.7	2.9	44.0	2.9	0
27	43	50	+	37.3	2.8	50.3	2.9	+
28	64	61	-	65.3	2.8	61.0	2.8	0
29	73	79	+	74.0	2.5	78.9	2.7	0
30	64	71	+	64.3	2.8	71.0	2.6	0
31	14	10	-	12.7	1.9	9.7	1.7	0
32	36	36	0	35.0	2.8	36.7	2.8	0
33	20	11	-	17.0	2.2	11.0	1.8	-
34	57	52	-	56.7	2.9	51.3	2.9	0
35	17	38	+	15.7	2.1	38.3	2.9	+
36	34	36	0	34.7	2.8	36.3	2.8	0
37	35	43	+	36.0	2.8	42.7	2.9	0
38	20	16	-	22.0	2.4	15.7	2.4	0
39	28	37	+	28.3	2.6	37.3	2.8	+
40	46	45	0	49.0	2.9	45.3	2.9	0
41	58	51	-	60.0	2.8	51.0	2.9	-
42	30	34	+	30.0	2.7	34.3	2.7	0
43	70	64	-	69.0	2.7	64.3	2.8	0
44	59	55	-	59.7	2.8	54.7	2.9	0
45	63	72	+	67.7	2.7	72.0	2.6	0

Table 1.3

Item Difficulties on the Computation Test

Item	Original Analysis			Reanalysis				
	%-Diff			1964		1970		
No.	1964	1970	Ch*	%-Diff	SE	%-Diff	SE	Ch*
1	3	9	+	2.3	0.9	9.0	1.7	+
2	12	7	-	9.7	1.7	7.0	1.5	0
3	7	12	+	7.3	1.5	11.7	1.9	0
4	2	2	0	2.0	0.8	1.7	0.7	0
5	10	13	+	10.0	1.7	13.0	1.9	0
6	14	9	-	12.7	1.9	8.7	1.6	0
7	17	16	0	14.7	2.0	16.3	2.1	0
8	12	11	0	9.7	1.7	11.3	1.8	0
9	13	17	+	12.3	1.9	17.3	2.2	0
10	11	15	+	8.3	1.6	15.0	2.1	+
11	20	34	+	16.3	2.1	33.7	2.7	+
12	10	23	+	10.0	1.7	23.0	2.4	+
13	32	47	+	29.0	2.6	47.0	2.9	+
14	32	50	+	29.0	2.6	50.0	2.9	+
15	16	19	0	14.7	2.0	19.0	2.3	0
16	26	45	+	24.7	2.5	45.0	2.9	+
17	30	31	0	32.3	2.7	31.3	2.7	0
18	23	45	+	23.3	2.4	45.0	2.9	+
19	32	47	+	28.3	2.6	47.0	2.9	+
20	28	46	+	25.3	2.5	46.3	2.9	+
21	11	9	0	9.7	1.7	9.0	1.7	0
22	38	46	+	32.0	2.7	46.0	2.9	+
23	21	18	0	24.7	2.5	18.0	2.2	-
24	45	53	+	46.3	2.9	53.3	2.9	0
25	25	22	0	25.3	2.5	22.0	2.4	0
26	42	48	+	43.0	2.9	48.3	2.9	0
27	33	46	+	27.7	2.6	46.0	2.9	+
28	31	44	+	29.7	2.6	44.0	2.9	+
29	32	28	-	28.3	2.6	27.7	2.6	0
30	17	17	0	16.3	2.1	17.0	2.2	0
31	17	31	+	16.0	2.1	31.3	2.7	+
32	54	59	+	50.0	2.9	59.0	2.9	+
33	28	42	+	30.3	2.7	42.0	2.9	+
34	60	79	+	61.0	2.8	79.0	2.4	+
35	23	31	+	22.7	2.4	30.7	2.7	+
36	5	8	0	4.7	1.2	7.7	1.6	0
37	62	39	-	64.0	2.8	39.0	2.8	-
38	66	76	+	66.3	2.7	75.7	2.5	+
39	71	72	0	73.7	2.5	72.3	2.6	0
40	88	88	0	86.3	2.0	88.3	1.9	0
41	54	63	+	56.7	2.9	63.3	2.8	0
42	47	46	0	49.0	2.9	46.3	2.9	0
43	76	75	0	77.3	2.4	74.7	2.5	0
44	80	91	+	85.0	2.1	91.3	1.6	+

The standard error (SE) for each item in the reanalysis was computed from the formula:

$$SE = \sqrt{\frac{d(100 - d)}{n - 1}}$$

where d is the %-difficulty of the item, and

n is the sample size.

The difference of the item difficulties was determined as well as the standard error of the differences, ($\sqrt{SE(64)^2 + SE(70)^2}$).[2] A change in item difficulty was deemed to have occurred when the difference in item difficulty estimates exceeded two standard errors. A summary of changes is given in Table 1.4. These may be compared with the original results in Table 1.1.

Table 1.4

Summary of Changes in Difficulty Values After Reanalysis of the Data

Test	No. of Items Decreasing in Difficulty	No. of Items Increasing in Difficulty
Reasoning	2	6
Computation	2	20

[2] In this study, subscripted variables are represented by placing what would be subscripts in parentheses.

The Interpretation of Change

Although the stated purpose of the 1970 project was to determine change, the Department also attempted tc analyze specific weaknesses. In particular, efforts were made to identify types of errors, for example, "zero difficulty", and to ascribe reasons for lower performance on specific items, for example, lack of use of analogy. Overall, the Department appears to have attempted to interpret change at three different levels.

The first interpretation of the results was in terms of the global abilities "reasoning" and "computation". Comparisons were made by determining median modal-age grade equivalents on the two administrations. This approach has several deficiencies. In the first place, the NACOME report (1975) recommended the abandcnment of grade-equivalent scores for a number of reasons: they have no natural statistical interpretation, they lead to the popular expectation that all students should perform at grade level or above, and they are open to the misinterpretation that children scoring above grade level could perform satisfactorily at the grade level indicated. In the second place the value to practitioners of knowing how well their students perfcrmed on tests of mathematical "reasoning" and mathematical "computation" provides little guidance as to what actic n should be taken in the classroom. This is particularly so in this instance since the Department's report (1970) itself notes: "The sub-tests overlap to a certain extent; almost all 'reasoning' items require some skill in computation and many of the

'computation' items require a certain amount of problem-solving ability" (p. 1).

A second interpretation of the results was in terms of students' ability to answer specific items correctly. In this instance each item is taken to be representative of an entire class of items. In general, for example, one is not interested in whether children can correctly determine the value of (1/4)/(1/2) or not, but whether they can perform the operation of division on unit fractions. This raises the problem of item peculiarity: the results may have been different if the specific item had been (1/6)/(1/3). The current move toward mastery learning and criterion-referenced testing in which a student responds to a number of similar items, indicates that the use of the traditional form of survey test to deal with very narrowly defined curricular areas is inadequate.

Finally, the results were interpreted in terms of performance in content areas. For example, the Director concluded that "pupils ... do much more poorly, however, in problems involving interest rates or fractions" (British Columbia Department of Education, 1971, p. G62). The Director did not cite specific evidence to support his conclusion of poorer performance on fractions. His report did show that, of the twenty-six items on the computation test for which the Department claimed poorer performance in 1970, five dealt with fractions. Yet there were four other items dealing with fractions on the same test for which performance was unchanged. Furthermore, there were nine more questions on the

reasoning test in which operations on fractions were required but none of these items showed changes in performance over time. Clearly there is a problem of establishing criteria on which to base such generalizations.

The value of grouping items on the Stanford Achievement Test appears to have been recognized. In the 1973 revision of the test the publishers provide an item analysis service (Rudman, 1977). The national "p-value", that is, the percentage of students who answered the item correctly, is given for each item along with p-values for the class, school, and system. Each of the latter three is specially marked if it differs significantly from the national value. The items are grouped by instructional objective and item group mean p-values determined for each of the four reporting categories. For example, at the grade four level, a number of items are grouped under the objective "addition and subtraction algorithms"; others, under "multiplication and division algorithms" (Rudman, 1977, p. 181).

The analysis of change based upon groups of items having some common mathematical underpinning would be valuable for the practitioner. It would identify areas of relative strength and weakness in manageable curricular units. Change in the overall difficulty of the groups of items could be determined by comparing group means across time.

The Problem of Scale

A fundamental requirement underlying the calculation of statistics such as the mean and standard deviation is that of an equal interval scale. On such a scale equal differences in the measures correspond to equal differences in the amount of the underlying attribute. Physical measurement scales such as length in centimetres and temperature in degrees Celsius are examples. One can determine not only that individuals differ in height but also by how much they differ. Ordinal measurement, on the other hand, allows only the arrangement of objects or individuals on a ranked basis without knowledge of true quantitative differences.

The scales which underly the measurement of achievement are not easy to classify in clear-cut terms. Glass and Stanley (1970, p. 13) suggest, for example, that I.Q. scores might be called "quasi-interval". Three people having I.Q.'s of 50, 110, and 120 yield more than a simple ranking--it would also be expected that the second and third individuals are much more alike with respect to I.Q. than the first and second. Glass and Stanley do not offer recommendations on how to deal with such measures.

Ahmann and Glock (1971, p. 246) give a hypothetical example of a standardized vocabulary test on which two pairs of students achieve raw scores of 68 and 88, and 17 and 37, respectively. They point out that it is unlikely that the 20 raw-score unit difference in each case is indicative of the same true spread in vocabulary: "It is all too true that educational measurement habitually yields somewhat unequal

units" (p. 246). They suggest that the increase from 68 to 88 is likely a greater achievement than an increase from 17 to 37. "It is typical of achievement tests to find the 'rubber units' contracted at the lower end of the raw-score distribution and expanded at the upper end" (p. 246).

Fischer (1976) states the problem more graphically. In Figure 1.1, the horizontal axis represents a certain cognitive ability, and the vertical axis is the expected raw score on a test which measures this ability. The solid curve is the graph of the function relating test score to ability. X(1) and X(2) represent the abilities of two children before "treatment" and X(1)' and X(2)' represent their abilities after treatment. Assuming that they have derived exactly the same benefit from the treatment, the difference in their abilities is constant.

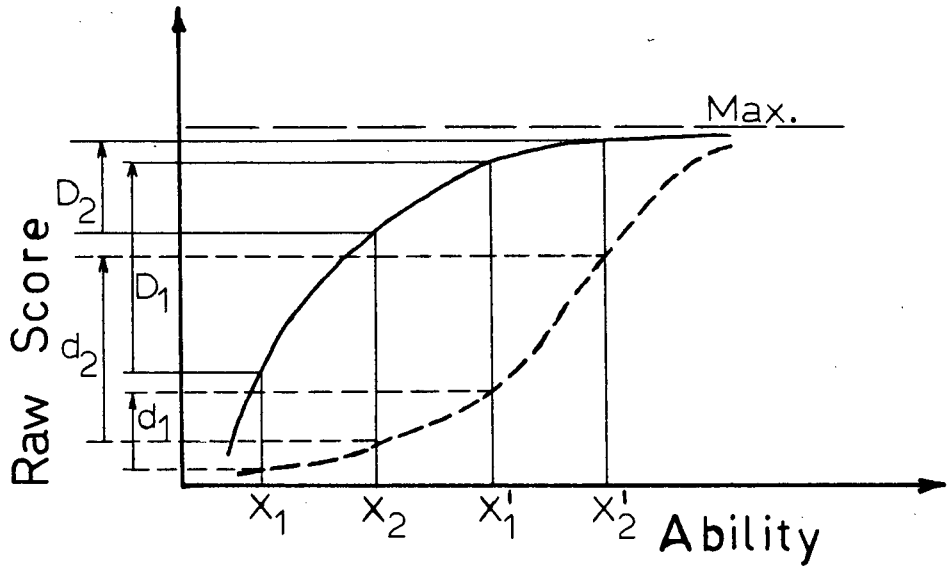


Figure 1.1. A hypothetical example of the relationship between raw score and ability.

Although the benefit of treatment is the same for each individual the raw score difference will not reflect this fact. Since D(1) is greater than D(2) the lower ability child will appear to have gained much more. If, however, the easy test items are replaced by more difficult ones the function relating ability and test score might be the dashed curve. The corresponding raw scores would show d(2) greater than d(1) implying that the higher ability person shows greater improvement. Hence the interpretation of raw scores is dependent upon the distribution of the difficulties of items making up the test.

Kerlinger (1964, p. 427) adopts a pragmatic point of view. He suggests that the best procedure is to treat ordinal measurements as though they were interval measurements except for the case of gross inequalities of intervals. He further advises that the interpretation of statistical analyses based on interval measures where the data are basically ordinal be made cautiously. He suggests that the competent research worker should be aware of transformations which can change ordinal scales into interval scales when there is serious doubt as to interval equality.

If one is willing to assume that the normal distribution underlies the observed variable measures, the problem can be resolved. Standard procedures can be used to perform a nonlinear transformation of raw scores (e.g., Magnusson, pp. 235-238). However, Stevens (1951) characterizes such usage as an act of faith, pointing out that: "It is certainly not unreasonable to believe that this

faith is often justified. What haunts us is the difficulty of knowing when it isn't" (p. 41).

An Alternative Approach

In 1960, Georg Rasch of the Danish Institute for Educational Research proposed several probabilistic models for the analysis of intelligence and achievement tests (Rasch, 1960). The particular model to be applied in the present study is known as the simple logistic model. Rasch, using the results of a Danish military test, argued from his data that items on the test could be ordered by their degree of difficulty as indicated by their percentage of correct sclutions. He argued further that the respondents could be ordered by their ability to solve the test items as indicated by their raw scores on the test. The problem which Rasch faced was that of setting up a model which allowed measurement on a ratio scale rather than on an ordinal scale.

Raw scores or percentages based upon raw scores are inadequate for a ratio scale in which one wishes to be able to state, for example, that person A has twice the ability as person B. If A obtains a score of 60% and B obtains a score of 30% and it is suggested that A has twice the ability of B, then it is not possible, on the same basis, to determine the score of person C who has twice the ability cf A. The same argument applies in the case of item difficulties.

In setting up the model, Rasch considered several desirable characteristics. The relative abilities of two

individuals should be uniquely determined regardless of the particular item used as a stimulus. Correspondingly, the relative difficulties of two items should not depend on the ability of the particular person to whom they were administered. Furthermore, if one person was twice as able as another and one question was twice as difficult as another, the more able person should solve the more difficult problem with the same "expenditure of effort" as the less able solved the less difficult (Rasch, 1960, p. 73).

For a very capable person faced with a very easy item, the probability that the person would be able to solve the problem should be very nearly unity. Only factors such as fatigue or boredom should result in a wrong answer. By the same token, a very dull person should have very little chance of correctly answering a very difficult item. Finally, for a problem neither too easy nor too difficult for the respondent, the outcome should be uncertain, and one would expect the probability of a correct solution to be around 0.5.

In general, the probability of a correct solution is a function of the ratio A/D, where A is the ability of the person and D is the difficulty of the item. Setting A/D = R, the problem becomes a matter of selecting a function of R such that the requirements of the preceding paragraphs are met. The simplest function occurring to Rasch was the function R/(1+R). Substituting A/D for R results in the expression: A/(A+D). Inspection reveals that for A equal to zero the probability of solving an item is zero. For all persons of high A meeting questions of low D, the probability is close to

unity for correct solutions. And, if A=D the probability is 0.5. Finally, the probability that a person will not solve the problem correctly is 1 - A/(A+D) = D/(A+D).

Rasch extended the model to a hypothetical situation in which a test of k items is administered to a group of n persons. He found that the model had several important properties. The estimate of a person's ability, A, can be derived solely from his or her raw score, r, regardless of which items contributed to that score. Secondly, the estimate of an item's difficulty, D, can be derived from the number of times the item was correctly answered without regard for which persons solved the item correctly (Rasch, 1960, pp. 76-77).

Finally, for the specific problem at hand, there is one result of fundamental importance deriving from the Rasch model: the difficulty/ability scale is an equal interval scale. The traditional p-values and test raw scores serve as sufficient estimators for assigning positions on the common underlying metric both to items and to persons.

Purpose of the Study

The purpose of this study, then, was to apply the Rasch model in an attempt to determine change in the mathematics achievement of Grade 7 students in British Columbia. This was accomplished in several steps.

(1) The Rasch model was applied to the data on hand from the 1964 and 1970 administrations of the Stanford Achievement Tests to determine item difficulties for each

year. Individual item difficulties were compared using a procedure elucidated in Chapter III. Items were grouped according to content into a number of meaningful and mutually exclusive subsets. Mean item difficulties for each group were compared from 1964 to 1970 in order to draw from the data conclusions more general than those made possible by simple item comparisons.

(2) In order to gain some insight into the performance of students in 1979, the 1964 and 1970 tests were administered to a sample of Grade 7 students in British Columbia in 1979. The same procedure as in (1) was followed in order to assess the extent of change between 1970 and 1979.

Significance of the Study

The conclusions reached in the NACOME report emphasized the need for better means of measuring change in achievement. In the literature viewed prior to this study, the application of the Rasch model to the problem had not been attempted. In theory, the Rasch model has a number of characteristics which make it particularly suitable for measuring change. By investigating a real situation, this study reveals some of the advantages and difficulties of applying the model.

In addition to contributing to the art of change analysis, the study should provide useful information on the state of mathematics achievement at the end of elementary school education in British Columbia. In helping to chart

real change in mathematical performance it should provide data for curricular emphasis or alteration. The findings of the study are a useful supplement to the British Columbia Mathematics Assessment (Robitaille & Sherrill, 1977).

CHAPTER II

REVIEW OF THE LITERATURE

The review of the literature is divided into two major sections. In the first section, the methodology and findings of previous attempts to measure change in mathematics achievement in the elementary grades are outlined. In the second section, the Rasch logistic model is described, and studies related to controversial issues surrounding the model are discussed.

STUDIES ON CHANGE IN MATHEMATICS ACHIEVEMENT

The paucity of information relating to changing student performance in Canada was noted by Hedges (1977):

> [The] widespread argument on the question of comparative student achievement is aggravated by the almost total absence of long-term evaluation studies based on standardized tests. (p. 3)

The few measurements of change in mathematics achievement have generally been carried out on three distinct levels: (1) school district, (2) provincial or state, and (3)

national. Each of these will be considered in turn. Because Canadian curricula and methods have always closely paralleled those in the United States, patterns of achievement among American school children have been used tc yield information on the Canadian situation. This fact is of particular importance at the national level, where nc Canadian studies exist.

Studies at the District Level

Hedges (1977) carried out a study of achievement in language arts and mathematics over a 40-year span. The study covered Grades 5 to 8 in the schools of St. Catharines, Ontario. Data were available frcm testing programs conducted in 1938 and 1952-54, each of which used the Dominion Arithmetic Test of Fundamental Operations, 1934 edition, for Grades 5, 6, and 7, and the Dominicn Group Achievement Test, Part II, 1934 edition, for Grade 8. The same tests were administered again in 1975-76. An attempt was made to adjust scores for changing socio-economic backgrounds and age-grade patterns of students. Furthermore, Hedges found that school objectives had changed sufficiently to require the creaticn of a "fair" test for 1976 which better reflected the arithmetic portion of the curriculum. By comparing the adjusted mean test scores, Hedges found that students in Grades 5 to 7 performed better than their earlier ccunterparts in fundamental operations in arithmetic when based on the test of fair comparison. On the other hand, he found that the 1975-76

Grade 8 students performed considerably less well than the other two groups in both arithmetic computation and arithmetic reasoning regardless of whether the original test or the fair test was used.

Hedges conceded that the results were anomalous. He argued that the most likely reason for the conflicting results across grade levels was the elimination of high school entrance examinations and the consequent decrease in emphasis on mathematics in Grade 8 over the previous twenty years in Ontario.

In a review of Hedges' research, Winne (1979) pointed out several flaws in the data analysis. Since local norms were not available for the St. Catharines schools in 1934, Hedges resorted to using median provincial norms, arguing only that the local and provincial student populations were similar. Secondly, Winne found it impossible to replicate Hedges' figures by using the described adjustment to the mean scores. Nevertheless, Winne concluded that the report was a healthy sign that the question of change was being examined objectively.

A shorter term study was conducted in 1974 by the North York (Ontario) Board of Education (Virgin & Darby, 1974). They wished to compare the mathematics achievement of students at that time with that of students in 1972. To this end, the Mathematics Computation subtest of the Metropolitan Achievement Test was administered to samples of approximately 1500 students in schools that had participated in the 1972 study at each of Grades 3, 5, and 6. Grade equivalent scores

based on 1971 American norms were used in the analysis. For Grade 6, the expected mean grade equivalent at the time of testing in 1972 was 6.5, while the achieved value was 7.2. In 1974, it was 7.4. The researchers concluded that the 1974 level compared favourably with that of 1972.

The 1974 North York study was replicated in 1975 in the same school district (Virgin & Rowan, 1975). A 20% stratified random sample was chosen, with the result that 1442 students in Grade 6 were selected to write the same test as in 1972. The mean grade equivalent score was 7.1 with an expected value of 6.7. The researchers argued that the decline from previous years could be explained partly by the more representative sample. They also suggested that the changed sampling technique was partially responsible for an increased range of school mean grade equivalents.

A study of Grade 3 pupils in the city of Edmonton showed a slight decline in arithmetic achievement from 1956 to 1977 (Clarke, Nyberg, & Worth, 1977). In this instance, all Grade 3 pupils wrote the California Achievement Test (Arithmetic, 1950 edition) in both 1956 and 1977. Twelve of the eighty items were deemed inappropriate because of high difficulty level or irrelevance to the Alberta curriculum, and both the 1956 and 1977 tests were rescored to eliminate those items. The mean raw score dropped from 55.98 to 55.56 in 1977. At the same time, the standard deviation increased from 5.97 to 7.58. The authors suggested that the decrease in means was due to lower achievement on a few items which had received less emphasis in recent years.

Hammons (1972) carried out a study in Caddo Parish, Louisiana to determine whether any significant change had occurred in arithmetic computation and reasoning among Grade 8 students during the period 1960 to 1969. A representative sample of 1000 to 1500 students was selected for each of the odd-numbered years from 1961 to 1969. The standardized test used was the California Achievement Test. Analysis of variance and trend analysis revealed a significant declining trend of proficiency in computational skill, but there was no significant change in achievement in reasoning.

Hungerman (1975) carried out a study to compare the computational skills of Grade 6 students in a southeastern Michigan school district in 1975 with those of a similar group in 1965. Ten schools were represented, with 305 students in 1965 and 386 in 1975. The test used in both years was the California Arithmetic Test (Part II--Fundamentals) which contained four sections, each of 20 questions: addition, subtraction, multiplication, and division. Fifteen separate analyses of covariance were performed, one for each section and one for the total computation score, using three different I.Q. covariates. Results differed slightly depending on the covariate selected. For reporting purposes here the "total I.Q." will be used.

Hungerman's results showed no significant differences in total computation scores. However, on the subtests, the 1965 group was significantly favoured ($p < 0.01$) on addition and subtraction. The 1975 group performed significantly better ($p < 0.05$) on division, while the groups

were not significantly different on multiplication. For individual items of computation, the 1975 group scored higher than the 1965 group on 10 addition items, 11 subtraction items, 13 multiplication items, and 16 division items. They also scored higher on all 33 whole number items, but lower on 20 of 30 fraction items, and on 5 of 8 decimal items. Hungerman suggested that a stable teaching staff and lack of any major socio-economic change were positive influences in maintaining computational skills.

In an attempt to extract more information from her data, Hungerman (1977) used profile analysis to determine relative performance within the categories of whole numbers and fractions. She also used median I.Q. scores to divide the subjects into high I.Q. and low I.Q. categories. Profile analysis generally confirmed the results of the analysis of covariance except for division in which no significant difference was found in performance across all items. In the analysis by content, the 1975 group performed better than the 1965 group ($p < 0.01$) on the whole number questions, while this was reversed on operations on fractions ($p < 0.01$). Performance of the low I.Q. subgroup in general changed little from 1965 to 1975; the high I.Q. subgroup contributed most of the total change.

Studies at the Provincial or State Level

The 1970 study in British Columbia cited in Chapter I (British Columbia Department of Education, 1970) examined the performance of students at the end of the elementary

school program. All Grade 7 students in the province were administered the Arithmetic Reasoning and Arithmetic Computation tests from the Stanford Achievement Test, Advanced Battery, Form L, (1953 revision) in 1964 and 1970. In 1964, the median modal-age grade equivalents in British Columbia were 1.8 and 1.1 years greater than the United States norms on the reasoning and computation tests, respectively. By 1970, the excesses over American modal-age grade equivalents had dropped to 0.8 and -0.1 on the two tests. The authors pointed out that the comparison in both cases was made on pre-1964 norms and suggested that new American norms in 1970 would be considerably lower.

Using a sample of three hundred papers stratified by total score, the 1970 British Columbia report cited changes in difficulty of items on the two Stanford tests. Particularly on the computation test, more items were more difficult in 1970 than were less difficult. However, as pointed out earlier, the conclusions were not based upon adequate sampling statistics.

In 1975, a study (Russell, Robinson, Wolfe, & Dimond) of the characteristics of elementary school mathematics programs in Ontario was released. Part of the intent of the study was to identify apparent trends in performance levels of students who had taken arithmetic tests as part of a continuing testing program. The researchers found only six counties in the province in which standardized tests had been administered over an extended number of years. In all cases some combination of obstacles made authentic

comparison of the results difficult. Nevertheless, the investigators cited performance in one of these jurisdictions as showing a slight decline in Grades 4, 5, and 6 from 1968 to 1974. They argued, citing available statistics, that such a decline could be explained by a decline in the mean age of students at each grade level across that period of time.

Russell et al. (1975) also asked teachers in 85 schools, selected on a stratified random basis, for their perceptions of trends in student performance. For Grades 6 and 8, approximately three-quarters of the sample felt that performance was either better or about the same in recent years. About one-half of the principals of the same schools, however, felt that performance had declined. On the basis of the questionnaire and the limited standardized test information available, the researchers argued that it was reasonable to conclude there had been no decline in standards on a provincial level in the period 1965 to 1975.

In Nova Scotia, a long-term study of competence in basic educational skills was carried out from 1955 to 1974 (McDonald, 1978). The Metropolitan Achievement Test battery was administered approximately every three years to provincial random samples of Grade 3 pupils. Comparisons were made on the basis of median grade equivalent scores. The results indicated no decline in performance. However, since three different editions of the tests were used--1947, 1959, and 1970--and the equating of test scores from one edition to the other was not discussed, it is difficult to tell whether an absolute level of performance was maintained, or whether the

pupils simply performed at the same level as the norming group in each instance.

Roderick (1973) used results on the Iowa Tests of Basic Skills to compare the performance of Grade 6 and Grade 8 students in the state of Iowa in 1973 to that of 1965, 1951-55, and 1936. Comparisons were made in the following topic areas: whole number computation; fractional number computation; decimals, percentages, and fractional parts; measurement and geometry; and problem-solving. Representative samples of schools in the state were selected. The particular statistical test used was not indicated. Roderick found the 1936 student performance superior to that of 1973 in all areas tested for each grade level. He found the 1951-55 students superior to the 1973 students in whole number computation and fractional number computation at Grade 6, and in decimals and percentages, and problem-solving at Grade 8. Students in 1965 in both grades were superior to the 1973 students in problem-solving, the only topic included in the 1965 testing. He concluded that the modern mathematics curriculum was seriously deficient with respect to many long-term curricular goals.

The results in a neighbouring state, however, do not reflect the same pattern. In 1950, Beckmann (1978) constructed a test based on the topics identified by the Commission on Post-War Plans of the National Council of Teachers of Mathematics as those which should have been mastered by a mathematically literate person. He administered his 109-item test to a sample of Grade 9 students across Nebraska in 1950, again in 1965, and finally, in 1975. The

test was administered to over a thousand students in each case, and the same schools were used insofar as possible each time. Beckmann found a significant gain ($p < 0.001$) between 1950 mean scores and 1965 mean scores. This was followed by a significant loss ($p < 0.001$) from 1965 to 1975. The net result left the 1975 students at about the same level as the 1950 students.

Scores in mathematics computation achieved by Grade 8 students in New Hampshire showed a consistent decline on grade equivalents from 1963 to 1967. In an investigation of whether the introduction of the modern mathematics program was having an effect on computational skills, Austin and Prevost (1972) classified Grade 8 students into three groups--traditional, transitional, and modern--depending upon the type of textbook used for teaching mathematics. Analysis of variance carried out in 1965 on raw scores on the Arithmetic Computation and Arithmetic Concepts subtests of the Metropolitan Achievement Test showed no difference among groups on the Concepts subtest. For the Computation subtest the modern group performed less well than the traditional group ($p < 0.01$), and also less well than the transitional group ($p < 0.05$). In 1967, students wrote the Arithmetic Computations, Concepts, and Applications subtests of the Stanford Achievement Test. Analysis of variance showed the modern group superior to the transitional group ($p < 0.01$) on the Applications subtest; the modern group superior to both the traditional and transitional groups ($p < 0.01$) on the Concepts subtest; and the modern group superior to the

transitional group ($p < 0.01$) on the Computations subtest.

In a follow-up study (Austin & Prevost, 1972) the 1965 eighth-grade group was tested in Grade 10 in 1967. Tests used were the Stanford High School Numerical Computation Test, and the Stanford High School Mathematics Test, Part A. The only significant difference showed the modern group superior to the traditional group ($p < 0.01$) on the Mathematics test. Using the two studies as evidence, the authors concluded that the type of mathematics text used did not differentially affect the ability of students to do arithmetic.

Studies at the National Level

In a novel approach to the problem, Maffei (1977) sent out 600 questionnaires to a stratified random sample of public high school mathematics chairpersons across all states of the United States. Each chairperson was asked to give the questionnaire to an experienced and effective mathematics teacher. The teachers were asked to state whether the mathematics achievement of students in their school was on the decline and, if so, to check the reasons for the decline. Seventy-nine percent of the teachers sampled believed there had been a decline. Most of the reasons cited for the decline centred on the student: less self-discipline, lower mathematical entry skills, lower reading comprehension skills, and higher absenteeism. The respondents also felt that mathematics teachers were less likely to set minimum academic pass standards.

A major study (NAEP, 1975) of the mathematical

skills of American children and adults was carried out in 1972-73 by the National Assessment of Educational Progress (NAEP). This organization was founded to survey the educational attainments of persons at ages 9, 13, 17, and 26 to 35 (adult) in ten learning areas, including mathematics. Over 90 000 individuals, statistically representative of the total population of the United States, were surveyed. The emphasis in the statistical analysis was at the item level, that is, on the percentage of respondents who correctly answered each item. The proportions of the most identifiable common errors for each age group for each exercise were also reported.

Although the NAEP data were intended to provide a baseline for future assessments in achievement, comparisons of the results were used in one instance to draw conclusions about the influence of the modern mathematics program. Carpenter, Coburn, Reys, and Wilson (1975) pointed out that the 13- and 17-year-old groups would have been taught throughout their school careers under the modern mathematics program, whereas the adult population would not. They argued that a detrimental influence of the new program on computational skills would be indicated if the younger groups performed less well than adults on computational questions. In fact, the 13-year-olds performed almost as well as adults on most computational tasks, and the 17-year-olds did better than the adults. Hence, they argued, no detrimental influence of the modern mathematics program was evident.

Finally, the general description of the trend in

mathematics achievement stated in the NACOME report (1975) provides a succinct summary of the studies to that time. The project team attempted to collect enough data to determine the truth or falsity of the charges of declining student competence in mathematics. They examined achievement data from four major sources: state assessment reports (particularly New York and California), performance on standardized test batteries and reports on norming samples from developers of standardized tests, research studies such as the National Longitudinal Study of Mathematical Abilities (NLSMA), and the National Assessment of Educational Progress. They came to two broad conclusions: (1) there has been a tendency for traditional classes to perform better on computation while modern classes do better in comprehension, and (2) mathematics achievement has shared in the general decline in basic scholastic skills since 1960. They noted, however, that the national picture was more complex than apologists or critics would make it out to be.

The studies cited are summarized in Table 2.1.

Investigator & Year	Time Span	Region	Grade Level	Content	Statistic Used	Findings
Hedges (1977)	1938/54/76	St. Catharines, Ontario	5 - 8	computations & reasoning	adjusted mean raw scores	Grades 5 - 7: 1976 > 1954 > 1938 Grade 8: 1952 > 1938 > 1976
Virgin et al. (1974,1975)	1972/74/75	North York, Ontario	6	computations	mean grade equivalents	1974 > 1972 > 1975
Clarke et al. (1977)	1956-1977	Edmonton, Alberta	3	mathematics achievement	mean raw scores	slight decline from 1956 to 1977
Hammons (1972)	1960-1969	Caddo Parish, Louisiana	8	computation & reasoning	ANOVA	1960 > 1969
Hungerman (1975,1977)	1965-1975	School district (Michigan)	6	computations	ANOVA & profile analysis	+,-: 1965 > 1975 ÷ , whole #'s: 1975 > 1965
B. C. Dept. of Ed. (1970)	1964-1970	British Columbia	7	reasoning & computation	median grade equivalents	1964 > 1970
Russell et al. (1975)	1965-1975	Ontario	6,8	mathematics achievement	mean raw scores & questionnaires	no change
McDonald (1978)	1955-1974	Nova Scotia	3	computation & concepts	median grade equivalents	no change
Roderick (1973)	1938/53/73	Iowa	6,8	basic arithmetic skills	???	1938 > 1973 (Grades 6,8) 1953 > 1973 (Gr. 6: whole #'s, frac.) 1953 > 1973 (Gr. 8: dec., prob-solv.)
Beckmann (1978)	1950/65/75	Nebraska	9	basic math. knowledge	t tests	1965 > 1950 1965 > 1975
Austin & Prevost (1972)	1963-1967	New Hampshire	8	computation, concepts, & applications	ANOVA	Mod > Trad (computations) Mod > Trad, Trans (concepts) Mod > Trans (applications)
Maffei (1977)	unspecified	U. S. A.	sec.	math. achieve.	questionnaire	79% of teachers believed decline
Carpenter et al. (1975)	1965-1973	U. S. A.	8,12	computation	logic based on item analysis	no detrimental effect of modern program on computational skills
NACOME (1975)	1960-1975	U. S. A.	all	mathematics	meta-analysis	overall decline in mathematical skills traditional > modern in computation modern > traditional in comprehension

THE RASCH LOGISTIC MODEL

In a concise exposition of the model, Rasch (1966a) indicated that there were just three assumptions underlying the model:

> (a) To each situation in which a subject (s=1,2...n) has to answer an item (i=1,2...m) there is a corresponding probability of a correct answer ($X_{si}=1$) which we shall write in the form
>
> $$Pr\{X_{si}=1\} = \frac{\lambda_{si}}{1+\lambda_{si}}, \quad (\lambda_{si} \geq 0)$$
>
> (b) The situation parameter λ_{si} is the product of two factors,
>
> $$\lambda_{si} = \pi_s \cdot \omega_i$$
>
> where π_s pertains to the subject and ω_i to the item.
>
> (c) Given the values of the parameters, all answers are stochastically independent. (p. 50)

The subject parameter, π_s, is a measure of the ability of the subject with respect to the kind of item being answered, and may take any non-negative value, with higher values indicating greater ability. The item parameter, ω_i, which may also be any non-negative value, is a measure of the easiness of the item. The model may also be set up using an item parameter, $1/\omega_i$, which measures the difficulty of the item, with higher values indicating greater difficulty. It is this form of the model which is described in Chapter I of the present study. In this form, by replacing π_s with A(s) and $1/\omega_i$ with D(i), the Rasch equation reduces to:

$$P(si) = \frac{A(s)}{A(s) + D(i)}$$

where $P(si)$ is the probability of subject $\underline{s}$ correctly solving item $\underline{i}$,

$A(s)$ is the ability of subject $\underline{s}$, and

$D(i)$ is the difficulty of item $\underline{i}$.

The underlying scale for $A(s)$ and $D(i)$ is a ratio scale ranging from 0 to $+\infty$.

A second approach is to use the equivalent logistic form of the model, derived as follows:

Dividing numerator and denominator of the right hand side by $D(i)$,

$$P(si) = \frac{A(s)/D(i)}{1 + A(s)/D(i)} \qquad [1]$$

The probability of failure on the item,

$$Q(si) = 1 - P(si)$$

$$= 1/[1 + A(s)/D(i)]$$

The expression for the odds on success, $O(si)$, becomes

$$O(si) = P(si):Q(si)$$

$$= A(s)/D(i) .$$

Taking the logarithm of both sides:

$$\ln[O(si)] = \ln[A(s)] - \ln[D(i)]$$

And, setting $\ln[A(s)] = a(s)$ and $\ln[D(i)] = d(i)$,

$$\ln[O(si)] = a(s) - d(i)$$

Thus, $O(si) = e^{a(s) - d(i)}$ or $A(s)/D(i)$

Substituting into [1]:

$$P(si) = \frac{e^{a(s) - d(i)}}{1 + e^{a(s) - d(i)}}$$

In this case, the underlying scale for a(s) and d(i) is an interval scale theoretically ranging from $-\infty$ to $+\infty$. The unit of measure for each parameter is the "logit" and, in practice, the usual range for each is approximately -4 to +4 logits, where negative values for a(s) indicate low ability persons, and negative values for d(i) indicate easy items. This approach and scale appear to be dominant in the literature at the present time.

A numerical example should help to clarify the model. Suppose person X of ability 1.30 logits is confronted with an item of difficulty -0.20 logits. The probability that the person will successfully complete the item is $P = e^{(1.30 - -0.20)}/(1 + e^{(1.30 - -0.20)}) = 0.82$. The odds on success for person X would be $e^{(1.30 - -0.20)} = e^{1.50} = 4.5$. If persons Y and Z with abilities 1.60 and 1.90 were to attempt the same item, their odds on success would be $e^{1.80} = 6.05$, and $e^{2.10} = 8.17$, respectively. Thus, the odds on success for Y compared to those for X are 6.05/4.50 = 1.35 times greater. The odds on success for Z compared to those for Y are 8.17/6.05 = 1.35 times greater. Hence, the equal intervals on the ability scale of 0.30 between X and Y and between Y and Z result in the same ratios for their odds on success. A similar procedure may be followed to determine the odds on success for one person faced with items of varying difficulty.

The Estimation of Parameters

Suppose that the responses to a test consisting of k items are arranged in a k X (k-1) matrix with items (1,2,...,k) in the columns, and raw scores (1,2,...,k-1) in the rows. The raw scores of 0 and k are excluded from the estimation procedure since they yield no information about the abilities of individuals achieving those scores; the test was simply too difficult or too easy for those individuals. Rasch (1966b) deduced from his model that the row and column totals are jointly sufficient statistics to estimate raw score (ability) and item difficulty parameters. Of more fundamental importance, he found that the row and column totals may be used as separate and independent estimators for ability and difficulty parameters, respectively. Hambleton and Cook (1977), citing work done by Andersen and by Wright and Douglas (1977a), maintain that the most attractive feature of the model is that total test score is a sufficient statistic for estimating ability.

Wright and Panchapakesan (1969) outlined two computer procedures for the estimation of parameters, the first using an unweighted least squares procedure, and the second using maximum likelihood. They recommended the latter since, according to them, it gives better estimates and the standard errors of estimate are better approximated. They set out an iterative procedure using raw scores and proportions of correct item responses as initial estimates. The method sets the obtained matrix of responses as an ideal and calculates the values of item and score parameters which best approximate

that matrix.

The computer program BICAL (Wright & Mead, 1978) is based on a modification of the Wright and Panchapakesan (1969) procedure. The recommended 1969 procedure failed to mention the necessity of removing from the calibration sample persons who answered all or none of the items correctly, and items correctly answered by all or none of the persons. In either case the parameter estimates will be infinite in extent and cannot be handled by the program. Secondly, the procedure as described results in biased estimates. The estimates of a person's ability are confined to the raw score equivalents. That is, unlike the nearly continuous difficulty range, where locations can be made ever more precise by increasing the sample size, the number of distinct positions on the ability scale is equal to the number of possible raw scores on a test, that is, k-1. Each such estimate represents a central measure of the true abilities of the persons achieving that score. Wright and Douglas (1977a) determined that a correction factor of $(k-1)/k$ can be applied to the biased estimates and that the results are extremely close to the correct values. This is the corrected procedure used in BICAL.

A clarification of the meaning of "ability" is perhaps in order at this point. For the Rasch analysis as outlined in the preceding paragraph, a person's ability is simply a transformation of the raw score obtained on a test. It is a measure of the knowledge and skills that an individual brings to a testing situation. It is not a reflection of an underlying capacity or potential for success. It does not

measure an innate quality. Virtually synonymous with "ability" in this sense is "achievement", or "attainment".

The Standard Error of Parameters

The Rasch procedure yields estimates not only of the difficulty and ability parameters, but also of the standard error of estimate associated with each. Classical procedures for dealing with item standard errors differ from those dealing with person standard errors. Each parameter will be examined separately.

Standard Error of Item Difficulties

In the traditional approach to test analysis, the estimate of the easiness of an item is given by the p-value of the item, that is, the proportion of correct responses to the item, say p. If the true difficulty of the item is P, the variance of the item is given by $P(1 - P)/n$, where n is the number of examinees. Thus the variance is greatest $(0.25/n)$ for test items on which half the respondents are successful. For items of greater or lesser difficulty, the variance is less, and decreases to zero for extreme values of p, that is, zero and one (Magnusson, 1966, chap. 2).

Larson, Martin, Searls, Sherman, Rogers, and Wright (1973) noted that the NAEP was considering transformations of scores to alleviate interpretive problems of change in percentages. If a sample from a population of respondents is used to estimate the p-value of an item, the standard error of the easiness estimate is given by $p(1 - p)/(n - 1)$. Hence

the standard error is a function of sample size and item easiness. A change of 10 percent will have different interpretations when it is a change from 25 to 35 percent as opposed to a change from 50 to 60 percent.

In the classical model the p-values for extremely easy or extremely difficult items are estimated with the least standard error. A change in p-value of a given magnitude thus will be deemed significant for items at either end of the difficulty continuum before such a conclusion is reached for items of average difficulty. This situation seems anomalous, particularly in the case of difficult items on multiple-choice achievement tests, where the element of guessing may be important. One might expect a larger variation due to chance for such items compared to items which correspond to the abilities of the test takers.

In the Rasch model the standard error of estimate for item difficulty is least for items whose difficulty matches the mean ability of the sample. That is, when the probability of passing the item is as close as possible to the probability of failing for the maximum number of persons in the sample, the most precise estimate of item difficulty is obtained (Whitely & Dawis, 1974). It should be noted that, while the model suggests that the difficulty parameter for an item is invariant with respect to the sample, the standard error of that parameter estimate depends upon the distribution of abilities in the sample. That is, the difficulty estimate varies in precision with the sample.

Standard Error of Person Abilities

In the Rasch model the standard error of ability is smallest for measures derived from central scores, becoming larger as scores become more extreme (Wright, 1977a). Thus the best estimate of a person's ability will be obtained from a test containing items for each of which the probability of success is 0.5. On very difficult questions the problem of guessing may confound the issue; on very easy questions the problem of boredom may arise. Ability parameters again are independent of the sample, but standard errors are sample-bound.

Wright (1977a) stated that the classical standard error of a score "has the preposterous characteristic of becoming zero at scores of zero of (sic) 100%" (p. 112). This is in contrast to Whitely and Dawis (1974), and Hambleton and Cook (1977) who suggested that the classical model provides a single standard error of measurement applicable to all examinees regardless of score. This discrepancy may be resolved by reference to Magnusson (1966, chap. 6). When classical parallel tests are considered, the standard error of measurement refers to the distribution of scores around a true score, and is constant for all persons. For randomly parallel tests, that is, tests comprising randomly selected items from an item population, the standard error of measurement refers to the distribution of an individual's test true scores around his or her population true score. In this case the standard error of measurement varies according to score.

Regardless of varying interpretations of the

classical standard error of measurement, by using the Rasch model Whitely and Dawis (1976) concluded that:

> Specifying measurement error for each score level, rather than the test as a whole, has an additional advantage; ability change at different score levels may be compared on a comparable statistical basis. A typical trait test is not equally precise for all populations. Extreme scoring populations can be expected to change more than mid-range scoring populations because of the greater measurement errors for populations at the tails of the total distribution. The standardized difference score (a z-ratio computed by dividing the differences in Rasch ability estimates by the measurement error associated with each score) permits comparison of change at different ability levels, since ability change is adjusted for the individual measurement errors. (p. 177)

The Evaluation of Fit

There are numerous ways of looking at the problem of whether the matrix determined by using the item and raw score parameters is a close enough match to the observed distribution of scores in the matrix. Historically, attempts at evaluation of fit seem to have been extreme, either casual or stringent.

Rasch (1960) used simple approximate methods (Rasch, 1966b) to determine graphically how well the model applied to his data. He grouped respondents according to raw scores into five ability groups ranging from low to high. He then, for example, plotted two-way graphs of item difficulty for each pair of adjacent score groups. Agreement with the model was shown visually by the fact that the four lines of best fit were parallel and had a slope near unity. The difference in intercepts is an indication of the differences in the mean

abilities of the group pairs.

Anderson, Kearney, and Everett (1968) used a procedure set out in a 1965 study by Brooks. The sample was divided into six ability groups yielding, for each group, an estimate of each item difficulty and the mean difficulty across items. For each item a *t* test was used to determine whether the regression of the six estimates on the six means departed significantly from unit slope. Two probability levels were used, 0.05 and 0.01, the former being the more stringent as it was the failure of fit that was being tested.

It should be noted that Brooks (cited in Tinsley, 1971) found, on his analysis of the responses of Grade 8 and Grade 10 students to the Lorge-Thorndike Intelligence Test, that of eight subtests, the two subtests which fit the Rasch model best were those dealing with number concepts. These are the subtests most closely related to the subject matter of the present study.

Wright and Panchapakesan (1969) proposed a different procedure for testing the fit of the model. For each cell in the *k* item by *m* non-empty raw score matrix, a standard deviate can be determined by subtracting the estimated expected value, P(si), from the observed value and dividing by the standard deviation of the observed value. Then the statistic for testing the overall fit is the chi-square statistic obtained by adding all the squared standard deviates across all cells, with degrees of freedom $(k-1)(m-1)$. For each item an approximate chi-square statistic can be formed by summing the squared standard deviates over the non-empty score groups,

with m-1 degrees of freedom. The authors cautioned against mechanically deleting all items for some significant value of chi-square since the test is approximate.

The foregoing procedure has become generally accepted for determining item fit. Various terms have been used to refer to it. For example, it is called "chi-square" by Kifer and Bramble (1974) and by Fryman (1976), each of whom rejected items having chi-square values whose probability of occurrence by chance was less than 0.05. When the chi-square value is divided by the appropriate number of degrees of freedom it is referred to as the "mean square fit", for example, by Forbes (1976) who eliminated items having a mean square fit in excess of 2.5 and item-total score correlation below 0.25.

The magnitude of the critical fit statistic cited by Forbes (1976) is an indication that different procedures are used by various authors to determine this statistic. It appears that Forbes' value was computed across groups of raw scores rather than for each raw score. It is likely that this statistic is the "between group fit mean square" given in BICAL (see Appendix C). The same criterion seems to have been used in part by Hashway (1977) who accepted only those items having an item fit mean square less than 2.20 and a discrimination parameter between 0.80 and 1.20.

The above-mentioned studies of Forbes (1976) and Hashway (1977) indicate that fit criteria other than the chi-square may be used. The point-biserial correlation may be invoked in either traditional or Rasch procedures. The

discrimination index is an indicator of how well the item separates the high ability persons from the low ability persons and is related to the traditional point-biserial coefficient. Wright and Mead (1978) found that for multiple runs using simulated data with exactly equal discriminations the standard deviations of observed discrimination parameters were frequently as large as 0.20. This evidence casts doubt upon the validity of Hashway's (1977) procedure. The issue of discrimination will be considered later in the chapter.

Rentz and Rentz (1978) suggested that the mean square fit statistic is the best single indicator of fit. They pointed out, however, that sample size affects mean square values. They recommended that the practitioner look for relative sizes instead of absolute numbers. They particularly cautioned against using probability values corresponding to the mean square noting that with sample sizes sufficient to estimate parameters these probabilities will usually be less than 0.05. They also made the point that the test developer may be faced with a decision to select-the-best or reject-the-worst. That is, if there is a large item pool available, one can be strict in selecting the best by using a combination of mean square and slope criteria. If, on the other hand, a high number of items must be selected from a limited number of items, a reject-the-worst attitude must be adopted.

The extreme case appears to be Choppin (1976), one of the first to support establishing item banks based on the Rasch model (e.g., Choppin, 1968), who stated:

> Past experience with the Rasch model convinces me that, for tests of typical homogeneity, the model fits well enough to be useful. Hence I now use it on any test that looks reasonably homogenous (sic) without too much concern about item fit. (p. 239)

The Need for Recalibration

A second problem emerges once criteria have been established on which to eliminate non-fitting items. The calibration of item difficulties is centred on the mean difficulty level. If the bulk of the non-fitting items tend to be more (less) difficult than the mean, the difficulty indices of the remaining items are higher (lower) than they would be if the items were recalibrated as a group. Ideally, the downward shift in difficulty for each of the remaining items should be a constant equal to the sum of the difficulties of the deleted items divided by the number of items remaining. Hence comparisons of the differences between item difficulties should yield the same results whether or not recalibration is carried out.

Problems more serious than a linear shift of scale may exist. One might expect that removal of non-fitting items would improve the estimates of parameters for the remaining items since the estimation process maximizes the fit for the entire collection of items. Theoretically, the remaining group of items needs to be recalibrated to improve the estimates; in practice, this necessity has been disputed by several researchers.

Anderson et al. (1968) used two separate criteria for rejecting items on two different tests. They found the

Pearson correlation between original item difficulties and the recalibrated values to be 0.9999 in all four cases. They concluded that recalibration was unnecessary.

Willmott and Fowles (1974) concluded that the difficulty indices of the fitting items were the same regardless of whether they appeared in the context of the entire set or by themselves. Their judgment was subjectively based upon visual inspection of graphs. Since they did not correct for the rescaling factor in their graphs, their conclusions are, at best, unconvincing.

Arrayed against these two studies are a number of others which advocate the recalibration of items. Brooks (cited in Tinsley, 1971), although not recalibrating his own data, suggested that estimates be recalculated after deleting non-fitting items. Tinsley (1971) recommended an iterative procedure for item calibration, especially when the sample consists of fewer than 500 subjects. Wright and Panchapakesan (1969) stated that the final step in item calibration is the reanalysis of retained items to obtain final difficulty estimates. Kifer and Bramble (1974) followed Wright and Panchapakesan's advice. Fryman (1976) carried the procedure to extremes by recalibrating four times.

On balance the evidence seems to indicate that recalibration should be carried out. Failure to do so will not adversely affect estimates of person abilities to any degree provided any linear scale shift is taken into account. However, if the difficulty of the item is the primary concern, then any improvement in the estimate is worth striving for.

Implications of the Model

The framework for analysis of the model set out by Rentz and Bashaw (1977) is clear and concise. Their approach is to consider the model in the form of an IF-THEN statement, that is, as a logical conditional. The assumptions stated by Rasch form the IF portion and Rasch's principle of specific objectivity forms the THEN portion. The clearest statement of the meaning of specific objectivity is contained in Rasch (1966b) where he uses the term to describe the situation in which:

> Comparisons of any two subjects can be carried out in such a way that no other parameters are involved than those of the two subjects ... [and] ... any two stimuli can be compared independently of all other parameters than those of the two stimuli. (pp. 105-106, italics in original)

Thus the model has the form:

$$\text{IF} \left\{ \begin{array}{l} 1.\ Pr\{X_{si} = 1\} = \dfrac{\lambda_{si}}{1+\lambda_{si}} \\ 2.\ \lambda_{si} = \pi_s \cdot \omega_i \\ 3.\ \text{stochastic independence} \end{array} \right\} \text{THEN} \left\{ \begin{array}{l} \text{specific} \\ \text{objectivity} \end{array} \right\}$$

There are certain implications which follow from each of the two portions of the statement, that is, from the antecedent or hypothesis, and from the consequent or conclusion. Those deductions which derive from the hypothesis will be called antecedent conditions and those from the conclusion consequent conditions.

Antecedent Conditions

Each of the three assumptions will be considered in order of increasing significance for achievement testing.

The first assumption simply serves to adopt a probability or stochastic model as opposed to a deterministic model as a convenient descriptive device. The choice does not imply that, in reality, the observed phenomenon occurred by chance as opposed to being causally determined (Rasch, 1966b). It should also be noted that the equation deals only with the probability of a correct answer, the implication being that responses are dichotomous or dichotomizable as correct/incorrect.

The third assumption, that of stochastic independence, assumes that the probability of obtaining a particular response pattern across a number of cells may be calculated as the product of the individual cell probabilities. For example, the probability that a person correctly answers a particular subset of questions on a test is determined by multiplying the probability of success on each of the items of the subset by the probability of failure on each of the remaining items. This assumption demands that the item responses given by one person do not affect the responses of any other person. It requires, further, that the response given by a person to one item does not affect the responses given to later items by that same person (Whitely & Dawis, 1974).

The inter-person independence suggests that under group testing conditions there is no copying or cheating. The

inter-item independence condition precludes the use of multi-part questions on which the answer to one section influences decisions to be made on another. It also implies that the test is a test of power rather than of speed. For a subject who fails to complete the test, the responses to the omitted items may be thought of as inter-dependent since they, as a group, relate to a particular external phenomenon.

Care must be taken to eliminate the influence of speededness on test responses. There is the usual problem of interpreting subject intentions on any omitted item: does the person not know the correct answer or is it simply a case of lack of time to give a response. Rasch (1966a) indicated that his model adequately described two subtests out of four analyzed. He subsequently discovered that in both cases of failure the subjects were under time stress. When subjects were stratified by overall working speed, the model applied.

The first and third assumptions generally set the stage by specifying the type of model (stochastic versus deterministic) to be used, and by setting out basic conditions for test construction and administration.

The second assumption is the one which leads to the greatest controversy and which has generated the largest amount of empirical research. Generally there are three topics of interest deriving from assumption number two: (1) the concept of <u>unidimensionality</u> of the ability being measured, (2) the influence of <u>guessing</u> on multiple-choice tests, and (3) the question of item <u>discrimination</u>. Each will be considered in turn.

The unidimensionality condition

Rasch does not appear to have used the term "unidimensionality". That term seems to have been imposed on the model when it was subsumed under the rubric "latent trait theory" by latent trait theorists (e.g., Birnbaum in Lord & Novick, 1968). It means simply that all the items constituting a test are homogenous in the sense that they measure only a single ability (Hambleton & Cook, 1977). The demonstrability of this condition is open to question.

Hambleton and Cook (1977) suggested that factor analysis of the test items might provide evidence for clustering items on one dimension. Wright (1978), however, maintained that factor analysis is not a valid indicator of dimensionality for the model since the stability of supposed factors over samples is weak. Furthermore, Rentz and Rentz (1978) pointed out that some of the test content in various studies reported appeared to have heterogenous and multi-dimensional content, yet the model with its implied unidimensional condition was applicable. They maintained that factor analysis is not satisfactory since it is itself a model with its own concepts of dimensionality. They also stated that there are no adequate preliminary tests for unidimensionality; the direct test is the fit of items to the model.

Rentz and Rentz (1978) stressed the necessity for using care in dealing with abilities which may become differentiated with progress through school. For example, mathematical ability in the lower grades may separate into

arithmetic competence and algebraic competence in later grades. If items from both subject areas are to be calibrated together it must be done on a calibration sample that has had the opportunity to learn both . The measurement must be done within a frame of reference which requires joint applicability of persons and items.

The interpretation of unidimensionality put forward by Whitely and Dawis (1974) is perhaps the most useful in visualizing the concept. Suppose that the matrix of probabilities for the cells of a test of $\underline{k}$ items appears as in Figure 2.1. Suppose also that the items have been ordered by increasing difficulty.

		Items				
		1	2	3	.....	k
Raw	1	P11	P12	P13	...	P1k
Scores	2	P21	P22	P23	...	P2k
	3	P31	P32	P33	...	P3k
	.	.	.	.	.	.
	.	.	.	.	.	.
	k-1	P(k-1)1	P(k-1)2	P(k-1)3	...	P(k-1)k

Figure 2.1. Raw scores by item matrix.

Unidimensionality requires that items are ordered in the same way within each score group, that is,

$$Pi1 > Pi2 > Pi3 > \ldots > Pik,$$

and that each item orders subjects by membership in a score group in the same way, that is,

$$P1j < P2j < P3j < \ldots < P(k-1)j .$$

The influence of guessing

The model assumes that the probability of a person with very low ability correctly answering an item of average difficulty is near zero. This condition may well apply to a test comprising open-ended questions only, but the situation becomes complicated when multiple-choice questions are used. Rasch (1960) recognized the problem but did not deal with it effectively. He analyzed two multiple-choice tests, arguing in one case that "it becomes possible to change the test form from multiple choice to free answers" (p. 62), and in the second case "with so many answers offered the deficiencies of a multiple-choice test are practically eliminated" (p. 62).

Other latent trait models attempt to estimate a separate "guessing" parameter for each item to account for item misfit at the low end of the ability continuum. Hambleton and Cook (1977) note that estimates of such a parameter generally are smaller than expected if the assumption is made that low ability examinees guess randomly on high difficulty items. They cite Lord (1974) in suggesting that this is probably due to the ability of item writers to develop attractive but incorrect alternative answers.

Before considering the problem of guessing further, it may be instructive to introduce the notion of the item characteristic curve (ICC). This is the function that relates the probability of success on an item to the ability measured by the test of which it is a part (Hambleton & Cook, 1977). In Figure 2.2, curve A is the item characteristic curve for an item of difficulty -1; B is the curve for difficulty 1.5. In each case the probability is 0.5 that a person with ability matched to the item difficulty will succeed on the item. In all cases, regardless of the ability of an individual, the probability of success is higher on item A than on item B since item B is more difficult.

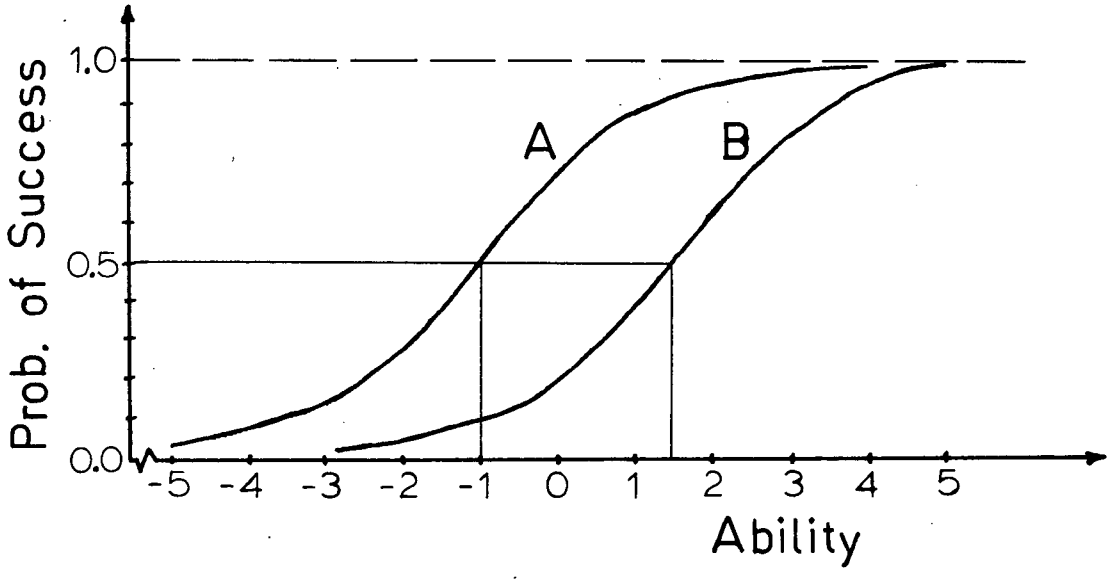


Figure 2.2. Item characteristic curves (ICC's).

In the Rasch model all ICC's have the same shape. A modification could be made if a guessing parameter were to be included in the model. Typical curves for five-alternative multiple-choice items might be as shown in Figure 2.3.

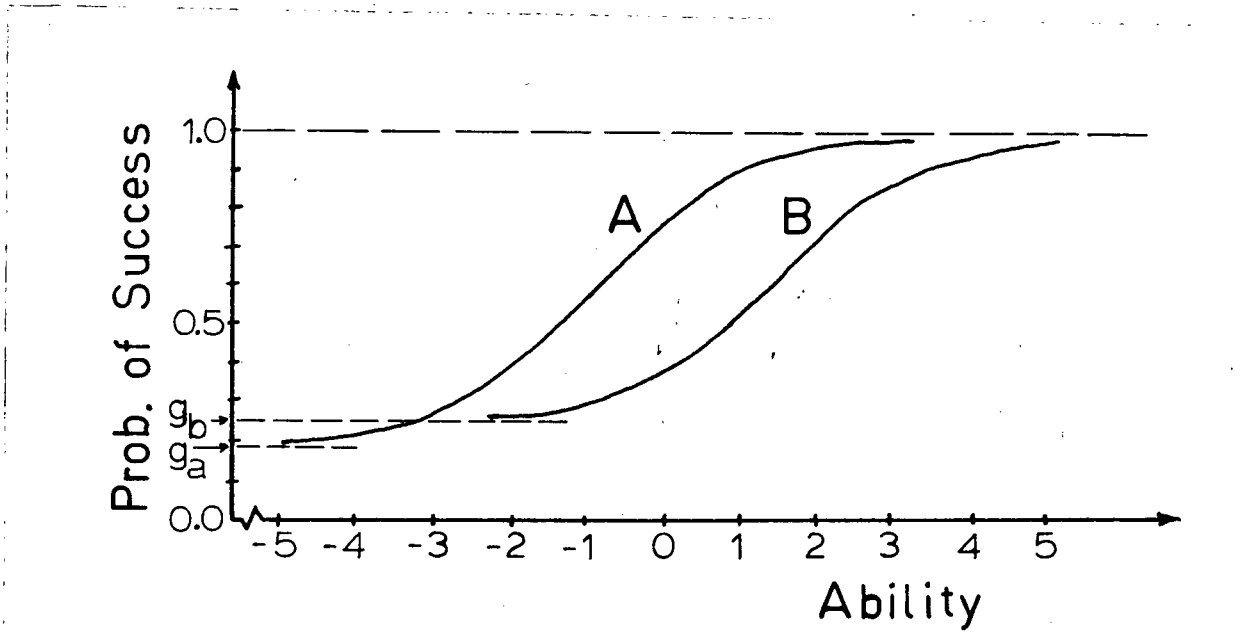


Figure 2.3. ICC's with a guessing parameter.

In each case the curve asymptotically approaches some hypothetical lower limit which is a function of the number of alternatives. Such a model might have the form:

$$P(si) = g + (1 - g)\frac{e^{a(s) - d(i)}}{1 + e^{a(s) - d(i)}}$$

where g is the asymptotic intercept on the probability axis for the item. The values of g may differ across items

depending on how much guessing each item provokes. Wright (1977a) made the observation that if one allows a parameter to measure the guessing potential on an item then a comparable person parameter representing a person's inclination to guess might equally well be admitted. He argued that such additional parameters "wreak havoc with the logic and practice of measurement" (p. 103).

Waller (1975) devised a procedure to remove the effects of random guessing by eliminating correct responses in the matrix for items deemed too difficult for particular ability examinees. The procedure is an iterative one requiring successive calibrations using increasing values of a cut-off value for the probability that an item is answered correctly by chance alone. Waller found that the overall chi-square value for testing the fit first decreased with increasing cut-off probability and then increased. The minimum chi-square pinpointed the required probability levels for best estimation of item difficulties. Simulated data verified the effectiveness of the proposed procedure. Notably, although Waller acknowledged his debt to Wright for advice, the procedure has not been incorporated into the latest version of BICAL (Wright & Mead, 1978).

Panchapakesan (1969), using simulated data, proposed a model which would provide for "intelligent" guessing. In her model the number of distractors eliminated by an examinee was a function of the probability that he or she would correctly answer the item. The probability levels at which the number of effective distractors changed was arbitrarily

set. Simulations of a 20-item multiple-choice (5 alternatives) test with a sample size of 1000 and varying ranges of abilities were carried out. She concluded that if the calibrating sample was able enough the effect of guessing would be negligible (p. 112). She cited Ross (1966) who also found guessing to be a negligible factor in his research, and she suggested that may have been so because the average ability of the subjects was greater than the average difficulty of the tests.

Panchapakesan proposed the following criterion for eliminating examinees from the calibration sample:

$$\text{If} \quad r = \frac{k}{m} + 2\sqrt{\frac{k(m-1)}{m^2}}$$

where $\underline{k}$ is the number of items,

$\underline{m}$ is the number of alternatives, and

$\underline{r}$ is the score below which examinees are eliminated.

In her equation, k/m is the expected score based on random guessing, and $k(m-1)/m^2$ is the variance of that score. Thus the procedure eliminates scores less than two standard deviations above the expected score due to random response by all examinees. It is difficult to reconcile this recommendation with her stated initial intention of allowing only for "intelligent" guessing. Nonetheless, the procedure appears to be reasonable, and provides a straight forward guideline for all multiple-choice tests.

Tinsley and Dawis (1975) acknowledged the possible effects of guessing on item calibration. They referred to Panchapakesan's criterion but decided, on the basis of their

initially small sample sizes (89 to 319, mode 269), not to follow the recommended procedure. Wright and Mead (1978) suggested that, in achievement testing, it is desirable to set the lower limit "somewhat above the guessing level" (p. 65).

The question of item discrimination

As previously outlined the Rasch model assumes all item characteristic curves have the same shape. This means, in terms of traditional item anlaysis, that all items have equal discrimination. Whitely and Dawis (1974) interpreted this to mean that the rate at which the probability of passing the items increases with total score must be equal for all items (p. 166).

Again, as for guessing, a modified model could be set out with item discrimination as a parameter. For example, if c is the discrimination parameter, then the model might be:

$$P(si) = \frac{e^{ca(s) - d(i)}}{1 + e^{ca(s) - d(i)}}$$

Typical ICC's for this function might be as shown in Figure 2.4.

In Figure 2.4 item A has the typical Rasch shape, and the value of c is unity. The value of c for item B is greater than unity, hence it better distinguishes higher ability examinees from those of lower ability. For item C the value of c is less than unity; it discriminates less well than item A. The measure of item discrimination is a function of

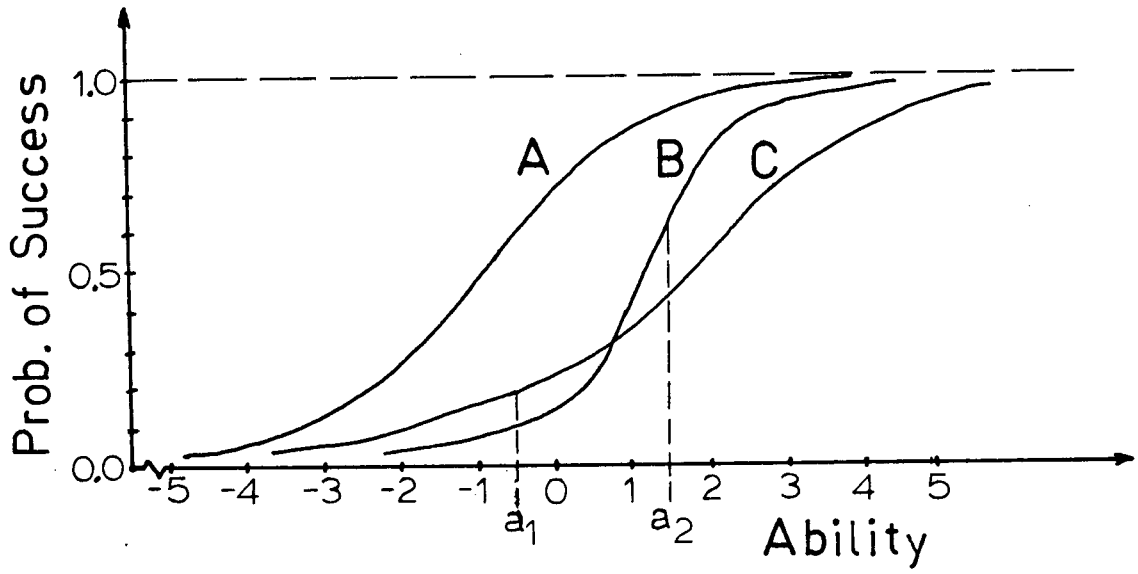


Figure 2.4. ICC's with a discrimination parameter.

the slope of the ICC at the point where the probability of success is 0.5.

Admitting a parameter for item discrimination complicates the model even more than allowing a guessing parameter. The Rasch model requires that any individual stands a better chance of succeeding on an easy item than on a difficult one. In Figure 2.4 this does not hold. Individuals with ability a(1) will likely do better on item C than on item B whereas for individuals with ability a(2) that situation is reversed.

Unfortunately for the Rasch model, items do differ in their discriminations. The lower limit for discrimination values of items on achievement tests is zero since negatively discriminating items are usually discarded, and the upper bound is likely around 2 (Hambleton & Cook, 1977). The

question then becomes one of the robustness of the model, that is, how much deviation the model can tolerate and still provide useful estimates of difficulty and ability.

Panchapakesan (1969) established certain criteria for the identification of items with deviant discriminations. Computer simulations indicated that her criteria could consistently eliminate seriously deviant items but they could not identify items whose discriminations ranged from 0.8 to 1.25. Furthermore, for a simulated test of 20 items whose discriminations ranged from 0.8 to 1.2 on a sample size of 500 the mean-square for overall fit was not significantly large. Panchapakesan also considered the question of the effect of varying item discriminations on the measurement of the abilities of the examinees. Her simulations showed, when item difficulty and discrimination were uncorrelated, that even for the extreme range of discrimination used ($0.4 < c < 2.5$) the bias in ability estimates was less than the standard error of measurement. She concluded, "In practical applications the model is robust even when the condition of equal discrimination is not met" (p. 100).

Dinero and Haertel (1976) also investigated the applicability of the Rasch model with varying item discriminations. Focussing upon the bias in measuring ability, they simulated 30-item tests on 75 subjects using five discrimination variances ranging from 0.05 to 0.25 drawn from three distributions: normal, uniform, and positively skewed (the most likely in practice). They concluded that the Rasch calibration procedure is robust with respect to

departures from homogeneity in item discrimination. They went on to discuss the seemingly counter-intuitive requirement of equal discrimination under the Rasch model as opposed to maximum discrimination in the classical model. They suggested that, although higher discrimination on an item results in better placement of an individual, this is achieved at the cost of loss of range. As an extreme example, an item having perfect discrimination yields information about only a single point on the ability scale. They also cautioned that estimates of discrimination depend on the fit of the item, that is, the worse the item fit the larger the error of estimate of its discrimination may be.

Several other studies using simulated data show similar results. Cartledge (1975) found that items with slopes in the region of 0.90 to 1.10 were treated as similar items fitting the model. She concluded that even when the slopes vary as much as from 0.80 to 1.20 the model is robust. The results of Hambleton (1969) accord with Cartledge's first finding, indicating that a range of 0.20 yields consistent fit to the model. However, when guessing was also introduced as a parameter, Hambleton found consistent rejection of the null hypothesis that the model fit the data.

On the other hand, as previously mentioned, Wright and Mead (1978) found that simulated runs using exactly equal discriminations frequently yielded observed standard deviations of the estimates as large as 0.20. The inference which might be drawn is that a range of 0.60 to 1.40 is acceptable.

In an application of the model tc actual results on objective achievement tests, including mathematics, Soriyan (1971) found that discrimination indices cf items fitting the mcdel were quite unequal, in the range of 0.50 to 1.25. It should be noted that this piece of research differs in kind from that on simulated data. In the latter case the discrimination parameters are known , in the former the values are those estimated after fitting the model.

Consequent Conditions

The preceding four sections were concerned with implications deriving from the assumptions, that is, antecedent conditions. They dealt with individual items and item fit to the model. Implications following from the consequence of the model, that is, of "specific objectivity", have been termed consequent conditions. They deal with sets of items and persons, that is, with tests and samples , and lead to questions of test fit. The two ccnsequent conditions in particular to be investigated are "sample-free item calibration" and "test-free person calibration".

Sample-free item calibration

One of the consequent conditions of the model is that estimates of the difficulties of items can be made without regard for the abilities of the persons in the calibrating sample. More correctly, difficulty estimates are made by taking into account the distributicn cf the abilities of the persons in the sample, thereby freeing the difficulty

estimates from the particulars of the abilities (Wright, 1967). Considerable research has been carried out to determine whether item calibrations made using real persons conform to the model. The studies which follow are cited chronologically in order to show the progressive nature of the research.

In 1964, Brocks (cited in Tinsley, 1971) analyzed the results of 509 Grade 8 students and 544 Grade 10 students on the Lorge-Thorndike Intelligence battery. The battery comprised five verbal and three non-verbal tests made up of multiple-choice items with five alternatives. Brooks plotted difficulty levels of all items on the test as determined from the Grade 8 sample against those difficulty levels determined from the Grade 10 sample. He devised his own statistic for determining the fit of the points to a straight line with unit slope and concluded that the difficulties were invariant with respect to the ability of the calibrating sample. Tinsley (1971), however, believed that it was not possible to judge the quality of Brooks' conclusion since the significance of Brooks' statistics could not be evaluated.

Anderson, Kearney, and Everett (1968) analyzed the responses to a 45-item intelligence type screening test for recruits to the Australian armed forces. Sample sizes were 608 and 874. The Pearson product-moment correlation between estimated item difficulties was 0.958, indicating difficulty invariance. When non-fitting items were removed and the analysis repeated, the correlation between the original difficulty level and the recalibrated values was 0.9999 for

all cases. The authors cited this as evidence that it is not necessary to recalibrate item difficulties after deleting non-fitting items. The sample of 608 was broken down into six ability groupings and item difficulties were determined for each group. The correlation between item difficulties by groups and by total size was 0.996. The authors concluded:

> Invariance in Rasch's model appears to be established, in that neither the sample from which the scale values were derived, nor the presence of items that failed to meet the model, appears to have any real influence on the resultant scale values. (p. 237)

Tinsley (1971) applied the model to analogy tests. He made ten comparisons between various types of respondents on four different types of analogy tests with sample sizes ranging from 89 to 630. Based on Pearson product-moment correlations he concluded that six of the ten comparisons ($r \geq 0.88$) supported the hypothesis of difficulty invariance. In two cases the sample sizes were deemed too small, and he suggested that the other two cases may have been invalid because of the test construction procedure. Tinsley also concluded that the deletion of non-fitting items increases the invariance of the item difficulty estimates.

Passmore (1974) applied the Rasch model to a large sample (6287) of nursing students who wrote the two-part National League for Nursing Achievement Test in Normal Nutrition. Two subtests of items fitting the model were identified and two non-overlapping samples were determined for each subtest by dividing the scores at the median subtest score. The item difficulty estimates correlated 0.994 and 0.997 for the two subtests, respectively. Passmore also noted

that the differences between the two difficulty estimates for each item were not more than 2.0 standard errors of estimate for the 40 items on one subtest and no more than 1.25 standard errors for the 65 items on the other. He concluded that the Rasch model was found to provide sample-free test calibration.

In a 1976 study to determine the smallest sample size which would yield reliable item difficulty calibrations, the Northwest Evaluation Association (NWEA) of Portland, Oregon used the responses of 1400 students to a Grade 4 mathematics test (Forster, Ingebo, & Wolmut, undated (b)). Five random samples were drawn for each of four different sample sizes: 50, 100, 200, and 300. The mean calibrated item difficulties for each sample size were correlated with those for the total group of 1400 students. The standard deviation of the item difficulty values for each sample size was divided by the standard deviation of the difficulty values determined by using the entire group of 1400 respondents. These ratios were compared to the value of unity, which should apply if the metrics are equal. A third statistic used in the comparison was the absolute value of the difference between the two estimates of difficulty values. Conclusions regarding sample size will be indicated later; here the main point of interest is the diversity of procedures used in the comparison.

The NWEA also carried out a study to determine whether item difficulties could be determined without random sampling (Forster, Ingebo, & Wolmut, undated (a)). The researchers divided the sample of 1400 students on a Grade 4 reading test in two ways: (1) above average versus below

average, and (2) inner-city students versus others. On the basis of what they called "restricted correlations" they concluded that random samples were not required. This conclusion was confirmed in a replication study on 4000 Grade 4 students in reading and mathematics, and 4000 Grade 8 students in reading and mathematics.

Hashway (1977) criticized procedures for testing the difficulty invariance condition as indicated in the research to that date on several counts. He stated that the use of simple bivariate plots was inadequate since no tests of significance are applicable. Of more importance, he contended that the correlation coefficient was not the appropriate statistic (p. 42). He argued that the existence of a high correlation or a similar rank ordering does not imply equivalence of raw scores. To overcome this problem Hashway suggested a different procedure for testing the item difficulty invariance property. He proposed looking at the regression equation between difficulty estimates:

$$d(ik) = b(1)d(ij) + b(2)$$

where $d(ij)$ and $d(ik)$ are the estimated item difficulties based on sample groups j and k,

$b(1)$ is the slope of the regression line, and

$b(2)$ is the intercept of the regression line,

both of the latter parameters estimated using a least squares procedure.

If the item difficulty estimates are equivalent then $b(1)$ should not differ significantly from 1.0 and $b(2)$ should not differ significantly from 0.0. The sufficient statistic

for testing each hypothesis is the t statistic. Hashway used this procedure to reanalyze data reported by Whitely and Dawis (1976) in which they, by applying an analysis of variance procedure, had found against the item difficulty invariance property. Hashway's analysis negated the conclusions set out by the original investigators.

Hashway (1977) applied the regression procedure to his own research. He constructed two mathematics tests and administered each to samples of Irish students at two levels, approximately Grades 6 and 7, and at two times, fall and spring. Thus for each test four calibrations were available with twelve different paired comparisons possible. At the 0.05 level of significance he found none of the slopes differing from unity and none of the intercepts differing from zero. Furthermore he found that the maximum observed difference of item difficulties for each item on the four calibrations to be less than 1.6 times the smallest standard error of the item's four difficulty estimates. This compared favourably with an expected 10% of the items which would have occurred on the basis of random error alone. He concluded that the item difficulty invariance property holds.

In the discussion of his results Hashway (1977, p. 148) expressed concern over the procedure. He suggested that the approach he used may be a necessary but not a sufficient condition for item difficulty invariance. He suggested an alternative procedure based on the analysis of standardized difference scores. This procedure parallels that to be elaborated in the next section of the present study.

Finally, Rentz and Rentz (1978) expressed the following caveat:

> One should regard "sample-free" and "item-free" with a little discretion. Items cannot just be given to any group of people; the sample must be comprised of appropriate people. We can not calibrate algebra problems on 2nd-graders; whether we calibrate them on 8th graders depends on the experiences of the 8th-graders. Items with which you intend to measure 9th-graders should be calibrated on people with 9th-grade experiences. While we do not have to pay particular attention to representativeness of the sample in any kind of strict sampling way, we have to exercise good judgement to make sure the sample is appropriate. There is usually no reason why one can't get a sample that is reasonably representative. If in general the group is appropriate in terms of the people for whom the test is designed, then the particular sample does not matter. (p. 26)

Test-free person calibration

The second consequent condition holds that just as estimates of item difficulty can be made without regard for the ability of the calibrating sample, so may estimates of person ability be made regardless of the difficulty of items used to assess that ability. This follows from the fundamental equation relating item difficulty and person ability; it may be thought of as a duality principle operating in the model.

Wright (1967) analyzed the scores of 976 students who had written the 48-item reading comprehension section of the Law School Admission Test. He divided the calibrated items from the test into two subtests of 24 items each. The 24 easiest items were used to make up an Easy Test, and the 24 most difficult items constituted the Hard Test. Each of the 976 students was thus assigned two ability estimates, one

based on the results on the Easy Test, and another based on the Hard Test. Each ability estimate was accompanied by its standard error of estimate. To assess whether equivalent ability estimates were made for each person, Wright determined the difference between the two estimates and divided by the standard error of the difference to obtain a standardized difference for each person. He argued that, if the estimates were statistically equivalent, the distribution of the standardized differences should have a mean of zero and a standard deviation of unity. The results showed a mean of 0.003 and standard deviation of 1.014. Without making further statistical analyses Wright concluded that test-free person measurement was indicated.

Willmott and Fowles (1974) used a similar procedure to analyze the responses to the 50 fitting items on a 70-item General Certificate of Education O-level Physics test. Two ability measures for each person were determined--one from the easiest 25 items, the other from the most difficult 25 items. Of the 745 candidates, the ability estimates for 703 (94.4%) differed by less than two standard errors. Since the significance level of this test was five percent, the authors concluded that, once the poor items had been edited out, the items in the test yielded person measures which were test-free.

Whitely and Dawis (1974) set up a test of the model similar to that of Wright. They pointed out that Wright's use of the term "statistically equivalent forms" falls under Lord and Novick's (1968) concept of tau-equivalent measures. The

expected values for true scores are equal but the expected values of error variances are not necessarily equal. To test this equivalency of calibrated subsets Whitely and Dawis re-analyzed a portion of Tinsley's (1971) data. Responses from 949 subjects on a 60-item verbal analogies test were used to calibrate the items. Three different divisions of the item pool were then set up: (1) odd versus even items, (2) easy versus hard items, and (3) randomly selected subsets with no overlap. They found for two subset comparisons, odd/even and random, no significant differences in either the means of the ability estimates or in their variances. For the easy/hard comparison, however, both the means and the variances differed ($p < 0.05$). When comparisons were made on standardized differences, the only significant difference was in the variance on the easy/hard subtest comparison ($p < 0.01$). Whitely and Dawis concluded that the results indicated that the Rasch model would produce statistically equivalent forms for any item subset except under the most extreme conditions.

Passmore's study (1974), on the other hand, showed contrary results. The responses of 6287 nurses on two Rasch calibrated subtests of an achievement test in nutrition were used to test the ability invariance hypothesis. The abilities of examinees scoring higher than the median score on each subtest were estimated from the easy, hard, odd, and even items selected from each subtest. Correlations between ability estimates on the easy/hard and odd/even items on each subtest were low, the highest being 0.382. Passmore concluded that item-free measurement was not attained to any reasonable

or practical degree.

Rentz and Bashaw (1977) felt sufficiently comfortable with the ability invariance condition to apply the Rasch model to the problem of test equating. Using the results of the equating phase of the Anchor Test Study (Ioret, Seder, Bianchini, & Vale, 1974) they converted the scores on twenty-eight reading tests for Grades 4, 5, and 6 students to a transformed Rasch scale. Each child responded to two reading tests yielding an estimate of the difference in the ability scale origin for the two tests. At each grade level, seven different batteries were used. Each battery was published in two forms and these were administered to fourteen additional samples at each grade level. This provided a basis for equating the tests within each grade level. The final step was to combine data across grade levels, that is, to carry out "vertical equating" of the tests, yielding a common ability scale for all tests and all grades. The authors found that in comparing their results with results using equipercentile equating, most test pairs were in agreement.

Rentz and Bashaw (1977) did not indicate any direct evidence for the item-free person measurement condition. That is, they did not attempt to determine the degree of agreement between estimates for each child's reading ability as determined from the two reading tests. Instead, they concentrated on demonstrating the stability of the raw score to estimated ability conversion for each test.

Slinde and Linn (1978) explored the adequacy of the Rasch model for vertical equating by using item response data

from 1365 students on a 50-item College Entrance Board achievement test in mathematics. Fourteen items were deleted as being too difficult, too easy, or possibly speeded. Wright's procedure was followed by dividing the items into an easy subtest and a difficult one. Their results were similar to those of Wright (1967) and Whitely and Dawis (1974). However, Slinde and Linn continued their analysis by calibrating the items separately using the low ability, medium ability, and high ability students independently. When ability estimates were obtained on extreme groups by using difficulty estimates obtained from the opposite extreme groups there were discrepancies. For the middle group the authors pointed out that the examinees would do better by taking the difficult test when ability estimates were obtained from the high group, but would do better to take the easy test when the estimates were obtained from the low group. They concluded that the Rasch model did not provide a satisfactory means of vertical equating, but conceded that the comparisons may have been more extreme than apt to be encountered when equating tests over several grades.

Hashway (1977) was concerned with the statistical procedures used to test the score invariance property. He extended the standardized difference procedure of Wright (1967) and Whitely and Dawis (1974). In that procedure, it will be recalled, if two estimates are available for the ability of each person they are subtracted and the difference is divided by the standard error of the difference of the two estimates. This results in a standardized difference score

for each person. If the distribution of such scores is unit normal, that is, with mean of zero and variance of one, it may be assumed that differences are due to random error. Thus the first step, Hashway argued, is to compare the observed distribution with the unit normal distribution function. This can be done using either the chi-square or the Kolmogorov-Smirnov statistic. If the test statistic is not significant the variation in standardized difference scores may be assumed to be a result of random error only. If, however, the hypothesis of unit normality is rejected, a second step should be performed.

Hashway argued that there were two reasons why the observed distribution would be non-normal:

(1) there is greater concordance of estimates than expected from random error, or

(2) there is greater discordance than expected.

Rejection of the first condition means acceptance of the second. For the first condition to hold, the distribution must be leptokurtic, and centred on zero with variance less than unity. Rejection of this leptokurtic property would imply that there is greater discordance than expected and the ability invariance condition does not hold. Hashway found that the distribution of standardized differences observed in his own research was leptokurtic and concluded that measurement based on Rasch instruments seemed to provide a stable mapping function.

The Issue of Sample Size

A controversy with respect to the size of sample required for the application of the Rasch model has recently surfaced. Whitely (1977) maintained that the key is the sample size required for adequately testing the fit of items to the model since the consequent conditions depend on this fit. She argued that a reasonably powerful significance test is needed to assess item fit and this can only occur with large sample sizes. She suggested that a sample size of less than 800 fails to detect sizeable differences.

Wright (1977b) argued that the important consideration is the precision of the calibration of items and concluded that sample sizes of 500 are more than adequate in practice. He contended that sample size depends upon the desired standard error of item calibration, and on the effects of item imprecision on the measurement of person abilities.

The previously cited study of the NWEA (Forster et al., undated(b)) on adequate sample size found that a sample size of 200 provided nearly as accurate information as a sample size of 300. As a consequence the NWEA now uses 200 to 300 students in field testing new items for the NWEA item bank.

CHAPTER III

DESIGN OF THE STUDY

The purpose of the study was to apply the Rasch model to measure change in arithmetic achievement at the end of Grade 7 in the province of British Columbia. The procedure for making comparisons was devised using the data available from previous testing programs in 1964 and 1970, and then applied to data obtained from a sample selected and tested in 1979. The essential element in the analysis of change was the difficulty of the test items as established using the computer program BICAL (Wright & Mead, 1978). The item difficulties were used, in turn, to establish summary statistics for blocks of items grouped so as to provide measures of performance on particular topics within the elementary mathematics curriculum.

Sampling Procedures

In March, 1964, the British Columbia Department of Education administered the Stanford Achievement Test, Advanced Battery, Partial, Form L, to all students in Grade 7. A

random sample of one hundred test booklets was selected at that time from each of the top, middle, and lower third of the distribution of total scores. In May, 1970, the Department administered the two arithmetic tests from the same battery to all Grade 7 students in British Columbia. A random sample of 300 papers was selected in a manner similar to that of 1964.

In 1979, it was clear that the representativeness of the 1964 and 1970 samples could not be replicated. The procedures of the previous years resulted in samples truly reflecting individual achievement throughout the province. The full authority and resources of the Department of Education were used in the earlier periods, while limited funding and reliance upon voluntary cooperation of persons in the field were characteristic of the data collection in 1979. Nevertheless, a procedure was established which, it was felt, resulted in a calibration sample sufficiently representative of the achievement of the population of Grade 7 students for purposes of comparison.

In order to produce item calibrations and estimates of abilities with standard errors comparable to those of previous years, a calibration sample size of 300, the same as in previous years, was opted for in the 1979 testing. In order to make the selection of subjects as similar to previous years as possible, it was necessary to select these 300 subjects from a larger sample using the same stratification as before. The size of the larger sample was dictated by the financial and clerical resources available to the researcher. This was set at 1500 students: just under four percent of the

approximately 40 000 students in the Grade 7 population.

Ideally, the larger sample would have consisted of 1500 students drawn randomly from the population. This was judged to be impractical, requiring the testing of a very small number of students in each of a large number of classrooms. There were two essential factors to consider in selecting a representative group of students. The first was representativeness of teaching practice; the second, representativeness of student ability. In the former case, the important factor was the school; in the latter, the student within the school. Hence it was decided to construct the sample in two stages: (1) by securing a representative sample of schools to yield results on approximately 1500 students, and (2) by selecting a stratified random sample of 300 students from the total sample.

The education system of the province consists of 75 school districts. Following contemporary Ministry of Education procedures for selecting samples of classes from these districts, two blocking factors were used: geographic location and size of school. The geographic regions and the number in the sample from each region are shown in Table 3.1.

Within each region, schools were ranked in order of their enrolment of Grade 7 students. The number of Grade 7 classes was estimated by dividing the school enrolment in Grade 7 by the average class size for the district in which the school was located. The average class size for the region was estimated by dividing the total number of students by the estimated number of classrooms in the region. Finally, the

Table 3.1

1979 Sample by Geographic Regions

Region	Number of Districts	Estimated # of Students	Approximate % of Total	Targetted # in Sample
1. South Centre	16	5 900	14.9	223
2. Greater Vancouver	9	14 700	37.1	556
3. South Mainland	11	4 500	11.4	171
4. South Coast	13	6 800	17.2	258
5. Southeast	12	2 400	6.1	91
6. North	14	5 300	13.4	201
Total	75	39 600		1 500

number of classes required for the sample was determined by dividing the targetted number of students for the region by the estimated average class size for the region. The result was rounded up to the next whole number.

As an example of the foregoing, for Region 1 the estimated number of classrooms was 285. The average class size for the region was 5900/285=20.7 . The targetted number of students in the sample was 223, requiring 223/20.7=10.8, or 11 classes.

For each region, the rank ordering of schools by enrolment was sectioned into strata equal in number to the required number of classes divided by two. Hence, in the example, five strata were defined, with two classes drawn from each stratum, except for the lowest from which three were selected. Each stratum contained 100%/5=20% of the students in the region. Since schools were ordered by enrolment, the top stratum contained fewer schools than the bottom. Within

each stratum the first class was randomly chosen and the second was located symmetrically within the stratum with respect to the first. This procedure ensured that the selection of classes from schools was well distributed both across the entire enrolment range and within each stratum.

The resulting sample comprised 65 classes in 61 schools located in 35 districts.

Data Collection

Educational policy and funding for each school district in the province is determined by a district school board, and the district superintendent, responsible to the board, is charged with administering the affairs of the district. In late March, 1979, a letter was sent to the superintendent of each district, requesting permission to contact the principals of the schools in the sample (see Appendix D). It was emphasized that the study was designed to investigate performance on a province-wide basis and that strict confidentiality with respect to students, schools, and districts would be maintained. Permission was granted by 30 of the 35 superintendents. In one case, the one school selected in the district had no Grade 7 students enrolled. In another, application forms to conduct research had to be filled out by the researcher, and permission was ultimately denied because of the short advance notification. In two other cases, the project conflicted with district-wide testing programs and consent was therefore withheld. In the fifth case, contact with the superintendent was not maintained due

to his attendance at a lengthy conference. As a result, the sample was reduced to 52 classes in 49 schools.

In mid-April, a test package (see Appendix D) was sent to the principal of each school. Each package contained a covering letter to the principal explaining the purpcse of the study and asking for his or her cooperatien, a letter to the teacher/test administrator asking that he or she administer the tests tc the Grade 7 students, a set of detailed directions for administering the tests, forty copies of each of the reasoning and computation tests for each class selected in the schcol, and a stamped, self-addressed envelope in which to return the completed tests. Principals were asked to have the tests administered in the pericd between April 23 and May 4. They were also asked to select randomly the required number of classes (cne or two) if there were more than this number in the school.

By mid-May responses had been received from all the schools in the sample. Forty-seven of the 49 principals ccoperated in the study. One principal felt that the tests would give rise to too much student anxiety because cf the number of items containing imperial units cf measure. The staff of the second school had already been involved in another doctoral research study, and felt that their students were being over-tested. The end result was the return of 1277 completed papers from 50 classrccms across the province.

The 1277 returned papers were marked by hand and ordered by total score. One hundred papers were selected at random from each of the top 426 papers, the middle 425 papers,

and the bottom 426 papers.

Verification of the Data

For the years 1964 and 1970 item responses from each of the 300 test booklets were coded by the author onto optical mark read (OMR) cards. Responses were coded 1,2,3,4,5 according to the response selected. The response was coded zero (0) if the item was left unanswered or if it was double marked. In some cases items were double marked but other notations on the booklet made it clear which response the student wished to have counted. In such cases the indicated response was accepted. The OMR cards were then read into a computer file.

In 1964 and 1970 the Department of Education had hand-prepared master summary sheets for the 300 tests in the form of an examinee by item matrix. Entries were 1 (one) if the correct answer was given and blank if not. In order to ensure comparability of the rescored items with the Department's scoring, a FORTRAN program was written to transform the data in the computer file into a similar examinee by item matrix of 1's and 0's. A comparison of this output with the Department's tally sheets, and referral to the original test booklets, served to verify the correct/incorrect matrix. Since the subsequent analysis used only this information, it was not considered necessary to verify the responses to items incorrectly answered.

The 300 papers selected in the 1979 sample were individually examined and corrected for anomalies in the

selection of responses. The responses, alphabetical as on the test papers, were commercially keypunched directly from the test papers onto cards. The keypunch operator verified the entries by keypunching the entire set of booklets twice, thereby ensuring agreement. A further check was made on the accuracy of the keypunching by randomly selecting a 10% sample for comparison with the original test papers. A small computer program was written to transform the alphabetically coded responses into the numerical values of 0,1,2,3,4,5 as for previous years.

BICAL

Estimates of item difficulties and person abilities were obtained using the computer program BICAL (Wright & Mead, 1978). A description of the components of the program and an annotated example of output are contained in Appendix C.

The algorithm used by BICAL is the unconditional maximum likelihood procedure (UCON) and consists of the following steps:

(1) Determine the number of correct responses for each item, s(i), and the number of persons, n(r), at each score, r.

(2) Edit the data to remove items on which zero or perfect scores were achieved, that is, for which s(i) equals zero or N*, the number of persons in the sample. Edit the data to remove persons who achieved zero or perfect scores, that is, for whom r equals zero or K*, the number of items on the test. Let N and K be the number of persons and items remaining, respectively.

(3) For each raw score, r, assume a corresponding initial ability estimate, $a(r)^0$, such that $a(r)^0 = \ln[r/(K-r)]$.

(4) For each item, i, assume a corresponding initial difficulty estimate, $d(i)^0$, such that $d(i)^0 = \ln[(N-s(i))/s(i)]$.

(5) Centre the set of item difficulties at zero by subtracting the mean of the K item difficulties from each item difficulty.

(6) Through iteration, determine a revised set of item difficulties, d(i), by using Newton's method to solve each of N maximum likelihood equations.

(7) Through iteration, and using the revised set of item difficulties, determine a revised set of person abilities, a(r), by using Newton's method to solve each of K maximum likelihood equations.

(8) Repeat steps 5, 6, and 7 until stable values for item difficulties, d(i), are obtained.

(9) Correct for bias by multiplying each item difficulty, d(i), by (K-1)/K.

(10) Determine person abilities, a(r), for each raw score, r, using the unbiased item difficulties, d(i), determined in step 9.

(11) Correct for bias by multiplying each person ability, a(r), by (K-2)/(K-1).

(12) Determine the asymptotic standard error for each d(i) and a(r) from the second derivative of the log likelihood function.

Several fit statistics, and an index of discrimination, are determined for each item. The

interpretation of each of these is illustrated in the sample output in Appendix C.

It will be recalled that the Rasch model does not include a "guessing" parameter for multiple-choice questions. Various suggestions have been made to allow for this fact when calibrating the items. Waller (1975) suggested removing responses for items too difficult for examinees. Wright and Mead (1978) suggested accepting only scores of examinees somewhat above the guessing level. The suggestions for eliminating examinees are more practical than those for eliminating particular responses. Consequently, Panchapakesan's recommendation outlined in Chapter II was used to eliminate examinees from the item calibration procedure. That procedure establishes the score $\underline{r}$ below which examinees are eliminated, such that

$$r = \frac{k}{m} + 2\sqrt{\frac{k(m-1)}{m^2}}$$

where $\underline{k}$ is the number of items, and

$\underline{m}$ is the number of alternatives per item.

As an example, if all 89 items were to be calibrated, the use of this criterion would lead to the following decisions:

(1) Reasoning test: For items 1 to 30 there are five alternatives, yielding $\underline{r}$ = 10.38. For items 31 to 45 there are four alternatives, yielding $\underline{r}$ = 7.10. Total $\underline{r}$ for the test is 17.48. Therefore eliminate scores less than 18.

(2) Computation test: For all items there are five alternatives, yielding $\underline{r}$ = 14.11. Therefore eliminate scores

less than 15.

(3) Total test: The sum of the r's is 31.59. Therefore eliminate scores less than 32.

Editing the Data

The Deletion of Persons

For reliable estimation of a student's mathematical ability it is necessary that the student have sufficient time to respond to all items on a test. That is, the test should be one of power rather than of speed. Speed is a factor which both complicates the model and confounds the interpretation of the results.

The authors of the Stanford Achievement Tests state that the time limits are generous and practically all pupils should have sufficient time to attempt all the questions (Kelley et al., 1953, p. 2). Nevertheless, in a preliminary inspection of the 1964 and 1970 data, blocks of unanswered items indicated that a number of individuals probably did not have time to finish. It was necessary to establish a criterion for deleting persons who, it was suspected, simply did not have time to complete the tests.

For the measurement of arithmetic skills there were three timed sections: two on the reasoning test, and the computation test itself. It was assumed that students answered the questions in the order in which the questions were presented on the tests, and that only those items at the end of a timed portion might have been left blank because the

student did not have time tc finish. It was also ncted that the sample size of 300 was near the lower limit for effective calibration, and that the use of a cutoff score for guessing would further reduce the sample sizes. Hence, it was necessary to balance the undesirability of loss of subjects on which to calibrate items against the inclusion of inappropriate subjects. To this end, the following decision rule was decided upon: if, at the end of any of the three timed portions the subject omitted at least ten items in a row, that person was deleted frcm the data base. As a result, in the most severe case, the 1970 sample was reduced in size by four percent.

The Deletion of Items

The deletion of persons is made before the calibration process; the deletion of items is made after. Decisions concerning the deletion of items therefore are aided by the use of statistical data which indicate how well the items fit the model. In principle, such decisions should be easier to defend than the rather arbitrary decision on the deletion of persons. In practice, this is not the case.

Criteria to evaluate item fit used by various researchers have been identified earlier. The most frequently cited criteria were the chi-square or mean square fit, residual discrimination values, and point-biserial ccrrelations. No commonly accepted combination of criteria or significance levels was identified in the literature. The present study fell into the Rentz and Rentz (1978) category of

"reject-the-worst" because only a limited number of items were available and maximum information was to be sought from the analysis. Therefore, a reasonably permissive criterion was set for including items in the calibration process. Items were deemed to be non-fitting if:

(1) the mean square fit exceeded unity by four or more standard errors, and,

(2) the discrimination index was less than 0.70.

This criterion was established basically because of the practical demands of the study. On one hand it was desirable to eliminate items which clearly failed to fit the model; on the other hand, the most important aspect of the analysis lay in the comparison of groups of items. It was felt that, in a group, the presence of one or two items which fit not as well as theoretically desirable would not adversely affect the comparisons. Preliminary analysis indicated that the criterion would eliminate about ten percent of the items on each test, and this was judged to be a satisfactory resolution of the problem.

Testing the Unidimensionality of the Two Tests

It was argued in Chapter II that there was a need to regroup the items on the reasoning test and the computation test. This could be done meaningfully only if the ability underlying each test was the same. There is some evidence from other sources that this is indeed the case. Merrifield and Hummel-Rossi (1976) subjected the Stanford Achievement Test: High School Basic Battery, 1965 edition, to factor

analysis using the responses of a sample of 226 Grade 8 students on the nine tests of the battery. The three mathematics tests were found to lie in a compact cluster on one of two oblique factors. The authors suggested that the analysis indicated redundant information in the data.

Two procedures were used to test the unidimensionality of the tests. The first is referred to as the simple test and the second as the strict test.

Rentz and Rentz (1978) suggest that there are no separate adequate tests for unidimensionality. They argue that the test of fit to the model will tell whether the antecedent conditions of the model have been met. They suggest that one might still get good fit on a set of mathematics items even though some appear to measure algebra and others to measure arithmetic.

Wright and Panchapakesan (1969) state that the failure of any item to fit the model may be for two reasons: (1) the model is too simple, or (2) the item measures a different ability than the fitting items. Thus fit to the model is evidence that a set of items measures a unidimensional ability.

Based on these arguments, the simple test was to calibrate the entire set of 89 items as a single group for each of the three administrations. It was reasoned that the unidimensionality condition could be assumed if the items acted as a cohesive unit in all three instances with no clear separation of non-fitting items along subtest lines.

In addition to this test a much more stringent

statistical test was used. The basic question was: given two tests A and B, how could one empirically decide whether they measure the same underlying ability? A strong indication of unidimensionality would be given if, after administering the two tests to a single sample of persons, the rank ordering of persons by raw scores was identical on each test. One would not expect equal person scores on the tests because of possible differences in the mean difficulty of items constituting the two tests. But, the Rasch model provides a measure for equating tests under exactly this condition, that is, it carries with it the property of test-free person calibration. If each person taking the tests is assigned the same Rasch ability score by each test, the condition of unidimensionality may be assumed.

Each test was calibrated independently, yielding two estimates of each person's ability. Each ability estimate derived from the computation test was then adjusted by the difference between the mean sample abilities on the two tests in order to make the ability scales comparable (Panchapakesan, 1969, p. 168).

Some elaboration of this procedure may be in order. Since each test was calibrated independently, and was centred at the mean item difficulty for its own collection of items, it was not expected that the mean ability level on each test would be the same. For example, a difficult set of items might show a mean ability of -0.3, whereas an independently calibrated set of easy items might show the mean ability of the same group of subjects to be 0.2. A linear shift of 0.5

units applied to each subject on one of the tests would bring the ability estimates onto the same scale.

Once two comparable ability estimates were obtained, the procedure suggested by Hashway (1977) for evaluating the score invariance property was invoked. A standardized difference score for each examinee was determined and the distribution of such scores across the sample was compared to the unit normal distribution. The test for unit normality was the Kolmogorov-Smirnov statistic. In view of the fact that the described procedure really tested two conditions at once--unidimensionality and score invariance--the probability for rejection of the null hypothesis was set at 0.01. If the hypothesis of unit normality were rejected, the distribution was to be evaluated for shape. If it proved to be leptokurtic with variance less than unity, the condition of strict unidimensionality was to be assumed.

Summarizing the steps, for each of 1964, 1970, and 1979, the following procedure was carried out:

(1) BICAL was run on the 45 items of the reasoning test, with a minimum acceptable score of 18.

(2) BICAL was run on the 44 items of the computation test, with a minimum acceptable score of 15.

(3) The raw score for each person on each of the two tests was converted to a Rasch ability score using the conversion table in the BICAL output.

(4) Each person's computational ability score was incremented by an amount determined by subtracting the sample mean computation ability from the sample mean reasoning

ability. This procedure adjusted the abilities to a common ability scale.

(5) A standardized difference score was determined for each person using the two ability scores and the standard error for each as indicated in the BICAL output. The calculation was as follows:

$$D = [a(R) - a(C)] / \sqrt{se(R)^2 + se(C)^2}$$

where $a(R)$ = the examinee's reasoning ability,

$a(C)$ = the examinee's adjusted computation ability,

$se(R)$ = the standard error associated with $a(R)$,

$se(C)$ = the standard error associated with $a(C)$.

(6) The distribution of standardized difference scores was tested for unit normality using the Kolmogorov-Smirnov statistic.

Testing the Changes in Item Difficulty

Once the unidimensionality assumption was justified, the item difficulty values used were those determined by calibrating the entire collection of items as a single test. The fit criteria cited in a previous section were applied to each of the 1964, 1970, and 1979 calibrations. Items which failed to meet these criteria on at least two of the three administrations were deleted from the analysis. The data for each year were recalibrated after removing the non-fitting items. The process was repeated until no non-fitting items were common to two of the three samples. Item difficulties

used in the analysis were those calibrated on the final run.

Comparisons of item difficulties were based on the Rasch difficulty parameters, having the logit as the unit of measure for both item difficulties and person abilities. The difference between traditional comparisons and Rasch item difficulty comparisons can be clarified by referring to the top part of Figure 3.1, in which it is assumed that 89 items remain in the calibration. They are shown from left to right in order of increasing difficulty on a hypothetical 1964 administration. Suppose that the increase in percentage difficulty, that is, in the proportion of people incorrectly answering the item, is linear across the sample in 1964. The mean difficulty is 30%. Now suppose that a hypothetical 1970 sample is uniformly worse in terms of raw scores across all items. The mean difficulty is now 50%. Hence, the difference in mean difficulty is 20%.

When these data are analyzed by the Rasch model the distribution of item difficulties in each case is represented by the same curve, shown in the middle part of Figure 3.1, as each independent calibration is centred at the mean difficulty level, with a value of zero. The difference in mean difficulty will show up in the Rasch analysis as a difference in the mean abilities of the two groups since the two parameters are determined on a common scale. The mean ability of the 1964 sample would be about -0.9 logits, and that of the 1970 sample, 0.0 logits.

To see, graphically, how the item difficulties differ, the superimposed calibration curves can be separated

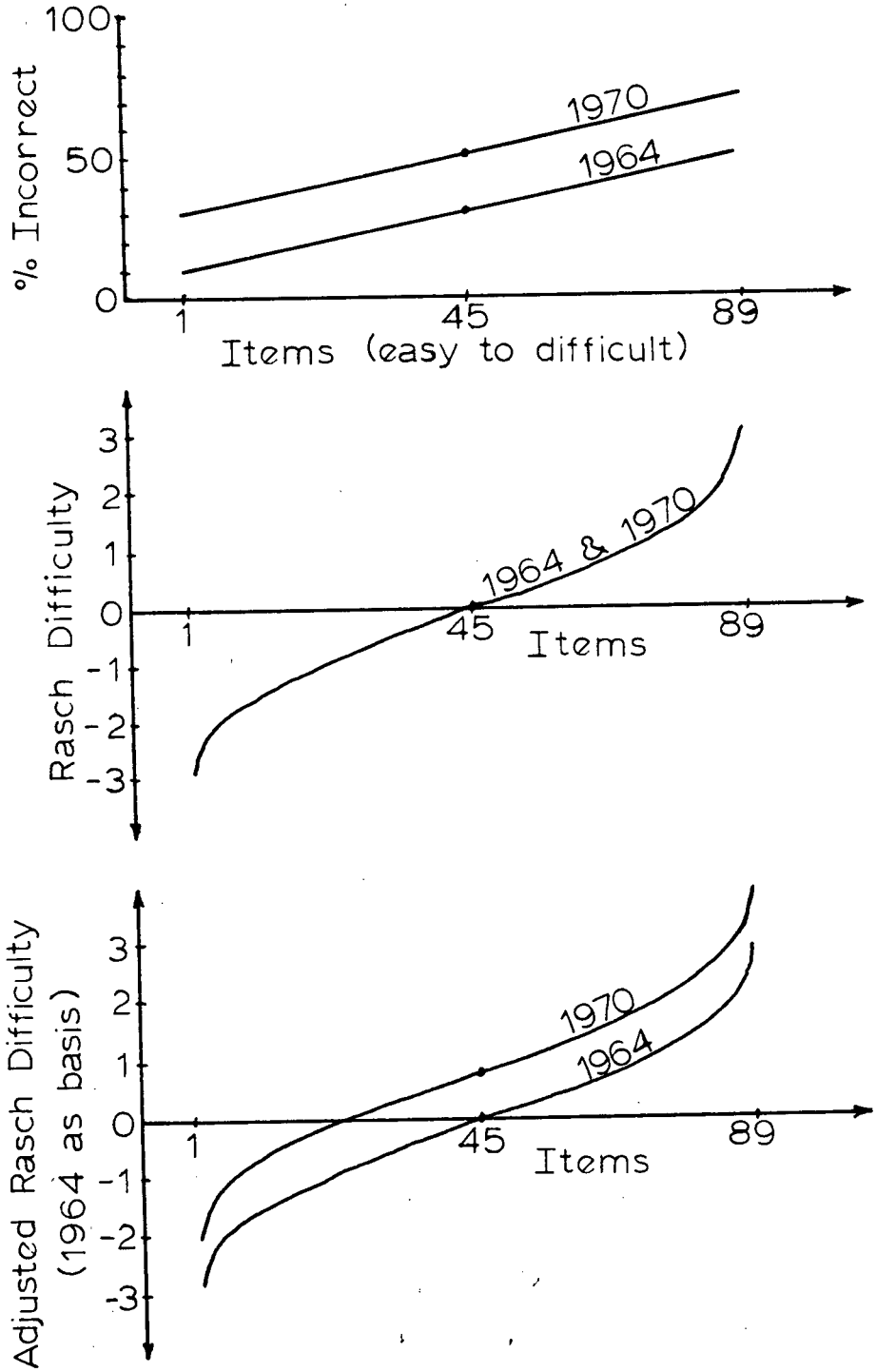


Figure 3.1. Two hypothetical distributions of traditional and Rasch item difficulties.

and shifted to the locations indicated by the mean abilities. The separated curves would be as shown in the lower part of Figure 3.1.

Once independent estimates have been established for two administrations, two types of comparisons can be made. In the first case, by the item difficulty invariance property of the model, it is expected, within the limits of random error, that the two difficulty estimates for an item will be equal. If this expectation is not met it can be concluded that changes in item difficulty have occurred relative to each other.

Suppose in Figure 3.2 that the 1964 testing situation is the same as in Figure 3.1. Suppose again that the mean 1970 percent difficulty is 50%. This time, however, because of changed curriculum emphasis or teaching practice, increases in item difficulties are not uniform, and the irregular line represents the graph of the 1970 item difficulties. This irregularity will be reflected in the Rasch item difficulties, as shown by the irregular curve superimposed on the regular 1964 curve. Thus, comparison of item difficulty based on the unadjusted Rasch estimates yield information on change in difficulty within the set of items, relative to the mean difficulty level of the item group. The vertical segments show the magnitudes being tested.

The second type of comparison which can be carried out is that of absolute difficulty across time. In each year the sample represented the overall distribution of both curriculum coverage and mathematical ability across the

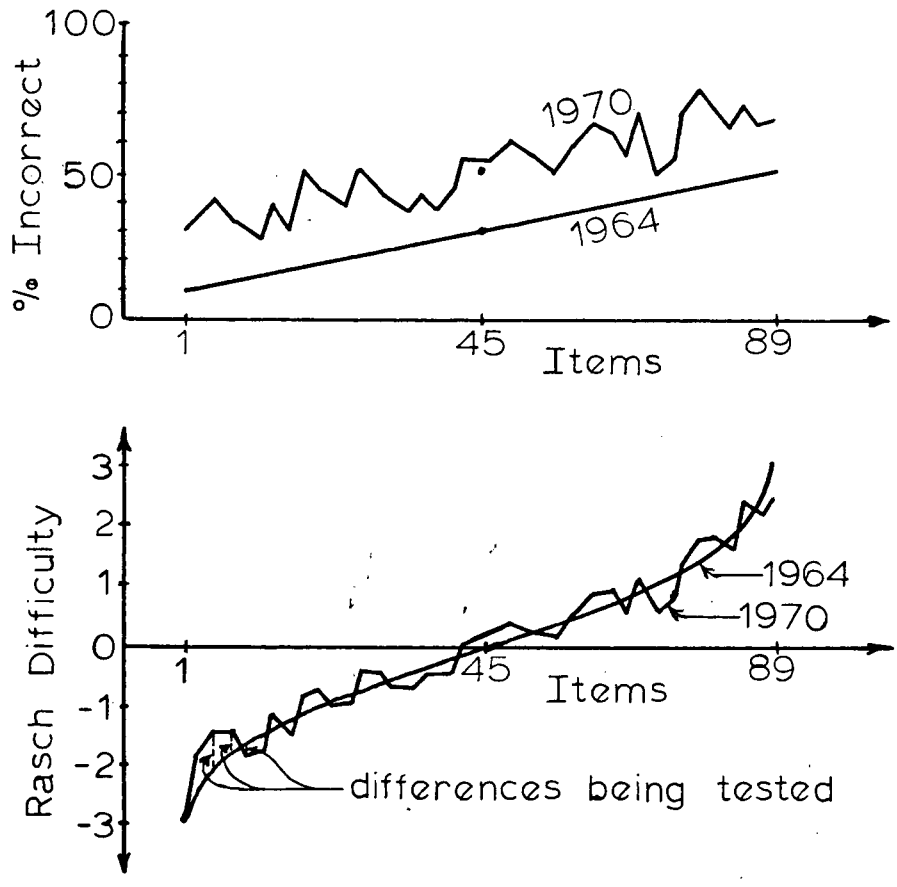


Figure 3.2. The testing of change in the relative difficulty of items.

province. Representativeness of curriculum coverage is the key to the first comparison of item difficulty within the aggregrate of items. Representativeness of mathematical ability is the key to the second comparison. In the second case the difference in mean ability of the 1964 and 1970 samples provides an estimate of the difference in the origin of the difficulty scales. By adjusting each item difficulty on one of the administrations by this amount, the item difficulties are placed on a common scale. Again, it is expected, within the limits of random error, that the difficulty estimates for each item will be equal.

As before, the procedure may be illustrated

graphically. Referring to Figure 3.3, suppose that the

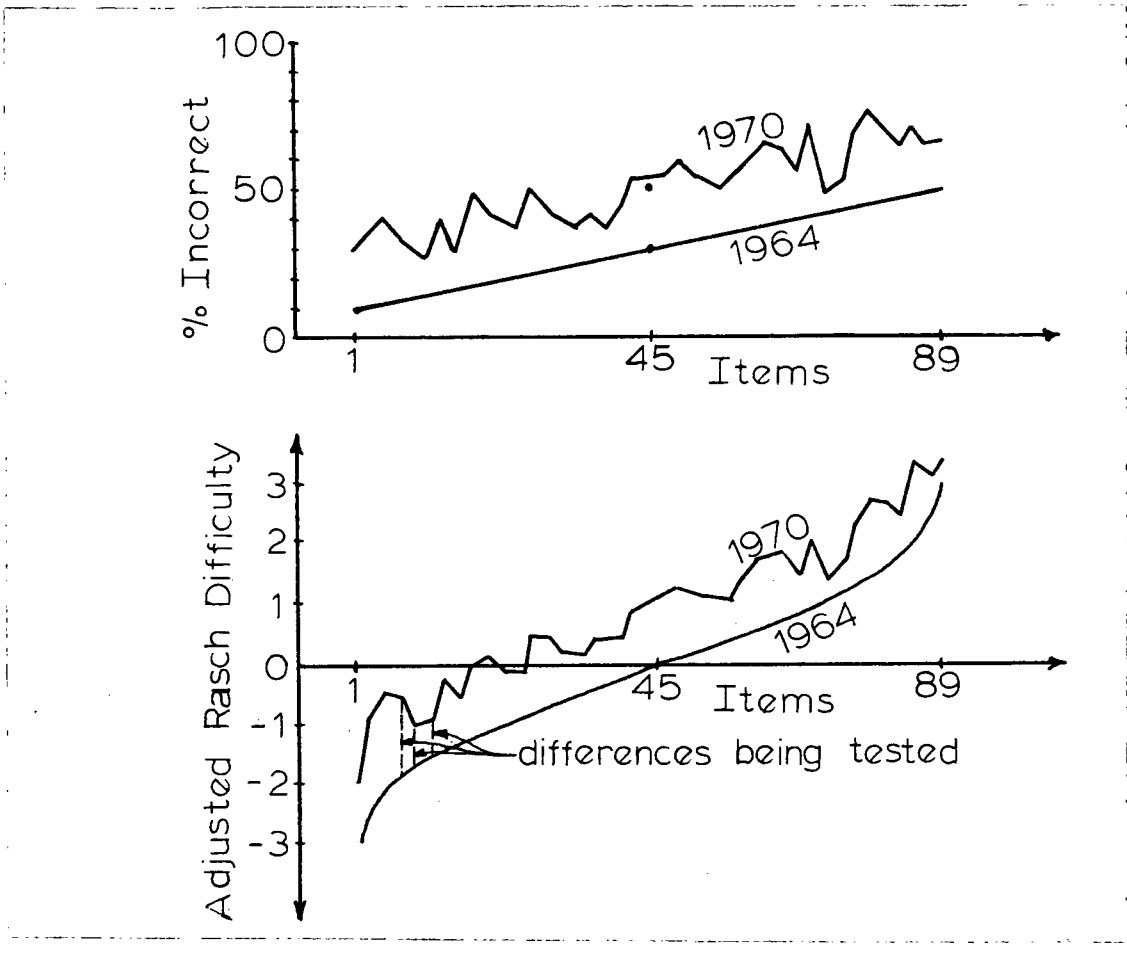


Figure 3.3. The testing of change in the absolute difficulty of items.

traditional difficulty distributions for the two administrations are the same as in Figure 3.2. The adjustment procedure described for Figure 3.1 is now used to separate the Rasch calibrations by the amount of the difference in the mean abilities. In the lower portion of the figure, it is again the magnitude of the vertical segments which is being tested for significance, but these segments now represent absolute

change because of the shifted location of one of the curves.

For each type of comparison Hashway's (1977) procedure constituted an omnibus test of the equality of item difficulties. This procedure paralleled that outlined in the previous section on testing score invariance. The Kolmogorov-Smirnov statistic was used to test the assumption of unit normality of the standardized difference scores of item difficulty. In each analysis rejection of the hypothesis of equal item difficulties permitted analysis of change for individual items.

For the assessment of relative change, the standardized difference score for each item was determined by:

$$D = [d(1) - d(2)] / \sqrt{se(1)^2 + se(2)^2}$$

where $d(1)$ = item difficulty in year 1,

$d(2)$ = item difficulty in year 2,

$se(1)$ = standard error of item difficulty in year 1,

$se(2)$ = standard error of item difficulty in year 2.

A positive value for D indicated that the item tended to be relatively more difficult in year 1.

For the assessment of absolute change, the standardized difference score for each item was calculated in a similar fashion except that the numerator of the function was replaced with $d(1) - [d(2) + k]$, with

$$k = m(1) - m(2)$$

where $m(1)$ = mean sample ability in year 1, and

$m(2)$ = mean sample ability in year 2.

The standard errors remained the same, as k was treated as a

constant applicable to all values of d(2). Again, a positive value for D indicated that the item tended to be absolutely more difficult in year 1.

Items were deemed to have changed in difficulty if the absolute value of the standardized difference score of the item was greater than 2. This procedure paralleled that of the reanalysis outlined in Chapter I, and allowed a comparison of conclusions reached using the Rasch approach as opposed to the traditional approach.

Testing Change in Content Area Difficulty

As was indicated earlier in the study there was a need for a reporting unit larger than the single item. Such reporting units, or content areas, should be small enough to provide useful curricular information yet large enough to avoid the overwhelming influence of just one or two widely fluctuating item difficulties.

Tentative units, subject to modification through the deletion of non-fitting items, were established using the following procedure. Ten categories of content were proposed by the author. The author, and two other persons experienced in the theory and practice of teaching mathematics, independently assigned items to the ten categories. Where all three persons agreed on the placement of an item, that designation was final. Agreement was reached immediately on 60 of the 89 items. By broadening or narrowing the descriptive title of five categories, the remaining items were assigned. In all cases, assignment required unanimous

agreement by the three judges.

The content areas and the items assigned to them are as follows (R refers to items on the reasoning test; C to items on the computation test):

1. Whole number concepts and operations (11 items)

   R: 14, 31, 41

   C: 1, 4, 5, 6, 7, 8, 9, 18

2. Applications using whole numbers (9 items)

   R: 1, 2, 4, 6, 15, 22, 28, 32

   C: 31

3. Common fraction concepts and operations (12 items)

   R: 33, 34, 44

   C: 10, 11, 12, 13, 15, 21, 22, 24, 36

4. Applications using common fractions (8 items)

   R: 8, 9, 11, 13, 16, 19, 21, 23

5. Decimals (11 items)

   R: 5, 17, 35, 37, 39

   C: 2, 3, 16, 17, 20, 32

6. Money (9 items)

   R:3, 7, 10, 18, 20, 29, 36, 43

   C: 38

7. Percent (8 items)

   R: 25, 30, 42

   C: 14, 19, 35, 41, 43

8. Elementary algebra (9 items)

   R: 12, 26, 38, 40

   C: 23, 37, 39, 40, 42

9. Geometry and graphing (8 items)

   R: 27, 45

   C: 25, 26, 29, 30, 34, 44

10. Units of measure (4 items)

   R: 24

   C: 27, 28, 33

Change for each group was assessed by comparing the mean values of the constituent item difficulties on the two administrations. Since the Rasch procedure placed estimated item difficulties on an equal interval scale the calculation of the mean values was psychometrically defensible. Each difficulty value contributing to the mean had its own associated standard error of calibration. Each item was treated as a stochastically independent unit as required by the third assumption of the model. Hence, the variance of the sum of item difficulties was assumed equal to the sum of the variances, and the variance of the mean was determined by dividing the sum of the variances by the number of items making up the group. If the absolute value of the difference of the means was greater than two standard errors of the difference of the mean, it was concluded that change had occurred in the overall group difficulty, and in the direction indicated by the means.

Two types of comparisons, paralleling those for individual items, were made. Unadjusted score comparisons yielded information on changing emphasis within the curriculum. Comparisons of scores adjusted for calibrated ability differences indicated trends across time.

CHAPTER IV

RESULTS

The structure of this chapter parallels that of Chapter III. In Chapter III the procedures for gathering and analyzing data were made explicit; in the present chapter the results of those procedures are given in detail. In order to interpret and assess the significance of the results, considerable discussion is included in this chapter. However, a general discussion of the model and its suitability for measuring change is reserved for Chapter V, the final chapter.

Verification of Data

For 1964 and 1970 a comparison was made of the computer-generated matrix of 1's and 0's resulting from the rescoring of the tests, with the Department's tally sheet of 1's and blanks. It was found that the markers in 1964 had made 35 scoring errors (12 on one examinee), and two tabulation errors had been made on the tally sheet, for an overall error rate of 0.14%. A check of the original booklets revealed that all the scoring errors consisted of the acceptance of the correct response on a multiply-marked item.

A similar comparison for 1970 showed a total of six Departmental scoring errors.

In 1979, the responses had been commercially keypunched onto computer cards. A 10% random sample of thirty test papers selected from the 1979 sample yielded no discrepancies between keypunched responses and actual responses. From these results it was assumed that the keypunching firm's guarantee of accuracy was confirmed.

The Deletion of Persons

In order to eliminate subjects for whom the test appeared to be too long, it had been decided to remove those individuals who had omitted ten or more items at the end of any of the three timed portions of the tests. As a result, four persons were removed from the 1964 data base, twelve from that of 1970, and twelve for 1979. All subsequent analyses were therefore based upon samples of 296, 288, and 288 for the years 1964, 1970, and 1979, respectively.

Summary Raw Score Statistics

Summary statistics for the raw scores in the three samples are shown in Table 4.1. The tests contained 45 items on the reasoning test and 44 items on the computation test. There is a consistent decline in the mean score and a consistent increase in the variability. The reliability of

the tests is consistently high.

Table 4.1

Summary Raw Score Statistics

Year	Reasoning Test			Computation Test			Total Test		
	Mean	S.D.	Rel[1]	Mean	S.D.	Rel[1]	Mean	S.D.	Rel[2]
1964	31.31	6.39	0.83	30.69	6.08	0.83	62.00	11.66	0.85
1970	30.63	6.82	0.84	27.98	6.74	0.85	58.61	12.75	0.87
1979	28.03	7.95	0.88	25.03	6.84	0.84	53.06	14.07	0.89

[1] Hoyt estimate of reliability
[2] Cronbach's composite alpha

Tests of Unidimensionality

One of the key requirements of the study was to demonstrate that the reasoning and computation tests measured the same ability. Three procedures were used to investigate the unidimensionality of the two tests.

As part of the initial data analysis, the Pearson product-moment correlation coefficient between reasoning and computation raw scores was determined for each sample. The values were 0.745, 0.769, and 0.809 for the years 1964, 1970, and 1979, respectively. Corrected for attenuation, the coefficients were 0.898, 0.910, and 0.941, respectively. Using appropriate procedures (Glass & Stanley, pp. 308-310; Forsyth & Feldt, 1969), all six coefficients were found to be different from zero at the 0.001 level of significance. Thus, scores achieved by individuals on the two tests in any given

year were highly positively correlated.

It must be noted that the correlation coefficients were based on raw scores rather than on Rasch abilities. An inspection of the sample test characteristic curve in Appendix C shows the transformation of raw scores to Rasch ability scores in the interval from -2 to +2 logits to be approximately linear. Generally, virtually no student abilities fell below -2 logits, and only about 10% were above +2 logits. Hence, if Rasch abilities were used instead of raw scores, the significance of the correlation coefficients should not be materially diminished.

The second test for unidimensionality consisted of examining the non-fitting items when all 89 items were calibrated as a single group for each sample. It had been argued that, if items did not show misfit of predominantly reasoning or computation items, it could be assumed that they measured the same ability.

The computer program BICAL was used to calibrate the items in each sample. All 89 items were included and, in each case, the minimum score of 32 was used to eliminate examinees near the guessing level. In all three cases, visual inspection of the output of items ordered from best to worst fit mean square showed no discernible separation between reasoning (R) and computation (C) items. Figure 4.1 shows the pattern of the worst fitting 20% of the items for each year. Reading from left to right the items become more ill-fitting. The sequence of reasoning and computation items appeared to be randomly ordered.

Year	Worst Fitting 20% of Items
1964	CCCCCRRRCRCRRRRCCC
1970	RCRRRCRCCRRRCRCCCR
1979	CRRCCRRCRRRCCCRCCR

Figure 4.1. Patterns of ill-fitting items.

The outcome may have resulted from the equal numbers of items from each category. Had, for example, the tests consisted of 79 straight-forward computations and 10 word problems, the word problems may have shown lack of fit as a unit. That is, the reading requirement in word problems may have ordered examinees in a different manner than the bulk of the remaining items. Nevertheless, for the tests analyzed, no distinction was apparent between the two nominally different types of items. The results were accepted as a preliminary demonstration of unidimensionality.

The final test for unidimensionality consisted of comparing standardized difference scores with the unit normal distribution. The first step in the process was the determination of each person's ability on the reasoning test and the computation test. Table 4.2 shows the mean ability estimate on each test for each year.

Each person was then assigned an adjusted difference score. For example, in 1970, the score was determined as: [reasoning ability] - [computation ability] - 0.22. Finally, each difference score was divided by the pooled standard error for the two originally determined abilities. The two-tailed

Table 4.2

Mean Ability Estimates on the Reasoning and Computation Tests

Test	Year		
	1964	1970	1979
Reasoning	1.17	1.06	0.70
Computation	1.23	0.84	0.41
(difference)	-0.06	0.22	0.29

probability level for rejection of the null hypothesis of unit normality had been set at 0.01. The test used was the Kolmogorov-Smirnov (K-S) goodness of fit test. Table 4.3 shows the results for the three years.

Table 4.3

Distributions of Standardized Difference Scores on the Reasoning and Computation Tests

Year	Mean	S.D.	Skew	Kurt	K-S Z	p
1964	0.02	1.18	0.04	-0.07	1.28	0.078
1970	0.02	1.20	-0.17	-0.41	1.42	0.035
1979	0.00	1.14	-0.25	-0.16	1.21	0.108

In no case was the skewness or kurtosis different by more than two standard deviations of each from zero. The main contributing factor to the departure from normality appeared to be the standard deviation of the difference scores; the greater the divergence from unity, the lower the probability

of unit normality. Nevertheless, in no case was the hypothesis of unit normality rejected at the predetermined 0.01 level of significance.

While the requirements of the study were met, the probability levels were quite low. Several alternative procedures were followed to see if the concordance between ability estimates could be improved.

In the first attempt at improving the agreement between the ability estimates, the non-fitting items on each test in 1964 were eliminated from the analysis. In particular, Items 27, 29, and 45 were removed from the reasoning test and Items 2, 5, 33, and 40 from the computation test. The criterion for removal was a fit mean square of four or more standard errors from unity and a discrimination index less than 0.70. The BICAL program was re-run and standardized difference of ability scores were recalculated. The probability for rejection of the hypothesis of unit normality was 0.110, as compared with 0.078 in the first instance.

The same procedure carried out on the 1970 data base resulted in the elimination of Items 4, 27, 29, 43, and 45 from the reasoning test and Items 5, 33, 39, 40, and 43 from the computation test. The probability for rejection of the hypothesis of unit normality was 0.034, as opposed to 0.035 in the former analysis. The results in these two cases demonstrate the robustness of the ability estimates. They tend to confirm the conclusions reached in Chapter II that recalibration likely has little effect on the estimates of person abilities. Support for this position may be found in

Wright and Douglas (1977b) where, in simulated runs, random disturbances in item calibrations could be as large as 1 logit before distortions in estimates of abilities reached 0.1 logit.

A second alternative for improving the concordance between ability estimates was explored. Although the calibration procedure eliminates low-scoring examinees for purposes of item calibration, the ability estimates are given for all persons in the sample. Thus the mean ability estimate includes persons scoring below the chance level. It seemed plausible that the overall K-S probability should improve if person scores at the chance level on one or both tests were excluded from the sample. The 1970 data base was edited to remove such persons, and the mean ability estimates were recalculated. As a result, 15 persons were deleted from the sample. The reanalysis on 273 subjects produced a probability for rejecting the hypothesis of unit normality of 0.050. It was felt that the improvement from 0.035 was not important, and the reanalysis was not carried out on the remaining samples. The lack of substantial improvement may have been due, in part, to the small number of examinees eliminated from the sample.

Each of the three procedures tended to confirm the commonality of the ability trait on the two tests. This result, however, is applicable only to the very specific sets of items assigned by the Stanford Achievement Test developers to the categories "arithmetic reasoning" and "arithmetic computation". It was pointed out in Chapter I that at least

one observer had noted an apparent overlap in content. The results may have been different had the two tests contained items more truly representative of the two descriptors.

In summary, the standardized difference scores appeared to be stable. Neither the deletion of non-fitting items nor of persons scoring below the guessing level had a major effect on the distribution of scores. The hypothesis of unit normality was not rejected at the 0.01 level of significance for any of the three samples. The assumption that the abilities measured by the two tests were indistinguishable was considered to be upheld.

Item Calibration

Having accepted the assumption of unidimensionality, the two tests were combined and treated as a single test for the remainder of the analysis. For each year the minimum acceptable score was set at 32. BICAL runs showed no zero scores and no perfect scores on any of the three samples. The size of each calibration sample is shown in Table 4.4.

The use of a cut-off score to eliminate examinees scoring below the guessing level did not seriously affect the sample size for calibration. Generally, the tests were easy for the students in each administration, producing few low raw scores. The most serious case was the initial calibration of the items in 1979 when 25 out of the 288 examinees were removed. Nevertheless, this accounted for less than 10% of the sample. As Wright (1977b) points out, the standard error

Table 4.4

Number of Subjects in Each Calibration

Year	Total Number of Subjects	Subjects Scoring Less Than 32	Subjects in Calibration
1964	296	4	292
1970	288	7	281
1979	288	25	263

of the item calibration is dominated by the reciprocal of the square root of sample size. In this instance, the minimum standard error in the years 1964, 1970, and 1979 were 0.126, 0.129, and 0.133, based on samples of 292, 281, and 263, respectively. Rounded to two decimal places, the standard errors are indistinguishable.

The fit mean square standard error was 0.08 in 1964, 0.08 in 1970, and 0.09 in 1979. The items for which the fit mean square (FMS) was four or more standard errors greater than unity on the 1964, 1970, and 1979 analyses are shown in Table 4.5 (R=reasoning, C=computation). Their discrimination values (Disc) are also indicated in the same table.

The items marked with an asterisk (*) in Table 4.5 demonstrated lack of fit on two out of the three administrations. They accounted for 6 of the 8 items not meeting the fit mean square criterion in 1964, 6 out of 7 in 1970, and 4 out of 5 in 1979. Eight items in total were deemed not to fit the model. The characteristics of these items are given in Table 4.6.

Table 4.5

Items Not Meeting Fit Mean Square Criterion

1964			1970			1979		
Item	Disc	FMS	Item	Disc	FMS	Item	Disc	FMS
C 5*	0.47	1.32	R27*	0.11	1.32	R33	1.02	1.41
R27*	0.15	1.35	C43*	0.49	1.33	C43*	0.28	1.48
R29*	0.19	1.38	R45*	0.11	1.34	C40*	0.17	1.59
R45*	0.18	1.40	C33*	0.08	1.37	C 5*	0.35	1.74
R 4*	0.59	1.43	C40*	0.49	1.40	R 4*	0.53	1.74
C33*	0.07	1.45	C 7	0.49	1.42			
C 2	0.49	1.45	R29*	0.33	1.49			
C 1	0.79	1.59						

*: items showing lack of fit on two of three administrations

Table 4.6

Characteristics of Non-Fitting Items

1964				1970				1979			
Item	Diff	Disc	FMS		Diff	Disc	FMS		Diff	Disc	FMS
R 4	-0.92	0.59	1.43		-1.28	0.69	1.16		-1.34	0.53	1.74
R27	0.56	0.15	1.35		0.93	0.11	1.32		0.38	0.45	1.24
R29	2.37	0.19	1.38		2.44	0.33	1.49		2.07	0.58	1.26
R45	2.06	0.18	1.40		2.06	0.11	1.34		2.12	0.35	1.25
C 5	-1.33	0.47	1.32		-1.04	0.55	1.12		-1.48	0.35	1.74
C33	0.29	0.07	1.45		0.61	0.08	1.37		0.07	0.09	1.30
C40	3.33	0.50	1.28		3.20	0.49	1.40		2.93	0.17	1.59
C43	2.60	0.98	1.05		2.16	0.49	1.33		2.43	0.28	1.48

After deletion of the eight items the set of remaining items was recalibrated for each year using a minimum acceptable score of 27. In those three years there were 4, 5, and 14 subjects scoring below 27, resulting in the number of subjects in the calibration of 292, 283, and 274,

respectively. The fit mean square standard errors remained unchanged. Two more items demonstrated lack of fit on two of the three calibrations as shown in Table 4.7. Both items were of high difficulty and both demonstrated consistent lack of fit. Both were in the 20% of the original items having the highest fit mean square on each administration.

Table 4.7

Non-Fitting Items on Recalibration

Item	1964			1970			1979		
	Diff	Disc	FMS	Diff	Disc	FMS	Diff	Disc	FMS
R43	2.26	0.36	1.34	1.74	0.27	1.36	1.82	0.34	1.36
C39	2.53	0.59	1.28	2.20	0.44	1.33	1.95	0.44	1.38

The two non-fitting items were deleted from the analysis and the remaining 79 items were recalibrated. The minimum acceptable score was set at 26. This resulted in the deletion of 3, 5, and 13 subjects, leaving 293, 283, and 275 in the calibration sample for each year. Fit mean square standard errors were unchanged. The ill-fitting items for each year are shown in Table 4.8.

The only item showing lack of fit on two administrations was C13. In each instance the fit mean square was close to the critical value. In 1970 and 1979, the critical fit mean square values were 1.32 and 1.36; the values for Item C13 were 1.32 and 1.37. It was considered likely

Table 4.8

Ill-Fitting Items in Final Calibration

Item	1964			1970			1979		
	Diff	Disc	FMS	Diff	Disc	FMS	Diff	Disc	FMS
R 8				-0.75	0.64	1.42			
R28	2.12	0.14	1.37						
R33				-1.47	0.55	1.48			
C 2	-1.12	0.60	1.92						
C 4	-3.20	0.61	1.33						
C 7				-0.84	0.51	1.48			
C13				1.03	0.39	1.32	0.17	0.51	1.37
C21	-1.25	0.67	1.39						

that the deletion of this item would bring about only marginal improvement in calibration, and the deletion process was stopped. In summary, the ten items deleted were R4, R27, R29, R43, R45, and C5, C33, C39, C40, C43.

The value of a two-part criterion for assessing the fit of items to the model was borne out by an inspection of Table 4.5. Items were designated as non-fitting if, on two of the three administrations, they demonstrated a fit mean square four or more standard errors greater than unity, and a discrimination index less than 0.7. In Table 4.5, it can be seen that the same eight items would have been deemed non-fitting had only the fit mean square criterion been used. This is explained by the high negative correlation between fit mean square and discrimination (1964: -0.66; 1970: -0.90; 1979: -0.83). However, had such a single criterion been adopted, on recalibration, three more items including the two deleted using the two-part criterion, would have been deleted.

This, in turn, would likely have led to the deletion of a further three items on the second recalibration, with more deletions possible. Thus, the second part of the criterion formed a valuable check on the deletion process.

The problem of using a single criterion was encountered by Fryman (1976), who adopted the criterion of a chi-square probability less than or equal to 0.05 for rejection of an item. Applying this to an existing Mathematics Placement test of 100 items, and using a computer program different from BICAL, he eliminated 13 items on the first calibration, 18 on the second, 12 on the third, 7 on the fourth, and 8 on the fifth. At this point he stopped since his stated intent was to develop an instrument which could be completed in a maximum of one hour. In his conclusion, Fryman cited Kifer, who suggested using an additional criterion based on the slope (discrimination). That suggestion results in a criterion similar to the one used in the present study.

A useful statistic which might have given objective evidence for the improvement in the estimates of parameters after recalibration is the overall chi-square statistic described in Chapter II. This statistic compares the observed and expected values across the entire raw score/item matrix. In the documentation for the program BICAL, Wright and Mead (1978) indicate how the statistic can be constructed. It is regrettable that they did not include it in their program.

The Rasch model appears to be suitable for detecting test items whose psychometric properties are suspect. With one exception, the non-fitting items consistently demonstrated

high fit mean square and low discrimination. The only item which appeared to fit well in one year but not the others, was C43; in 1964, both fit mean square and discrimination were quite acceptable. It must be kept in mind that the criterion deleted items showing generally poor test characteristics on all three administrations. There still remained in the analysis those items showing lack of fit on a single administration. For example, in 1964, Items R28, C2, C4, and C21 did not meet fit criteria. In some cases the reason for the singularity of lack of fit was evident. For example, in 1964, Item C4 was answered correctly by 97.3% of the examinees, providing little scope for any discriminating power. On the other hand, for R28, a difficult item, there was evidence of consistent guessing across all three years, but only in 1964 did the figures exceed the criterion.

Six of the ten deleted items were high difficulty items. The explanation of this fact may lie in the tendency of students to guess on such items.

An inspection of some of the deleted items yielded possible explanations for lack of fit. For example, item C33 was:

Add 15 m. 8 cm.
4 m. 5 cm.

One would expect the item to show little discrimination between high ability and low ability students, since the correct answer could be obtained by adding the component parts, without any knowledge of the conversion factor from metres to centimetres. Item C40, on the other hand, turned

out to be one of the two most difficult items on the test. It required the multiplication of +4a and -3, an inappropriate item for Grade 7 students in British Columbia, since operations on integral algebraic expressions have never been part of the curriculum at that level. Results no better than chance could be expected, and were obtained. On the other hand, the reason for deletion of Item R4 remains obscure. There appears to be no obvious flaw in: "Dot's mother is going to buy tomato plants to set out. There are to be 14 rows with 18 plants in each row. How many plants will be needed?"

On the initial calibration, none of the items showed lack of fit on all three administrations. Six of the eight non-fitting items showed lack of fit on two successive administrations. This may indicate changing trends in curriculum related to each item. For example, C43, a difficult question on simple interest, showed consistently increasing fit mean square and decreasing discrimination across time. This may reflect a move away from teaching this topic.

Final item difficulties generally were located in the interval from -3.0 to +3.0 logits. Only one or two items in each calibration fell outside these limits. This was to be expected since the initial difficulty estimates were set at ln[%incorrect/%correct]. For a difficulty value of +3.0, for example, approximately 95% of the responses would have to be incorrect, and this is roughly the upper limit for the usual standardized test. Item difficulties and their standard errors are shown in Table 4.9.

Table 4.9

Item Difficulties and Standard Errors

Item	1964			1970			1979		
	Diff	Std Err	Adj Diff	Diff	Std Err	Adj Diff	Diff	Std Err	Adj Diff
R 1	-2.26	0.33	-2.26	-2.28	0.30	-2.03	-2.63	0.30	-1.94
R 2	-2.37	0.34	-2.37	-2.19	0.30	-1.94	-1.53	0.20	-1.25
R 3	-2.79	0.42	-2.79	-2.37	0.31	-2.12	-1.96	0.23	-1.27
R 5	-1.16	0.21	-1.16	-1.72	0.24	-1.47	-1.42	0.19	-0.73
R 6	-1.45	0.23	-1.45	-1.97	0.26	-1.72	-1.69	0.21	-1.00
R 7	-0.25	0.16	-0.25	-0.45	0.16	-0.20	-0.48	0.15	0.21
R 8	-0.82	0.19	-0.82	-0.75	0.17	-0.50	-0.83	0.16	-0.14
R 9	-0.66	0.18	-0.66	-0.38	0.16	-0.13	-0.27	0.14	0.42
R10	-0.32	0.16	-0.32	-1.00	0.19	-0.75	-0.99	0.17	-0.30
R11	-1.30	0.22	-1.30	-1.03	0.19	-0.78	-0.93	0.16	-0.24
R12	-0.32	0.16	-0.32	-0.69	0.17	-0.44	-0.50	0.15	0.19
R13	-0.72	0.18	-0.72	-0.93	0.18	-0.68	-0.29	0.14	0.40
R14	0.71	0.13	0.71	0.51	0.14	0.76	-0.19	0.14	0.50
R15	-0.13	0.15	-0.13	-0.19	0.15	0.06	0.13	0.14	0.82
R16	-0.20	0.16	-0.20	0.04	0.14	0.29	0.68	0.13	1.37
R17	-0.35	0.16	-0.35	-0.53	0.16	-0.28	-0.29	0.14	0.40
R18	-0.15	0.15	-0.15	-0.06	0.15	0.19	-0.44	0.15	0.25
R19	0.41	0.14	0.41	0.22	0.14	0.47	0.24	0.14	0.93
R20	0.46	0.14	0.46	0.40	0.14	0.65	0.11	0.14	0.80
R21	0.00	0.15	0.00	-0.24	0.15	0.01	-0.10	0.14	0.59
R22	-0.38	0.16	-0.38	0.40	0.14	0.65	0.48	0.13	1.17
R23	-0.20	0.16	-0.20	-0.28	0.15	-0.03	0.07	0.14	0.76
R24	1.86	0.13	1.86	1.47	0.13	1.72	1.11	0.13	1.80
R25	1.39	0.13	1.39	1.30	0.13	1.55	1.40	0.14	2.09
R26	0.98	0.13	0.98	0.79	0.13	1.04	0.66	0.13	1.35
R28	2.12	0.13	2.12	1.68	0.13	1.93	1.04	0.13	1.73
R30	2.09	0.13	2.09	2.16	0.14	2.41	2.19	0.16	2.88
R31	-0.85	0.19	-0.85	-1.61	0.23	-1.36	-2.07	0.24	-1.38
R32	0.64	0.13	0.64	0.42	0.14	0.67	-0.31	0.14	0.38
R33	-0.51	0.17	-0.51	-1.47	0.22	-1.22	-1.19	0.18	-0.50
R34	1.68	0.13	1.68	1.18	0.13	1.43	0.92	0.13	1.61
R35	-0.57	0.17	-0.57	0.42	0.14	0.67	0.47	0.13	1.16
R36	0.64	0.13	0.64	0.42	0.14	0.67	0.22	0.14	0.91
R37	0.69	0.13	0.69	0.66	0.13	0.91	-0.50	0.15	0.19
R38	-0.15	0.15	-0.15	-1.00	0.19	-0.75	-0.46	0.15	0.23
R39	0.27	0.14	0.27	0.42	0.14	0.67	0.11	0.14	0.80
R40	1.36	0.13	1.36	0.83	0.13	1.08	0.95	0.13	1.64
R41	1.90	0.13	1.90	1.12	0.13	1.37	0.48	0.13	1.17
R42	0.37	0.14	0.37	0.33	0.14	0.58	-0.29	0.14	0.40
R44	1.86	0.13	1.86	1.30	0.13	1.55	1.40	0.14	2.09

Table 4.9 - cont'd.

Item	1964			1970			1979		
	Diff	Std Err	Adj Diff	Diff	Std Err	Adj Diff	Diff	Std Err	Adj Diff
C 1	-2.79	0.42	-2.79	-1.67	0.23	-1.42	-1.56	0.20	-0.87
C 2	-1.12	0.21	-1.12	-1.78	0.24	-1.53	-2.07	0.24	-1.38
C 3	-1.56	0.24	-1.56	-1.29	0.20	-1.04	-1.28	0.18	-0.59
C 4	-3.20	0.51	-3.20	-3.42	0.51	-3.17	-2.93	0.34	-2.24
C 6	-0.82	0.19	-0.82	-1.61	0.23	-1.36	-1.53	0.20	-0.84
C 7	-0.69	0.18	-0.69	-0.84	0.18	-0.59	-0.73	0.16	-0.04
C 8	-1.30	0.22	-1.30	-1.25	0.20	-1.00	-1.04	0.17	-0.35
C 9	-0.85	0.19	-0.85	-0.69	0.17	-0.44	-0.64	0.15	0.05
C10	-1.40	0.23	-1.40	-0.90	0.18	-0.65	-0.27	0.14	0.42
C11	-0.51	0.17	-0.51	0.22	0.14	0.47	0.54	0.13	1.23
C12	-1.25	0.22	-1.25	-0.26	0.15	-0.01	-0.96	0.17	-0.27
C13	0.33	0.14	0.33	1.03	0.13	1.28	0.17	0.14	0.86
C14	0.31	0.14	0.31	1.10	0.13	1.35	0.64	0.13	1.33
C15	-0.72	0.18	-0.72	-0.69	0.17	-0.44	-0.29	0.14	0.40
C16	0.03	0.15	0.03	0.79	0.13	1.04	0.17	0.14	0.86
C17	0.48	0.14	0.48	0.22	0.14	0.47	0.17	0.14	0.86
C18	0.00	0.15	0.00	0.95	0.13	1.20	0.33	0.13	1.02
C19	0.27	0.14	0.27	0.96	0.13	1.21	0.87	0.13	1.56
C20	0.05	0.15	0.05	0.91	0.13	1.16	0.45	0.13	1.14
C21	-1.25	0.22	-1.25	-1.67	0.23	-1.42	-1.04	0.17	-0.35
C22	0.46	0.14	0.46	0.88	0.13	1.13	1.36	0.14	2.05
C23	0.07	0.15	0.07	-0.87	0.18	-0.62	-0.53	0.15	0.16
C24	1.21	0.13	1.21	1.25	0.13	1.50	1.22	0.14	1.91
C25	0.11	0.15	0.11	-0.28	0.15	-0.03	-0.53	0.15	0.16
C26	1.05	0.13	1.05	1.03	0.13	1.28	0.97	0.13	1.66
C27	0.23	0.14	0.23	0.90	0.13	1.15	2.36	0.16	3.05
C28	0.35	0.14	0.35	0.79	0.13	1.04	1.98	0.15	2.67
C29	0.25	0.14	0.25	-0.02	0.15	0.23	-0.50	0.15	0.19
C30	-0.57	0.17	-0.57	-0.78	0.17	-0.53	-0.88	0.16	-0.19
C31	-0.54	0.17	-0.54	0.16	0.14	0.41	0.40	0.13	1.09
C32	1.37	0.13	1.37	1.58	0.13	1.83	0.88	0.13	1.57
C34	1.93	0.13	1.93	2.68	0.16	2.93	2.00	0.15	2.69
C35	-0.06	0.15	-0.06	0.16	0.14	0.41	0.48	0.13	1.17
C36	-2.06	0.30	-2.06	-1.97	0.26	-1.72	-1.45	0.19	-0.76
C37	2.05	0.13	2.05	0.50	0.14	0.75	1.71	0.14	2.40
C38	2.17	0.13	2.17	2.45	0.15	2.70	2.09	0.15	2.78
C41	1.70	0.13	1.70	1.75	0.13	2.00	1.75	0.14	2.44
C42	1.32	0.13	1.32	0.90	0.13	1.15	1.08	0.13	1.77
C44	3.41	0.17	3.41	3.82	0.22	4.07	3.36	0.22	4.05

For each calibration the standard error of item difficulty ranged approximately from 0.130 to 0.400. On each test, items nearest in difficulty to the mean ability were calibrated with the lowest standard error. The effect of changing mean abilities was not great. For example, an item of average difficulty on each of the three administrations had a standard error of 0.148, 0.145, and 0.137 in 1964, 1970, and 1979. The constant decrease in standard error reflects the movement of the mean abilities toward the centre of the difficulty/ability scale.

Final summary statistics for abilities based on the 79 remaining items are given in Table 4.10.

Table 4.10

Summary Statistics for Abilities

Year	Mean	S.D.
1964	1.39	0.93
1970	1.14	0.96
1979	0.70	1.00

The decline in the mean raw scores on the test is reflected in the changing mean abilities. In 1964, the mean ability was 1.39, considerably above the zero point of the scale. By 1979, the mean ability score had declined to 0.70. In Rasch terms, this meant that, when confronted with any item, the odds on success for the average 1964 student compared with that of 1979 were twice as great. This is

arrived at by determining the value of e raised to the power 1.39 - 0.70. It is coincidental that one mean ability value also happens to be twice the other.

Because mean abilities were considerably greater than zero, the tests resulted in larger standard errors of person measurement for most examinees than would have been the case had the tests been less difficult. That is, if the mean abilities had been centred around zero, the measurement of person abilities would have been more precise. However, the differences were not great: for the worst case, in 1964, the standard error for the average student was 0.29 as compared with a possible 0.26.

Changes in Item Difficulty

For each pair of years two standardized difference values for each item were calculated: one reflected relative difficulty change within the set of items, and the second reflected absolute change of difficulty across time. For changes in relative difficulty the standardized difference score was determined by subtracting the two difficulty estimates and dividing by the pooled standard error of the difficulty estimates. The resulting distributions and Kolmogorov-Smirnov statistics are shown in Table 4.11.

In all cases the hypothesis of unit normality was rejected at the 0.01 level of significance. Thus, for all comparisons, the omnibus test indicated that the relative difficulty levels of some items had changed.

Table 4.11

Distributions of Standardized Scores Related to Relative Change in Item Difficulty

Comparison	Mean	S.D.	Skew	Kurt	K-S Z	p
1964-1970	0.00	2.38	-0.15	0.70	1.71	0.006
1970-1979	-0.10	2.30	0.34	1.22	1.85	0.002
1964-1979	-0.13	3.03	0.45	1.24	2.30	0.000

In Table 4.11, the negative mean standardized scores indicate that more items were relatively easier in 1979 than in 1964 or 1970. However, since the difficulties are centred on their own mean for each year, the algebraic sum of the shifts in difficulty for each year must be zero. This would imply that, although fewer items were relatively more difficult, their average change in difficulty was greater than that for the items which had become easier. For example, there was a sharp increase in the relative difficulty of Items C27 and C28 unmatched by a similar shift for any easier items.

The criterion for deciding that a particular item had changed in difficulty was a standardized difference value whose absolute value exceeded two. Items which changed in relative difficulty in at least one comparison are shown in Table 4.12. A plus (+) indicates that the item had become more difficult in the latter year, a minus (-) indicates less difficulty, and a zero (0) indicates no change.

Summary statistics for changes in relative item difficulty are given in Table 4.13.

Table 4.12

Items Changing in Relative Difficulty

Item	1964-70	1970-79	1964-79	Item	1964-70	1970-79	1964-79
R 2	0	0	+	C 1	+	0	+
R10	-	0	-	C 2	-	0	-
R13	0	+	0	C 6	-	0	-
R14	0	-	-	C10	0	+	+
R16	0	+	+	C11	+	0	+
R22	+	0	+	C12	+	-	0
R24	-	0	-	C13	+	-	0
R28	-	-	-	C14	+	-	0
R31	-	0	-	C16	+	-	0
R32	0	-	-	C18	+	-	0
R33	-	0	-	C19	+	0	+
R34	-	0	-	C20	+	-	+
R35	+	0	+	C21	0	+	0
R36	0	0	-	C22	+	+	+
R37	0	-	-	C23	-	0	-
R38	-	+	0	C25	0	0	-
R40	-	0	-	C27	+	+	+
R41	-	-	-	C28	+	+	+
R42	0	-	-	C29	0	-	-
R44	-	0	-	C31	+	0	+
				C32	0	-	-
				C34	+	-	0
				C35	0	0	+
				C37	-	+	0
				C42	-	0	0

+: items increasing in relative difficulty in latter year
-: items decreasing in relative difficulty in latter year
0: items showing no change in relative difficulty

Table 4.13

Number of Relative Difficulty Changes

Comparison	Easier in Latter Year	Harder in Latter Year
1964-1970	14	16
1970-1979	15	9
1964-1979	20	14

For the omnibus test of absolute change in difficulty a procedure similar to that outlined for relative difficulty was followed. In this instance, however, the difficulty of each item was adjusted to the 1964 scale by adding the difference between the mean sample abilities. To the difficulty levels of the 1970 items the value of 0.25 was added to compensate for the decreased level of ability of the 1970 sample. To the 1979 difficulty levels, 0.69 was added. These values were obtained by subtracting the mean abilities shown in Table 4.10 for the years in the comparison. The adjusted difficulties are shown in Table 4.9. Table 4.14 shows the results of the analysis on these adjusted difficulties.

Table 4.14

Distributions of Standardized Scores Related to Absolute Change in Item Difficulty

Comparison	Mean	S.D.	Skew	Kurt	K-S Z	p
1964-1970	1.11	2.40	-0.04	0.56	3.10	0.000
1970-1979	1.91	2.26	0.57	1.60	5.21	0.000
1964-1979	2.97	2.97	0.68	1.52	6.05	0.000

The null hypothesis of unit normality was rejected at the 0.001 level of significance in all three comparisons. Hence, in all comparisons, the absolute difficulty levels of some items had changed.

In Table 4.14, the mean values indicate a constant trend toward increasing difficulty. There is no doubt that the distributions reflect more than just random error in difficulty calibrations. In this case, there are no constraints on the movement of item difficulties as there were for relative difficulty. The difficulty values incorporate both change in relative difficulty due to changing emphasis within the curriculum and the effect of the changing ability of the samples.

To determine absolute changes in item difficulty, again the criterion of a standardized difference value whose absolute value exceeded 2 was used. Items changing in absolute difficulty are given in Table 4.15.

Summary statistics for changes in absolute item difficulty are given in Table 4.16.

The separation of the difficulty estimates into the categories of relative and absolute can provide valuable information on change. For example, from Table 4.12, it can been seen that Items R14, R32, R37, R42, C29, and C32 were relatively easier in 1979 than in both 1964 and 1970. However, the effect of the general decline in ability was to eliminate those changes, leaving five of the six items unchanged in absolute difficulty. On the other hand, the relative gain in performance on Item R37 was sufficiently large to outweigh the decline in general ability, and the item was easier in 1979 than in previous years, as seen in Table 4.15.

Table 4.15

Items Changing in Absolute Difficulty

Item	1964-70	1970-79	1964-79	Item	1964-70	1970-79	1964-79
R 2	0	+	+	C 1	+	0	+
R 3	0	+	+	C 3	0	0	+
R 5	0	+	0	C 7	0	+	+
R 6	0	+	0	C 8	0	+	+
R 7	0	0	+	C 9	0	+	+
R 8	0	0	+	C10	+	+	+
R 9	+	+	+	C11	+	+	+
R11	0	+	+	C12	+	0	+
R12	0	+	+	C13	+	-	+
R13	0	+	+	C14	+	0	+
R15	0	+	+	C15	0	+	+
R16	+	+	+	C16	+	0	+
R17	0	+	+	C18	+	0	+
R19	0	+	+	C19	+	0	+
R21	0	+	+	C20	+	0	+
R22	+	+	+	C21	0	+	+
R23	0	+	+	C22	+	+	+
R25	0	+	+	C23	-	+	0
R26	0	0	+	C24	0	+	+
R28	0	0	-	C26	0	+	+
R30	0	+	+	C27	+	+	+
R33	-	+	0	C28	+	+	+
R35	+	+	+	C31	+	+	+
R37	0	-	-	C32	+	0	0
R38	-	+	0	C34	+	0	+
R39	+	0	+	C35	+	+	+
R40	0	+	0	C36	0	+	+
R41	-	0	-	C37	-	+	0
R44	0	+	0	C38	+	0	+
				C41	0	+	+
				C42	0	+	+
				C44	+	0	+

Table 4.16

Number of Absolute Difficulty Changes

Comparison	Easier in Latter Year	Harder in Latter Year
1964-1970	5	24
1970-1979	2	41
1964-1979	3	49

There is a caveat regarding any conclusions drawn on changing difficulty for an individual item. The criterion for deciding change was a standardized difference for which the absolute value exceeded two. This is equivalent to setting an approximate alpha level of 0.05. With the large number of comparisons made, the probability is very high that a Type I error will have been made on at least one comparison.

Changes in Content Area Difficulty

The deletion of non-fitting items resulted in changes in the items constituting each content area. The items which failed to fit the model came from across the range of curriculum topics. Seven out of the ten topics lost one or two items. The most serious effect likely was on Topic #10, Units of Measure, which lost one of its four items, further weakening a group already containing few elements. The revised item groupings were as follows:

1. Whole number concepts and operations (WNC) - 10 items

   R:14, 31, 41

   C:1, 4, 6, 7, 8, 9, 18

2. Applications using whole numbers (WNA) - 8 items

   R: 1, 2, 6, 15, 22, 28, 32

   C: 31

3. Common fraction concepts and operations (CFC) - 12 items

   R: 33, 34, 44

   C: 10, 11, 12, 13, 15, 21, 22, 24, 36

4. Applications using common fractions (CFA) - 8 items

   R: 8, 9, 11, 13, 16, 19, 21, 23

5. Decimals (Dec) - 11 items

   R: 5, 17, 35, 37, 39

   C: 2, 3, 16, 17, 20, 32

6. Money (Mon) - 7 items

   R: 3, 7, 10, 18, 20, 36

   C: 38

7. Percent (Pct) - 7 items

   R: 25, 30, 42

   C: 14, 19, 35, 41

8. Elementary algebra (Alg) - 7 items

   R: 12, 26, 38, 40

   C: 23, 37, 42

9. Geometry and graphing (Geo) - 6 items

   C: 25, 26, 29, 30, 34, 44

10. Units of measure (Mea) - 3 items

   R: 24

   C: 27, 28

For each content area the mean difficulty was calculated. The standard error of the mean was determined by summing the squares of the constituent standard errors, dividing by the number of items in the group, and taking the square root of the result. Mean values, standard errors of

the mean, and means adjusted to the 1964 scale for each administration are shown in Table 4.17.

Table 4.17

Summary Statistics for Content Areas

Content Area	1964 Mean	1964 Std Err	1964 Adj Mean	1970 Mean	1970 Std Err	1970 Adj Mean	1979 Mean	1979 Std Err	1979 Adj Mean
1. WNC	-0.79	0.18	-0.79	-1.04	0.24	-0.79	-0.99	0.20	-0.30
2. WNA	-0.55	0.22	-0.55	-0.50	0.21	-0.25	-0.51	0.18	0.18
3. CFC	-0.18	0.19	-0.18	-0.09	0.16	0.17	0.03	0.15	0.72
4. CFA	-0.44	0.17	-0.44	-0.42	0.16	-0.17	-0.18	0.14	0.52
5. Dec	-0.17	0.17	-0.17	-0.03	0.17	0.22	-0.30	0.16	0.39
6. Mon	-0.03	0.21	-0.03	-0.09	0.19	0.16	-0.21	0.16	0.48
7. Pct	0.87	0.14	0.87	1.11	0.14	1.36	1.01	0.14	1.70
8. Alg	0.76	0.14	0.76	0.07	0.15	0.32	0.42	0.14	1.11
9. Geo	1.03	0.15	1.03	1.08	0.17	1.33	0.74	0.16	1.43
10. Mea	0.81	0.14	0.81	1.05	0.13	1.30	1.82	0.15	2.51

In all three administrations, the order of difficulty of the ten content areas was roughly the same as their numerical order, Whole Numbers being easiest with Geometry, Percent, and Units of Measure consistently being the most difficult. The standard error of the means tended to decrease to a minimum value as the mean group difficulty approached the mean ability for the year. In general, as the mean difficulty increased, the standard error decreased.

To determine relative change, the standardized difference of the content area means for the years in the comparison was determined. Any standardized difference of means for which the absolute value was greater than two was

taken to indicate a change.

To determine absolute change a similar procedure was followed. In this case, the comparison was made of the difficulties adjusted to the 1964 scale. The standard errors of the means were the same as in the test for change in relative difficulty.

The results of both relative and absolute comparisons are given in Table 4.18.

Table 4.18

Changes in Content Area Difficulty

Content Area	No of Items	1964-1970		1970-1979		1964-1979	
		Rel	Abs	Rel	Abs	Rel	Abs
1. WNC	10	0	0	0	0	0	0
2. WNA	8	0	0	0	0	0	+
3. CFC	12	0	0	0	+	0	+
4. CFA	8	0	0	0	+	0	+
5. Dec	11	0	0	0	0	0	+
6. Mon	7	0	0	0	0	0	0
7. Pct	7	0	+	0	0	0	+
8. Alg	7	-	-	0	+	0	0
9. Geo	6	0	0	0	0	0	0
10. Mea	3	0	+	+	+	+	+

The range of the standard error of the difference of means in the year to year comparisons was approximately 0.20 for the higher difficulty topics to 0.30 for those of less difficulty. Thus, in order to be found significantly different, the measures had to differ by roughly 0.40 to 0.60 logits. For each comparison, only one topic changed in

relative difficulty. On absolute difficulty, however, there was evidence for increasing difficulty, with six of ten content areas more difficult in 1979 than in 1964. The sole exception was the Elementary Algebra section which became easier from 1964 to 1970, but that advantage was lost from 1970 to 1979.

The interpretation of these results must be tempered by the knowledge that the content areas were made up of differing numbers of items. The statistical procedure for deciding that a change had occurred did not take this factor into account. The degree of confidence with which generalizations can be made is dependent upon the number and representativeness of the items subsumed under any one topic. For example, although Topic #10, Units of Measure, showed consistently increasing difficulty, that unit contained only three items; one on time which was unchanged in difficulty, and two on Imperial units of weight, the latter two dominating the former. More confidence should be placed in conclusions reached on say, Topic #3, Common Fraction Concepts and Operations, which contained 12 diverse items.

Comparison of Results Using Rasch and Traditional Procedures

To determine whether decisions on change would differ depending on whether the Rasch model or the traditional approach were used, a further analysis was made of the 79 retained items. P-values, in the form of the percentage of incorrect responses on each item for each year, were

calculated, along with the standard error associated with each P-value. The usual standardized difference scores were calculated. The results are shown in Table 4.19. If the absolute value of the standardized difference exceeded two, the item was presumed to have changed in absolute difficulty.

Discrepancies between decisions made using the two models were observed in 27 cases out of the 237 item comparisons made. The items on which discrepancies occurred and the nature of those discrepancies are shown in Table 4.20.

On 24 items, use of the traditional model would lead to the judgment of increasing difficulty, whereas a conclusion of no change would be made using the Rasch model. In three cases the Rasch model indicated decreasing difficulty, while the traditional approach indicated no change. The mean percent difficulty of the 24 items was 22.7, and that of the 3 items, 46.6.

An analysis of content area difficulty parallel to that previously described for the Rasch model was carried out using P-values, traditional standard errors, and the customary standardized difference of group means. In this case only the absolute change in difficulty was determined from year to year. The results are shown in Table 4.21.

Table 4.19

Traditional Analysis of Change

Item	Percent Incorrect			Standard Error			Stand. Difference		
	1964	1970	1979	1964	1970	1979	64-70	70-79	64-79
R 1	4.43	6.60	11.81	1.24	1.47	1.90	0.97	2.17	3.12
R 2	4.39	6.60	19.10	1.19	1.47	2.32	1.17	4.56	5.64
R 3	3.04	5.90	14.24	1.00	1.39	2.06	1.67	3.35	4.88
R 5	10.47	9.03	18.75	1.78	1.69	2.30	-0.59	3.40	2.84
R 6	8.45	7.64	17.36	1.62	1.57	2.24	-0.36	3.56	3.23
R 7	19.93	21.87	31.94	2.33	2.44	2.75	0.58	2.74	3.33
R 8	13.51	17.71	25.35	1.99	2.25	2.57	1.40	2.24	3.64
R 9	15.20	22.22	34.03	2.09	2.45	2.80	2.18	3.17	5.39
R10	18.92	14.58	23.96	2.28	2.08	2.52	-1.40	2.87	1.48
R11	9.12	14.58	23.96	1.68	2.08	2.52	2.04	2.87	4.90
R12	19.26	18.40	30.21	2.30	2.29	2.71	-0.26	3.33	3.08
R13	14.19	15.28	31.94	2.03	2.12	2.75	0.37	4.79	5.19
R14	36.49	37.50	34.03	2.80	2.86	2.80	0.25	-0.87	-0.62
R15	21.62	25.69	41.67	2.40	2.58	2.91	1.16	4.11	5.32
R16	20.61	28.82	51.04	2.36	2.67	2.95	2.30	5.58	8.06
R17	18.92	20.14	32.99	2.28	2.37	2.78	0.37	3.52	3.92
R18	21.62	27.43	30.56	2.40	2.63	2.72	1.63	0.83	2.46
R19	30.41	31.94	42.71	2.68	2.75	2.92	0.40	2.68	3.11
R20	31.42	35.42	40.28	2.70	2.82	2.90	1.02	1.20	2.24
R21	23.65	24.31	35.76	2.47	2.53	2.83	0.19	3.02	3.22
R22	18.58	35.42	47.22	2.26	2.82	2.95	4.65	2.89	7.71
R23	20.61	23.95	38.89	2.36	2.52	2.88	0.97	3.90	4.92
R24	59.80	57.29	60.07	2.85	2.92	2.89	-0.61	0.68	0.07
R25	50.34	53.47	64.93	2.91	2.94	2.82	0.76	2.81	3.60
R26	41.55	43.06	51.04	2.87	2.92	2.95	0.37	1.92	2.31
R28	64.86	61.11	59.38	2.78	2.88	2.90	-0.94	-0.43	-1.37
R30	64.19	70.14	78.13	2.79	2.70	2.44	1.53	2.19	3.76
R31	13.18	9.72	14.93	1.97	1.75	2.10	-1.31	1.90	0.61
R32	34.80	35.76	32.29	2.77	2.83	2.76	0.24	-0.88	-0.64
R33	16.89	11.11	21.53	2.18	1.86	2.43	-2.02	3.41	1.42
R34	56.42	51.04	55.21	2.89	2.95	2.94	-1.30	1.00	-0.29
R35	15.88	36.11	47.22	2.13	2.84	2.95	5.71	2.72	8.62
R36	34.80	36.11	43.40	2.77	2.84	2.93	0.33	1.79	2.13
R37	35.81	40.62	30.21	2.79	2.90	2.71	1.20	-2.62	-1.44
R38	21.28	14.93	32.29	2.38	2.10	2.76	-2.00	5.00	3.02
R39	28.04	36.11	39.58	2.62	2.84	2.89	2.09	0.86	2.96
R40	49.32	44.10	57.64	2.91	2.93	2.92	-1.27	3.28	2.02
R41	60.47	49.65	47.22	2.85	2.95	2.95	-2.64	-0.58	-3.23
R42	29.73	34.03	33.68	2.66	2.80	2.79	1.11	-0.09	1.02
R44	59.80	53.47	65.28	2.85	2.94	2.81	-1.54	2.90	1.37

Table 4.19 - cont'd.

Item	Percent Incorrect			Standard Error			Stand. Difference		
	1964	1970	1979	1964	1970	1979	64-70	70-79	64-79
C 1	3.38	9.37	17.01	1.05	1.72	2.22	2.97	2.72	5.55
C 2	10.81	9.03	14.93	1.81	1.69	2.10	-0.72	2.19	1.49
C 3	7.77	12.50	20.49	1.56	1.95	2.38	1.89	2.59	4.47
C 4	2.70	3.82	11.11	0.94	1.13	1.86	0.76	3.36	4.04
C 6	13.18	10.07	18.75	1.97	1.78	2.30	-1.17	2.98	1.84
C 7	14.53	16.67	28.47	2.05	2.20	2.66	0.71	3.42	4.15
C 8	9.46	12.50	22.92	1.70	1.95	2.48	1.17	3.30	4.47
C 9	13.18	18.75	28.82	1.97	2.30	2.67	1.84	2.85	4.74
C10	8.45	15.97	33.33	1.62	2.16	2.78	2.79	4.93	7.73
C11	16.55	31.94	48.26	2.16	2.75	2.95	4.40	4.05	8.67
C12	9.80	24.31	23.61	1.73	2.53	2.51	4.73	-0.19	4.53
C13	29.05	48.26	41.32	2.64	2.95	2.91	4.85	-1.68	3.12
C14	28.72	49.65	51.39	2.63	2.95	2.95	5.29	0.42	5.73
C15	14.19	18.40	32.64	2.03	2.29	2.77	1.38	3.96	5.37
C16	23.99	43.40	41.67	2.49	2.93	2.91	5.06	-0.42	4.62
C17	31.76	32.64	41.67	2.71	2.77	2.91	0.23	2.25	2.49
C18	23.99	46.18	44.79	2.49	2.94	2.94	5.76	-0.33	5.41
C19	28.38	46.53	55.56	2.62	2.94	2.93	4.60	2.17	6.90
C20	24.32	45.49	46.53	2.50	2.94	2.94	5.49	0.25	5.75
C21	9.80	9.37	21.87	1.73	1.72	2.44	-0.17	4.19	4.04
C22	31.42	44.79	63.89	2.70	2.94	2.84	3.35	4.68	8.29
C23	24.66	16.32	29.17	2.51	2.18	2.68	-2.51	3.72	1.23
C24	46.28	52.43	61.46	2.90	2.95	2.87	1.49	2.19	3.72
C25	25.34	23.96	29.86	2.53	2.52	2.70	-0.39	1.60	1.22
C26	42.91	48.26	56.60	2.88	2.95	2.93	1.30	2.01	3.33
C27	27.36	45.14	80.90	2.60	2.94	2.32	4.53	9.55	15.38
C28	29.39	43.40	75.69	2.65	2.93	2.53	3.55	8.35	12.63
C29	27.70	27.78	30.56	2.61	2.64	2.72	0.02	0.73	0.76
C30	16.22	17.36	25.00	2.15	2.24	2.56	0.37	2.25	2.63
C31	16.22	31.25	45.49	2.15	2.74	2.94	4.32	3.55	8.04
C32	49.66	59.03	55.90	2.91	2.90	2.93	2.28	-0.76	1.51
C34	61.15	78.47	75.69	2.84	2.43	2.53	4.64	-0.79	3.82
C35	22.64	30.90	47.92	2.44	2.73	2.95	2.26	4.24	6.61
C36	5.07	7.64	18.06	1.28	1.57	2.27	1.27	3.78	4.99
C37	63.51	37.15	70.83	2.80	2.85	2.68	-6.59	8.60	1.89
C38	65.88	75.00	77.43	2.76	2.56	2.47	2.42	0.68	3.12
C41	56.42	62.50	70.83	2.89	2.86	2.68	1.50	2.13	3.66
C42	48.65	45.14	58.68	2.91	2.94	2.91	-0.85	3.28	2.44
C44	84.80	90.97	90.62	2.09	1.69	1.72	2.30	-0.14	2.15

Table 4.20

Items Showing Discrepancies Between
Decisions Using the Rasch and Traditional Models

Item	Years	Decision		Item	Years	Decision	
		Rasch	Trad			Rasch	Trad
R 1	70-79	0	+	C 1	70-79	0	+
R 1	64-79	0	+	C 2	70-79	0	+
R 5	64-79	0	+	C 3	70-79	0	+
R 6	64-79	0	+	C 4	70-79	0	+
R 7	70-79	0	+	C 4	64-79	0	+
R 8	70-79	0	+	C 6	70-79	0	+
R10	70-79	0	+	C13	70-79	-	0
R11	64-70	0	+	C17	70-79	0	+
R18	64-70	0	+	C17	64-79	0	+
R20	64-70	0	+	C19	70-79	0	+
R28	64-79	-	0	C30	70-79	0	+
R36	64-79	0	+	C30	64-79	0	+
R37	64-79	-	0				
R38	64-79	0	+				
R40	64-79	0	+				

Table 4.21

Content Area Decisions Using the Rasch and Traditional Models

Content Area	1964-1970				1970-1979				1964-1979			
	Stand Diff		Dec		Stand Diff		Dec		Stand Diff		Dec	
	Rasch	Trad	R	T	Rasch	Trad	R	T	Rasch	Trad	R	T
1. WNC	0.00	0.76	0	0	1.58	1.60	0	0	1.82	2.36	0	+
2. WNA	1.00	1.43	0	0	1.55	2.25	0	+	2.61	3.71	+	+
3. CFC	1.36	1.59	0	0	2.24	2.65	+	+	3.75	4.27	+	+
4. CFA	1.23	1.73	0	0	3.25	3.55	+	+	4.36	5.28	+	+
5. Dec	1.62	2.26	0	+	0.73	1.09	0	0	2.40	3.33	+	+
6. Mon	0.67	0.88	0	0	1.29	1.81	0	0	1.93	2.67	0	+
7. Pct	2.47	2.44	+	+	1.71	1.98	0	0	4.19	4.49	+	+
8. Alg	-2.14	-1.86	-	0	3.85	4.11	+	+	1.77	2.27	0	+
9. Geo	1.32	1.36	0	0	0.43	1.02	0	0	1.82	2.33	0	+
10. Mea	2.56	2.43	+	+	6.10	6.03	+	+	8.29	8.90	+	+

CHAPTER V

DISCUSSION AND CONCLUSIONS

There are numerous contentious issues in the measurement of change. The specific problem to which this study was directed was the problem of scale. Because the Rasch model purports to yield measures of item difficulty and person ability on a common equal interval scale, this model was selected as the basis for the study. Once item difficulties were established using Rasch procedures, traditional statistical procedures utilizing standardized difference scores were used to assess change. A large part of the discussion in this chapter centres on how change decisions differ depending on whether the item difficulties used are those generated in the Rasch model, or traditional p-values.

The other major section of this chapter focusses on the changing achievement patterns in British Columbia as determined by the Rasch analysis. Some consideration is given to the problem of sampling variations for the three test administrations. The question of how declining performance on specific topics might be viewed by the educational community is partially resolved by reference to a previous study of achievement in British Columbia.

Comparison of the Rasch and Traditional Models

On the basis of the Rasch analysis, 29 of 79 items had changed in absolute difficulty from 1964 to 1970. In the preliminary traditional reanalysis in the initial phase of the study, 28 of the same 79 items were judged to have changed in difficulty over the same time span. The only item on which decisions differed was Item R38: easier in 1970 using the Rasch item difficulties, no change using the %-difficulties. This prompted the question of how decisions might differ, in general, depending on which model was used.

In the preliminary traditional reanalysis, the entire sample of 300 subjects had been used for each of 1964 and 1970. In order to make the results comparable with those obtained using the Rasch procedure, the %-difficulties were recalculated after deleting the same subjects as had been deleted for the Rasch analysis.

In 27 out of 237 comparisons, discrepancies were found to exist between the Rasch and traditional decisions, including three on the 1964-1970 comparison. In 24 of these cases, items were deemed to have changed in difficulty using the traditional model, while no change in difficulty was indicated using the Rasch model. Hence, the Rasch model appeared to be more conservative.

In attempting to determine the reason for differing interpretations, consideration was given to the mechanics for deciding change--the standardized difference. The formula for determining absolute change in difficulty was:

$$D = [\,d(1) - d(2)'\,] / \sqrt{se(1)^2 + se(2)^2} \qquad [1]$$

where d(2)' was the adjusted difficulty on the second administration. For the Rasch model to be more conservative, two things, or a combination of both, might have occurred:

(1) the separation between d(1) and d(2)' was relatively less in the Rasch model than in the traditional, or

(2) the pooled standard error was relatively greater in the Rasch model than in the traditional.

To determine the nature of the relationship between traditional %-difficulties and Rasch item difficulties, the graph shown in Figure 5.1 was drawn. The graph is a plot of the unadjusted Rasch item difficulties against %-difficulties for each of the three years. Inspection of the graph shows that, for each year, the transformation is very close to linear for items within approximately the 20% to 80% difficulty range. Beyond these limits the relationship is curvilinear, with the extreme Rasch scores becoming increasingly larger relative to the %-difficulty values. Hence, in equation [1], if d(1) had the larger absolute value of the two difficulty values, the difference between d(1) and d(2)' in Rasch units will be relatively as great as, or greater than, that in traditional units. This result alone should make the Rasch model less conservative than its alternative.

The second factor affecting the standardized difference is the standard error. Suppose a comparison is made of the difficulty of an item on two administrations, and an item moves toward the extremity of the curve on the second

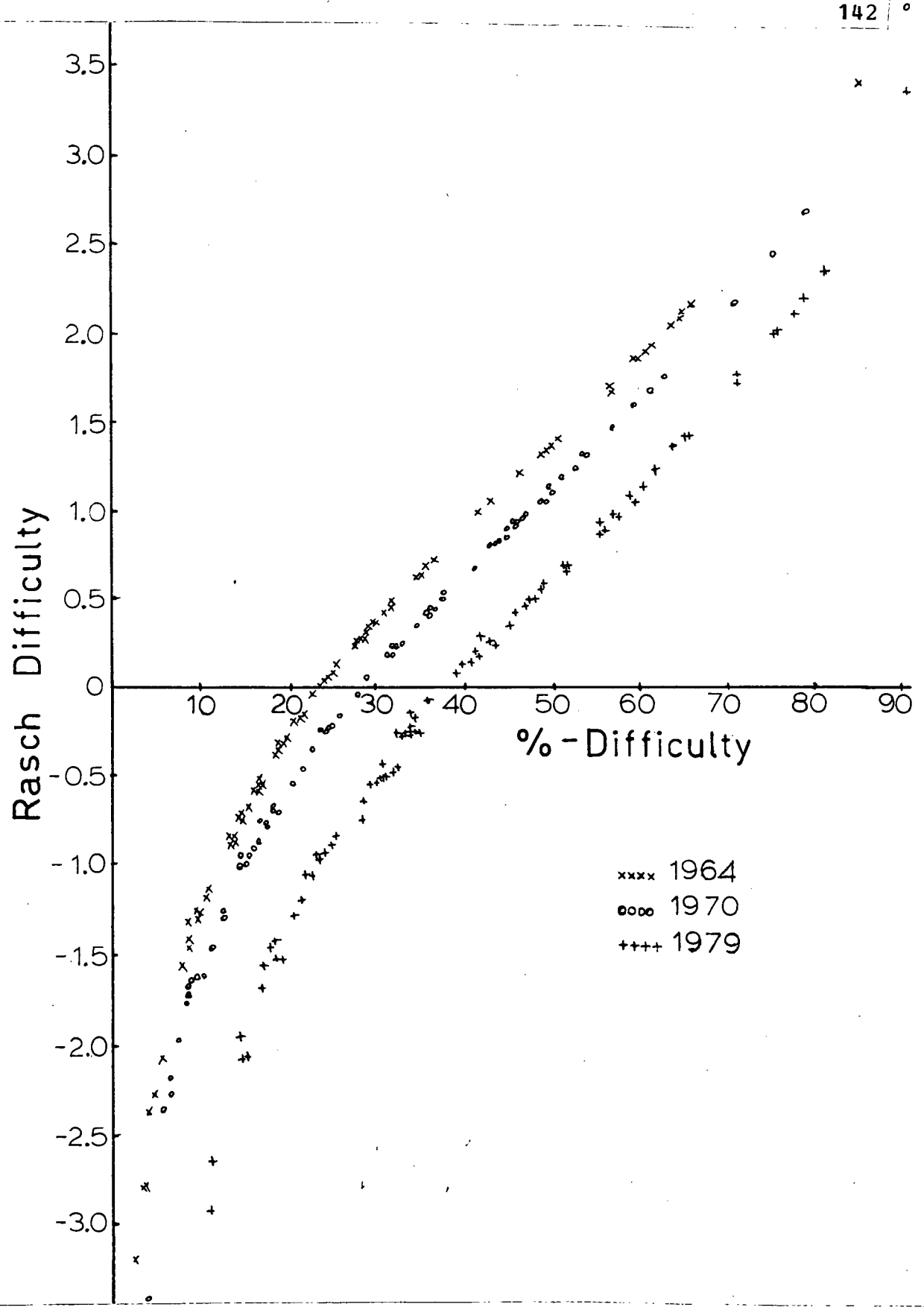


Figure 5.1. The relationship between %-difficulty and Rasch difficulty.

administration. As previously pointed cut, this has the effect of increasing the difference between the item difficulties as compared with the traditional model. But there is also the effect of the increasing standard error of the estimates toward the extremes. This will tend to reduce the value of D in equation [1]. The question then is: will the increase in the standard error be sufficiently large to offset the increase in difficulty value? To help answer this, the graph in Figure 5.2 was constructed. The graph portrays the relationship between item difficulty and its standard error on the data from the 1979 test administration. From Figure 5.1 it was noted that the transformation from %-difficulty to Rasch difficulty was basically linear in the interval from 20% to 80% difficulty. For the 1979 data that translated into the interval from -1.2 to 2.2 logits. For that same interval the relationship between Rasch difficulty and standard error is curvilinear, with increasess in standard error accelerating as the item difficulties move outward from the mean sample ability position. This tends to decrease the value of D in equation [1], making the Rasch test more conservative in this central region.

For the extreme regions it might be expected that the rapidly increasing standard error more than offsets the increased spread in item difficulty noted earlier, thereby maintaining the conservative nature of the model. However, a mathematical demonstration is required to resolve the issue, and that has not been done in this study.

Further consideration must be given to the standard

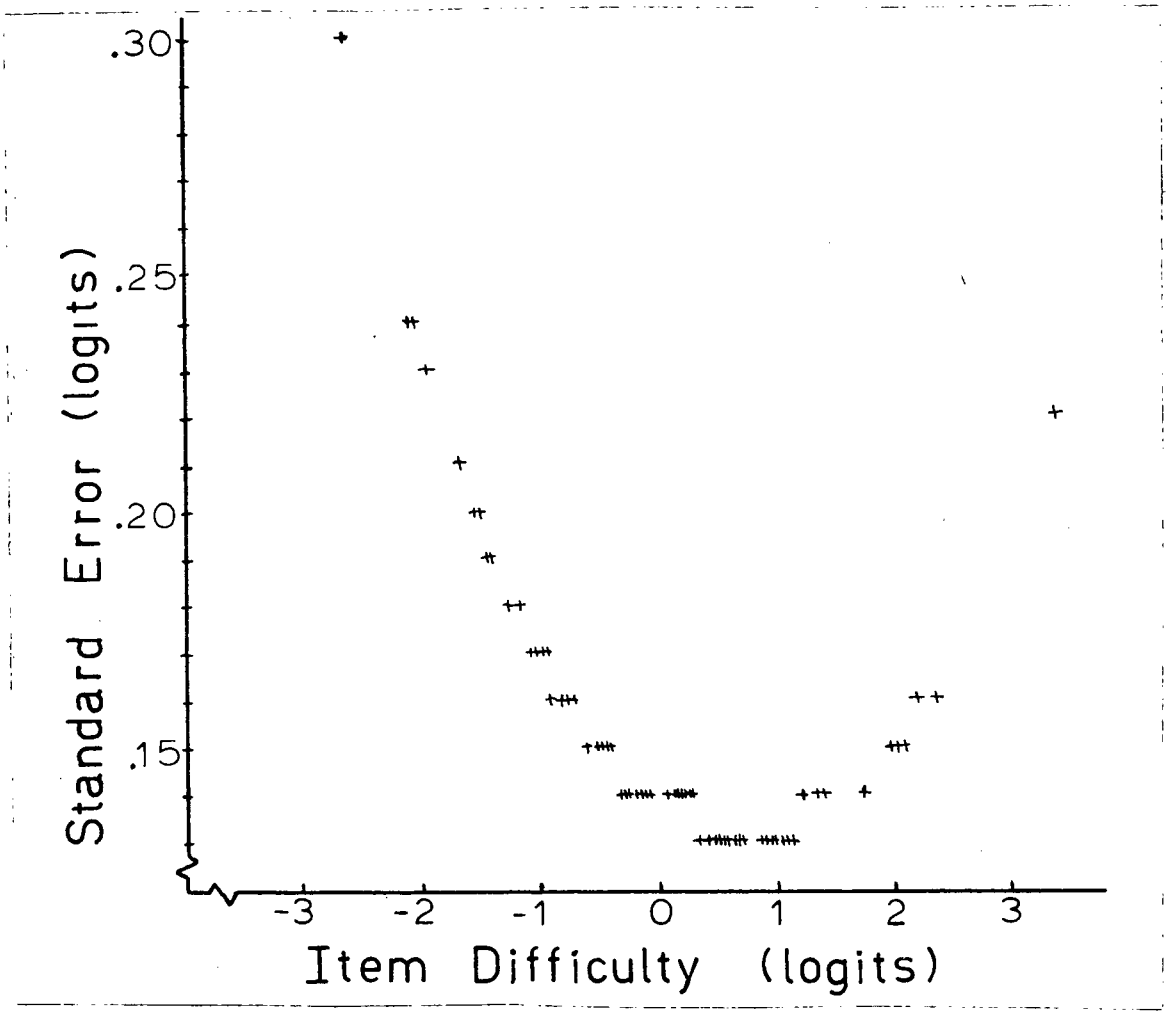


Figure 5.2. The relationship between Rasch item difficulty and standard error.

error. As pointed out in Chapter II, it is here that the Rasch model differs from the classical model. The classical standard error of estimate is maximum for items of 50% difficulty, and decreases to zero at either extremity. The Rasch standard error of item calibration is minimum for items centred at the mean ability level and increases toward the extremities.

Figure 5.3 portrays 79 items from a hypothetical test arranged in order from easiest to hardest. The upper portion of the figure shows the confidence band of $\pm$ 2 standard errors about the %-difficulty values. In the lower

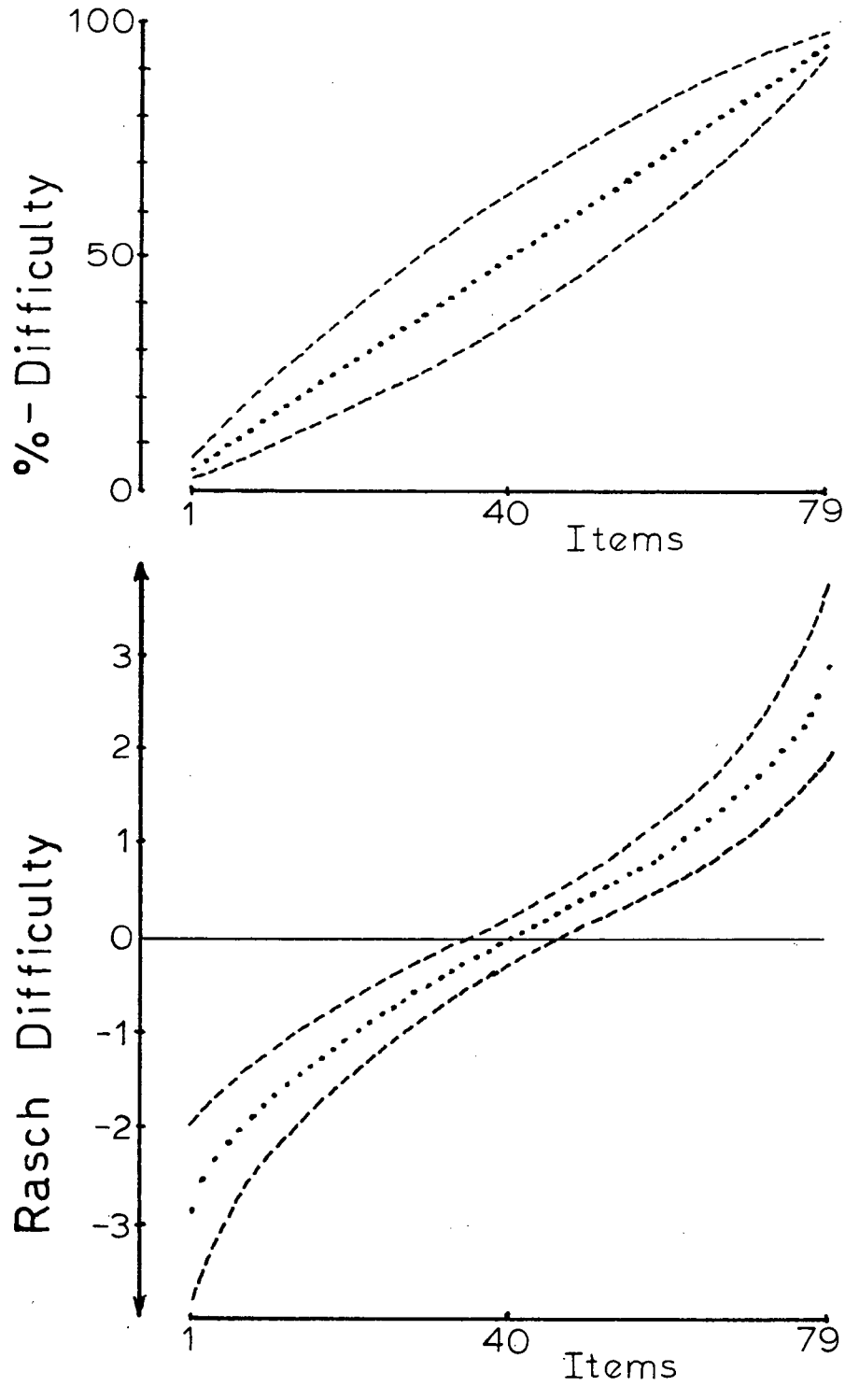


Figure 5.3. Variation in confidence bands within the traditional and Rasch models.

portion of Figure 5.3, the %-difficulty values have been transformed into approximate Rasch equivalents, assuming the idealized curve from Figure 5.1. The confidence band of ± 2 standard errors is shown as before. The scales have been exaggerated to make the differing effects clear.

The effect of the different behaviours of the standard errors is shown in Figure 5.4. A hypothetical case is given in the diagram in which it is assumed that the items have not changed in relative difficulty; they have all increased in difficulty by the same number of percentage points. In the upper portion of Figure 5.4, at time A the range in %-difficulty was 6 to 86 units with a mean of 46; on time B, the range was 14 to 94 with a mean of 54. The dashed lines are the limits of the region whose vertical height equals two standard errors of the difference of the difficulties for an item, assuming a sample size of 300. Because the standard error of estimate for the difficulty of an item decreases toward the extremes, so will the width of the envelope in Figure 5.4 decrease toward the extremes. In this example, the items having difficulties near the mean will not have changed in difficulty, whereas those on the extremes will. The crossover points are around the 47% and 57% difficulty levels at time A.

The results of transforming the %-difficulties for times A and B into Rasch difficulties are shown in the lower part of Figure 5.4. The separation between the curves is about 0.32 logits. The dashed lines again represent the envelope for two standard errors of the difference of the

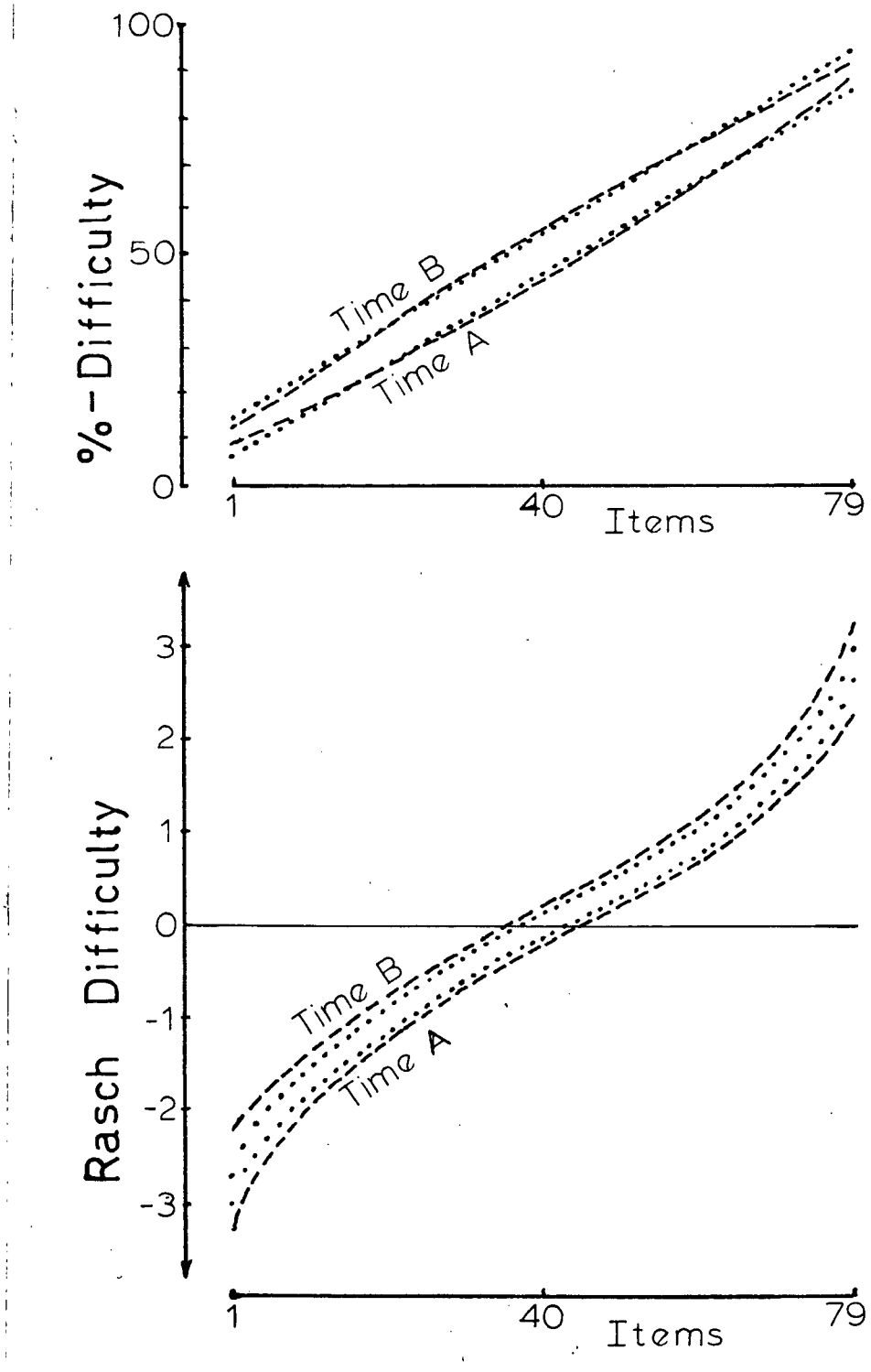


Figure 5.4. Effect of varying standard errors on decisions on changing item difficulty.

difficulties. Again the item of mean difficulty lies within the envelope, and the judgment of no change is made. In this case, however, the standard error of estimate of item difficulty increases toward the extremes, producing the same effect on the standard error of the difference of the difficulties. As a result, the envelope opens out toward the extremes. Since the vertical distance between the curves is constant, no item is concluded to have changed in difficulty. The net effect is to make the Rasch approach more conservative than the traditional approach.

If the argument based on the behaviour of standard errors shown in Figure 5.4 has merit, in this study conflicting decisions should have been concentrated on items having difficulties near the extreme ends of the distribution. Furthermore, since the items were generally easy in all three test administrations, with only about one-fifth of the 237 item difficulties exceeding 50% difficulty, it would be expected that the discrepancies in decision-making would occur predominantly for items at the easier end of the scale. This expectation was confirmed. The mean %-difficulty on the 24 items on which the Rasch model was conservative was about 23%, while the overall item difficulty on the tests was about 28%. That is, the conflicting decisions were made on items less difficult than average. For the three items on which the Rasch was less conservative, the mean %-difficulty was about 47%. It can only be conjectured that discrepancies here occurred through a combination of standard error variations and fluctuation in calibration exemplified in Figure 5.1.

The relatively conservative nature of the Rasch model in making change decisions at the item level was carried on into the decisions concerning change in topic difficulty. The standardized differences of mean group difficulties are almost invariably less in the Rasch approach than otherwise (see Table 4.21). This tendency led to conflicting decisions on two topics in the 1964-1970 comparison, one topic in the 1970-1979 comparison, and four topics in the 1964-1979 comparison. The difference is most dramatic in looking at change from 1964 to 1979. On the traditional comparison one would conclude that decline had occurred on all ten topics whereas the Rasch comparisons would yield more conservative results.

It would appear that the decision as to which model to use rests mainly on the user's view of which model most appropriately represents the confidence interval for items at the extremes of the difficulty range. For difficult items, the issue may be resolved in favour of the Rasch procedure by appealing to the argument that uncertainty increases as items become more likely candidates for guessing or random responses. For easy items, however, the situation is unclear. It has been suggested that boredom, or carelessness, may introduce uncertainty into the results on very easy items. This is certainly possible, but the suggestion does not have the same intuitive force as that for guessing on difficult items. The issue remains basically unresolved.

Change in Achievement in British Columbia

The present study was designed to explore the use of the Rasch model to measure change. Investigation has shown the Rasch model to be more conservative than the alternative traditional approach, each using the test item as the unit of analysis. That is of theoretical interest. The study was also intended to investigate and report upon real change which had taken place in the mathematics achievement of Grade 7 students in British Columbia. Principals and superintendents were promised a summary of the findings. That is of practical interest. Which change is "real"--Rasch or traditional?

Since the letter requesting the cooperation of school principals had indicated that a different kind of statistical model would be used to carry out the analysis, it was decided that a basic commitment to the Rasch model had been made. Consequently, all further discussion of results and trends are based on decisions reached using the Rasch model.

Sampling and Motivation Considerations

In contrast to 1964 and 1970, the data collection for 1979 relied on the voluntary cooperation of personnel in the field. Consequently, there was some concern that, without the persuasive force of authority, school districts would be reluctant to cooperate in a study which might not reflect well on the comparative achievement of students. These concerns proved to be unfounded, as almost 90% of the superintendents

agreed to allow the researcher to contact school principals. There still remained the question of how well schools would cooperate, but again the return rate of completed tests from schools was very high, over 95%. As a result, completed test papers were returned for 1277 students, approximately 86% of the design sample of 1500.

The loss of fourteen schools from the sample was not distributed equally across the six geographic regions. The greatest loss was from the Greater Vancouver region, with 15 out of 24 schools cooperating. The second greatest loss was from the North region, with 5 of 9 schools remaining in the sample. Regions 3 and 4 lost one school each; there were no losses from Regions 1 and 5. Most of the schools lost were located in larger school districts and situated in urban centres. While the loss of such schools may bias the sample, because of the wide variability of schools within each district, no firm conclusion can be made in this regard.

A second factor which may have a bearing on the validity of the year-to-year comparison is the changing sampling procedure. In 1964, the sample was drawn from the population stratified by performance on the entire Stanford Achievement Battery. In 1970, the selection was the same but performance was based on just the two arithmetic tests. In 1979, the criterion was performance on the two arithmetic tests, but the sample was constructed to reflect the geographic diversity and variation in school size in the province. Although the samples for 1964 and 1970 probably fairly represented the regions through the random selection

process, there is no guarantee, with a sample size of 300, that this was achieved. The same argument applies to school size. Nevertheless, these potential sample differences appeared to be overwhelmed by the magnitude and consistency of change over the years in this study. In another situation where evidence for change was less clear, the sampling variations might have been considered a more important factor.

A third complicating factor is that of student motivation. In 1964 and in 1970, the testing program was province-wide. In 1964 the results for each individual and each class were returned to the teacher. It is assumed that the pupils writing the tests were aware that this would be the case. It is not known whether results of the 1970 testing were forwarded to the classroom teacher, but the tests were administered under the authority of the Department. In both instances, students were under some pressure to perform as well as possible. That condition did not hold in 1979.

In 1979, the design of the study did not require the determination of class averages or summary statistics for districts. This, combined with the prevailing view regarding the need for confidentiality of test results, resulted in a decision not to return results to classroom teachers or principals. This may have reduced both the motivation of students to succeed on the test and that of teachers to ensure proper test conditions.

Change in Achievement by Content Area

With sampling and motivation reservations in mind, the results for each content area on each comparison may be examined. In the discussion, one of the topics, #10, Units of Measure, will be dealt with on its own at a later stage. Only the nine remaining topics are involved in each comparison to follow.

The comparison which prompted this study was that of 1964 to 1970. The results of the present study indicate that the Director's concerns for declining performance had some foundation. Standardized differences of mean topic difficulties show a general trend supporting the Director's conclusions. However, just one topic can be judged to have increased significantly in difficulty. On Topic #7, Percent, three of seven items (C14, C19, C35) increased in absolute difficulty, with two of those three (C14, C19) also increasing in relative difficulty. Changes on the latter two items were quite large: on each item, in 1964 about one-quarter of the examinees responded incorrectly; in 1970, that proportion increased to one-half. The questions were straight-forward, for example, C14: 30% of $40 = ? . It is difficult to understand how these items were so much more difficult while performance on word problems requiring the same computation was not significantly reduced.

The sole topic which prevailed against the general declining trend was Elementary Algebra, on which performance improved significantly from 1964 to 1970. Five out of seven items improved relatively (R38, R40, C23, C37, C42) and three

of these showed absolute improvement (R38, C23, C37). The characteristic common to the five items showing relative improvement was the inclusion of a variable. The results would seem to indicate a relatively greater emphasis within the curriculum on algebraic usage, and a consequent improvement in student performance on this topic.

From 1970 to 1979, three topics increased significantly in difficulty. Once again, the fact that all numerical changes, significant or not, were in the direction of increasing difficulty indicates an overall trend. The improvement in Elementary Algebra from 1964 to 1970 was lost from 1970 to 1979. Although the topic did not receive relatively less emphasis, performance declined on six of the seven component items.

Students' understanding and ability to manipulate common fractions decreased from 1970 to 1979. On Topic #3, Common Fraction Concepts and Operations, 9 out of 12 items were more difficult in 1979. On Topic #4, Applications Using Common Fractions, 7 out of 8 items showed increasing difficulty.

On the absolute comparison from 1964 to 1979, five topics increased in difficulty. In addition to the three topics previously discussed, the persistent trend of increasing difficulty resulted in a significant decline in performance on Topic #2, Applications of Whole Numbers, and Topic #5, Decimals. The proportion of items in each group which were more difficult in the latter year were as follows: Applications of Whole Numbers, 4 of 8; Common Fraction

Concepts and Operations, 9 of 12; Applications Using Common Fractions, 8 of 8; Decimals, 6 of 11; Percent, 6 of 7. The decline in performance in Elementary Algebra from 1970 to 1979 offset the improvement from 1964 to 1970, leaving a net effect of no change.

A discussion of change on Topic #10, Units of Measure, has been postponed until now because it serves to illustrate a fundamental problem with the test administration in 1979. In 1970, the Canadian government announced plans to convert from the imperial system of weights and measures to the metric system. The changeover was to be completed by 1981. In September, 1973, all pupils at the primary level were to begin using the metric system. The Council of Ministers of Education, Canada, agreed that instruction in Canadian public schools should be predominantly metric by 1978. The Metric Commission of Canada recommended that the changeover to the metric system be done with a minimum of conversion from imperial to metric units. This policy was to be followed in the schools where students were to be encouraged to "think metric" through immersion in the metric system.

Of eight unsolicited letters from teachers and principals who administered the tests in 1979, seven pointed out the difficulty of using an old test to assess students who were used to metric measures. The obvious problem lies with the units themselves, but one correspondent also suggested that the use of commas instead of spaces in larger numbers was a source of confusion on three items.

The metric problem had been recognized prior to sending out the 1979 tests. However, regardless of the directives of the Ministry of Education on metrication, it was unclear what was actually being done in the schools. For example, in the provincial mathematics assessment of 1977, Robitaille and Sherrill suggested that the majority of elementary teachers in the province were still using both metric and imperial units of measure in their teaching. It was also conjectured by the researcher that the emphasis on a new measurement system might have resulted in teachers looking at the inadequacies of the old system, thereby evoking an awareness of imperial units. The results from the 1979 tests appear to support the concerns raised by the teachers.

Fifteen items out of 89 on the tests involved the use of imperial units. Eleven of these increased in difficulty from 1970 to 1979. However, seven had increased in difficulty from 1964 to 1970, suggesting that more than a problem of units was involved. The final comparison shows that 14 had become more difficult from 1964 to 1979, although, oddly enough, the item requiring the reading of a gas meter in cubic feet had become easier.

Of the fifteen items, only two required a knowledge of a base other than ten. These were items involving addition and subtraction of two or more quantities in units of pounds and ounces. These two items had been placed under Topic #10 along with Item R24, requiring subtraction of hours and minutes, and Item C33, requiring the addition of metres and centimetres. The latter was the only metric item on the two

tests, and it was eliminated in the calibration process, leaving just three items in the topic. Although the topic showed relative and absolute increase in difficulty in all but one comparison, it was considered to be too restricted in content to permit further generalization.

On thirteen of the fifteen items using imperial measure, the unit was incidental to the computation process. For example, Item R16: "At 8 miles an hour, how many miles can a skater go in 4 1/2 hours?" While it is not obvious that the use of unfamiliar units invalidates the question, the effect may be to detract from the students' performance. This was a factor which particularly affected conclusions reached regarding Topic #4, Applications Using Common Fractions, for which all but one of its 8 items made use of imperial units. The problem also occurs, to a lesser extent, on Topic #2, Applications Using Whole Numbers, where 4 of 8 items involved imperial measure. The problem does not arise on any other topic.

Before outlining possible reasons for changing performance, several aspects of the tests and topics should be reviewed and clarified. In the first place, both the number and nature of the topics into which the items were grouped were arbitrarily determined. The initial classification of the items was that which seemed most natural to the researcher. Another investigator may well have reduced the number of topics or reassigned the items. Secondly, there was no choice in the scope and depth of the items assigned to each category. The content of each category was determined solely

by the items available from the test. There is no assurance that each topic is adequately represented by the items subsumed by it. Therefore, when performance on a topic is referred to, it must be thought of achievement on this topic as defined by these items from this test. The temptation to generalize beyond the data must be resisted.

It was not the intent of the study to attempt to identify, in any systematic way, correlates of change. No demographic data such as age, sex, socio-economic status, or language spoken at home, were gathered. The nature of changing population characteristics and their relationship to performance in school is complex. No doubt, societal factors such as increasing urbanization, television, increased permissiveness, drug usage, and marital breakdown all contribute to change in achievement in school, but no attempt has been made here to assess their influence.

When the factors which possibly influence change are narrowed down to the school and the subject within the curriculum, hypotheses can be formed with somewhat greater confidence. The first factor which may help to explain the consistent decline in performance on the test as a whole from 1964 to 1979 is the changing curriculum. As pointed out previously, the tests were based on the American curriculum of the late 1940's, and there is a high probability that the tests were also appropriate for the curriculum of British Columbia elementary schools in 1964. Canadian and American mathematics curricula have generally been comparable at any given time. In British Columbia, a series of textbooks

entitled Study Arithmetics had been used as supplementary texts for Grades 3 to 6 in the early 1940's. It was adopted as the authorized textbook series, and hence course of study, in 1947. It continued as the sole authorized textbook series until the Seeing Through Arithmetic series was phased in between 1962 and 1966. At the Grade 7 level, Mathematics for Canadians, 7 was used from the mid-1950's to the mid-1960's. It replaced Junior Mathematics, Book 1, which had been used since the early 1940's. Therefore, in general, there was a stability of curriculum lasting over twenty years on which teachers could base their teaching and testing. The Stanford Achievement Tests likely were broadly representative of that curriculum.

A revised elementary mathematics program was phased in during the years from 1962 to 1967. By 1970, all Grade 7 students had had seven years of the new program. That program expanded the elementary mathematics curriculum to include "modern" topics such as the terminology of sets, the number line, the properties of number systems, and numeration systems with bases other than ten. As well, the program included informal geometry as a fundamental component for each grade.

The large scale curriculum development projects of the 1950's and 1960's were replaced in the 1970's by an emphasis on local curriculum improvement. This development, together with a change in the provincial government, resulted in the decentralizion of the curriculum from 1973 to 1976. The Seeing Through Arithmetic series was replaced by a

multiple authorization of three different series. Teachers were expected to follow the Department's curriculum guide and to use the best parts of each series to provide the optimum program for their students. The teaching of metric units became a priority at all levels. At the Grade 7 level the concepts of functions and flow-charting were introduced.

Thus, it can be seen that in 1964 the tests used in this study likely sampled the totality of the students' mathematical knowledge, whereas by 1979 the students had been exposed to a much broader curriculum than that measured by the tests. Two points can be made. In the first place, the test results do not reflect the child's body of mathematical knowledge in 1979. Performance on topics common to previous years may be poorer, but the child's knowledge is likely to be broader.

Secondly, because of an expanded curriculum, one might expect performance on a portion of the curriculum to decline if the overall time devoted to learning mathematics in the elementary school remained the same. That is, if the time on a specific task decreased, lower performance might be expected. Table 5.1 shows how the time allotted to the study of arithmetic changed from 1958 (British Columbia Department of Education, 1957) to 1972 (British Columbia Department of Education, 1972).

It can be seen that, for the last four years of the elementary program, on which the bulk of the test items in this study were based, there has been little change in the time allotment. Consequently, if time on task is a

Table 5.1

Time Allotted to the Study of Arithmetic*

	Grade						
Year	1	2	3	4	5	6	7
1958	100	100	150	200	200	200	240
1972	150-160	150-200	200-210	200-220	200-220	200-220	200-240

* in minutes per week

significant factor, the observed decline in achievement should not be unexpected.

Another factor which might be expected to influence achievement is teaching method. If the proponents of such innovations as open area schools, discovery learning, and team teaching are correct, improvement in achievement should result from the adoption of any or all of the above. It appears that this question must remain unresolved for, in spite of the rhetoric of reformers and the exhortations of curriculum guide writers, instructional practices seem to have remained largely unchanged. An extensive survey of British Columbia mathematics teachers in 1977 (Robitaille & Sherrill, 1977a) concluded that:

> the teacher of mathematics is highly traditional in character....the most frequently used teaching techniques are total class instruction and teacher explanation. Among the most commonly used student activities are individual work and textbook exercises....these results indicate that few organizational innovations are being used in the mathematics classes of the province. (p. 44)

Turning to specific changes by topic, the

improvement in Elementary Algebra from 1964 to 1970 is explainable by the adoption of the modern mathematics program in the mid 1960's. That program placed much emphasis on the solution of open sentences. Thus performance became better relative to other topics, and the absolute achievement of students increased.

The increasing difficulty of questions on Percent from 1964 to 1970 might be attributed, in part, to the changed method for solving such problems. The previous program had placed reliance on rules, whereas the new program tried to apply a single structure using rate pairs, proportions, and the solution of equations. It is possible that both teachers and students found the new procedure more difficult.

From 1970 to 1979, the increasing difficulty with respect to Operations on, and Applications of, Common Fractions may be due to the introduction of the metric system in the mid 1970's. This may have served to remove some of the emphasis on common fractions, since, with metric units, the necessity for dealing with thirds, twelfths, sixteenths, and the like, is reduced.

There appears to be no identifiable reason for the declining performance on Elementary Algebra from 1970 to 1979, and on Applications of Whole Numbers, and Decimals from 1964 to 1979. The latter is particularly puzzling, as it might be thought that the introduction of the metric system would lead to increased facility in the use of decimal fractions.

A criticism that might be made of a study of this sort is that comparisons with past years do not help to assess

performance relative to the expectations of contemporary society. That is, performance may have declined over previous years, but may still be acceptable in its own time, or the opposite may hold. To gain some insight into how the 1979 performance might be viewed, an analysis was made of a previous study in British Columbia.

In the spring of 1977, the Learning Assessment Branch of the British Columbia Ministry of Education administered provincially developed mathematics tests to all students in the province enrolled in Grades 4, 8 and 12. The tests were constructed to measure minimum basic skills which the student might be expected to possess at each grade level. For each grade, results on each item were judged by a fifteen-member interpretation panel consisting of seven mathematics teachers at that grade level, two supervisors of instruction, two teacher educators, two school trustees, and two members of the public at large. The panel rated performance on each item on a five point scale indicating their satisfaction with the results, as follows:

5 - strength

4 - very satisfactory

3 - satisfactory

2 - marginally satisfactory

1 - weakness

On the assumption that the interpretation of the performance of Grade 8 students in 1977 would not differ substantially from that of Grade 7 students in 1979, this researcher assigned 59 out of the 60 items on the Grade 8 test

to the ten content categories used in the present study. Number 10, Units of Measure, became Units of Metric Measure. The rating for each item, obtained from Robitaille and Sherrill (1977b), ranged from 1 to 5, and the mean rating for the topic was calculated, with the results shown in Table 5.2.

Table 5.2

Mean Satisfaction Ratings of Content Areas on the 1977 Grade 8 Assessment

Content Area	No of Items	Mean
1. WNC	12	3.5
2. WNA	2	4.5
3. CFC	10	2.6
4. CFA	0	---
5. Dec	8	2.75
6. Mon	2	3.0
7. Pct	3	2.0
8. Alg	3	2.33
9. Geo	14	2.36
10. Mea(metric)	5	3.6

Little reliance can be placed on ratings of content areas which contain few items. For inclusion in the discussion, the number of items required in a given content area was arbitrarily set at 5. As a result, five topics may be considered: WNC, CFC, Dec, Geo, and Mea(metric). On three of these five topics, performance was judged to be less than satisfactory, that is, their mean satisfaction rating was less than 3.0. One topic showing weakness was Geometry, but the changing role of geometry in the curriculum since 1964 prevents any general conclusions to be made. On the other two

topics--Common Fractions Concepts and Operations, and Decimals--performance at the Grade 7 level had declined from 1964. The combination of the results from the two studies indicates the need for a serious reappraisal of the effectiveness of instruction on these two topics.

Limitations of the Study

The limitations of the study derive from three sources: the model itself, the nature of the data, and the purpose of the study.

Rentz and Bashaw (1977) identify two differing applications of the Rasch model: test construction and test analysis. In the former case the test maker can use the Rasch model as a guide, with the freedom to select the best set of items which fit the model. In the latter case the collection of test items is virtually fixed; it becomes necessary to rely on the robustness of the model to accommodate less than ideal items. This study falls into the second category.

In 1970 some teachers raised objections that the Stanford Achievement Tests were not based on the British Columbia modern mathematics curriculum (British Columbia Department of Education, 1971). In fact, as has been noted, the test was based on the American curriculum of the late 1940's. Thus, the assessment of change was based upon those elements of the curriculum of the past fifteen years common to that of thirty years ago. Hence the study is limited in the scope of the curriculum with which it deals. It does not attempt to assess change in the overall mathematics program.

Finally, the primary purpose of the study was to document and measure change. Although some suggestions are made concerning the reasons for change it was not the intent to undertake an investigation into the correlation between other factors and change in mathematical performance. Any identified changes, of course, cannot be generalized beyond the boundaries of the province of British Columbia.

Some Concerns and Suggestions for Future Research

The most suitable criterion to be used as a measure of item fit in the Rasch model is still an open question. The use of the fit mean square has recently been criticized by George (1979), who maintains that the use of the statistic does not detect unacceptable variations in discrimination. In a series of BICAL calibrations on items from an English achievement test, George concluded that dissimilarities in item discrimination could produce discrepancies in difficulty estimates based on high and low ability groups. While the implications of George's study are most serious for applications of the Rasch model such as test linking, and vertical equating, there is also reason for concern in comparing the performance of two groups of differing abilities on the same test, as was the case in the present study. In view of the controversy in the literature concerning the problem of fit, this difficulty will likely remain for some time to come.

A second matter which requires further investigation is the nature of the mathematical relationship between item difficulty and standard error. Hashway (1977) conjectured that the standard error may be a hyperbolic tangent function of the parameter. It has been pointed out that the basic element responsible for the sometimes conflicting decisions reached in the Rasch and traditional models was the differential behaviour of the standard error. The relationship needs to be clarified in order to permit a thorough analysis of the differing results when the two approaches are used.

Difficulty with standard error goes beyond the mathematical. There is a further question of which model best reflects the reality of testing. An argument has been made for preferring the Rasch model for difficult items, but the case at the other end of the difficulty scale is not clear. A well thought out rationale needs to be developed in order to decide which is preferable overall.

Finally, if the Rasch model is to be used for future comparisons of performance, procedures based upon the item as the unit of analysis are recommended over those using the person. In general, the standard error of estimate of the difficulty (ability) parameter is inversely proportional to the square root of the reciprocal of the number of persons (items). Since the number of examinees is larger than the number of items, estimates of item difficulties are more precise than estimates of person abilities. Furthermore, because the number of examinees can be increased without

limit, the distribution of the item difficulties is nearly continuous, while that of the abilities is discrete, having the same number of steps as possible raw scores. Hence, changes in item difficulties are more likely to be detected than changes in abilities. If the test is divided into content areas containing a few items in each, the distinction becomes even more important.

REFERENCES

Ahmann, J. S., & Glock, M. D. Evaluating Pupil Growth: Principles of Tests and Measurements (4th Ed.) Boston: Allyn and Bacon, 1971.

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Armbruster, F. E. The more we spend, the less children learn. The New York Times Magazine, August 28, 1977.

Austin, G. R., & Prevost, P. Longitudinal evaluation of mathematical computational abilities of New Hampshire's eighth and tenth graders, 1963-1967. Journal for Research in Mathematics Education, 1972, 3, 59-64.

Beckmann, M. W. Basic competencies -- twenty-five years ago, ten years ago, and now. The Mathematics Teacher, 1978, 71, 102-106.

Beltzner, K. P., Ccleman, A. J., & Edwards, G. D. Mathematical Sciences in Canada. Ottawa: Printing and Publishing Supply and Services Canada, 1976.

British Columbia Department of Education. Programme of Studies for the Elementary Schools of British Columbia. Victoria, B. C.: Department of Education, Divisicn of Curriculum, 1957.

British Columbia Department of Education. A Report on the Testing of Arithmetic. Victoria, B.C.: Department of Education, Research and Standards Branch, 1970.

British Columbia Department of Education. Ninety-ninth Annual Public Schools Report, 1969-70. Victoria, B.C.: Department of Education, 1971.

British Columbia Department of Education. Instructional services circular #761: elementary schccl--time allotment guidelines. Victoria, B. C.: Department of Education, September 9, 1972.

Carpenter, T. P., Coburn, T. G., Reys, R. E., & Wilson, J. Results and implications of the NAEP mathematics assessment: secondary school. The Mathematics Teacher, 1975, 68, 453-470.

Cartledge, C. McC. A comparison of equipercentile and Rasch equating methodologies (Doctoral dissertation, Northwestern University, 1976). Dissertation Abstracts International, 1976, 37, 5141A. (University Microfilms No. 76-2215)

Choppin, B. H. Item bank using sample-free calibration. Nature, 1968, 219, 870-872.

Choppin, B. H. Recent developments in item banking: a review. In N. M. De Gruijter & L. J. Th. van der Kamp (Eds.) Advances in Psychological and Educational Measurement. New York: Wiley, 1976.

Clarke, S. C. T., Nyberg, V., & Worth, W. H. General Report on Edmonton Grade III Achievement: 1956 - 1977 Comparisons. Edmonton, Alta.: Alberta Education, 1977.

Conway, C. B. Grade VII Aptitude-Achievement Survey, March 9th - 13th, 1964. Victoria, B.C.: Department of Education, Division of Tests and Standards, 1964.

Dinero, T. E., & Haertel, E. A computer simulation investigating the applicability of the Rasch model with varying item characteristics. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1976. (ERIC Document Reproduction Service No. ED 120240)

Fischer, G. H. Some probabilistic models for measuring change. In N. M. de Gruijter & L. J. Th. van der Kamp (Eds.) Advances in Psychological and Educational Measurement. New York: John Wiley and Sons, 1976.

Forbes, D. W. The use of Rasch logistic scaling procedures in the development of short multi-level arithmetic achievement tests for public school measurement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976. (ERIC Document Reproduction Service No. ED 128400)

Forster, F., Ingebo, G., & Wolmut, P. Can Rasch item levels be determined without random sampling? Monograph V, Vol. I. Portland, Ore: Northwest Evaluation Assocation, undated(a)

Forster, F., Ingebo, G., & Wolmut, P. What is the smallest sample size needed for field testing? Monograph II, Vol. I. Portland, Ore: Northwest Evaluation Association, undated(b)

Forsyth, R. A., & Feldt, L. S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. Educational and Psychological Measurement, 1969, 29, 61-71.

Fryman, J. G. Application of the Rasch simple logistic model to a mathematics placement examination (Doctoral dissertation, University of Kentucky, 1976). Dissertation Abstracts International, 1976, 37, 5626A. (University Microfilms No. 77-5689)

George, A. A. Theoretical and practical consequences of the use of standardized residuals as Rasch model fit statistics. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Glass, G. V, & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970

Hambleton, R. K. An empirical investigation of the Rasch test theory model (Doctoral dissertation, University of Toronto, 1969). Dissertation Abstracts International, 1971, 32, 4035A. (Microfilm available through the National Library of Canada, Ottawa)

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-95.

Hammons, D. W. Student achievement in selected areas of arithmetic during transition from traditional to modern mathematics (1960-1969) (Doctoral dissertation, The Louisiana State University and Agricultural and Mechanical College, 1972). Dissertation Abstracts International, 1972, 33, 2237A. (University Microfilms No. 72-28,349)

Hashway, R. M. A comparison of tests derived using Rasch and traditional psychometric paradigms (Doctoral dissertation, Boston College, 1977). Dissertation Abstracts International, 1977, 38, 744A. (University Microfilms No. 77-17,594)

Hedges, H. G. Achievement in Basic Skills: A Longitudinal Evaluation of Pupil Achievement in Language Arts and Mathematics. Toronto: Minister of Education, 1977.

Hungerman, A. D. 1965-1975: Achievement and Analysis of Computation Skills Ten Years Later. Ann Arbor, Mich.: The University of Michigan, 1975. (ERIC Document Reproduction Service No. ED 128202)

Hungerman, A. D. 1965-1975: Achievement and Analysis of Computation Skills Ten Years Later (Part II). 1977. (ERIC Document Reproduction Service No. ED 144839)

Kelley, T. L., Madden, R., Gardner, E. F., Terman, L. M., & Ruch, G. M. Stanford Achievement Test, Intermediate and Advanced Partial Batteries, Forms J, K, L, M, and N: Directions for Administering. New York: World Book Co., 1953.

Kerlinger, F. N. Foundations of Behavioral Research. New York: Holt, Rinehart and Winston, 1964.

Kifer, E., & Bramble, W. The calibration of a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974. (ERIC Document Reproduction Service No. ED 091434)

Kline, M. Why Johnny Can't Add: The Failure of the New Math. New York: St. Martin's Press, 1973.

Larson, R., Martin, W., Searls, D., Sherman, S., Rogers, T., & Wright, D. A look at the analysis of National Assessment data. In W. E. Coffman (Ed.) Frontiers of Educational Measurement and Information Systems--1973. Boston: Houghton Mifflin, 1973.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. Anchor Test Study Final Report. Project Report and Volumes 1 through 30. Berkeley, Cal.: Educational Testing Service, 1974. (ERIC Document Reproduction Service Nc. ED 092601 to ED 092631)

Maffei, A. C. Causes of recent decline in the mathematics achievement of public high school students: a national survey of high school mathematics teachers (Doctoral dissertation, University of South Carolina, 1977). Dissertation Abstracts International, 1977, 38, 2030A. (University Microfilms No. 77-22,420)

Magnusson, D. Test Theory. Reading, Mass.: Addison-Wesley, 1967.

McDonald, A. S. The Nova Scotia standards project. In Then and Now: comparisons of school achievement over time. Symposium presented at the annual meeting of the Canadian Educational Researchers' Association, London, Ontario, June 1, 1978.

Merrifield, P., & Hummel-Rossi, B. Redundancy in the Stanford Achievement Test. Educational and Psychological Measurement, 1976, 36, 997-1001.

National Advisory Committee on Mathematical Education. Overview and Analysis of School Mathematics, Grades K-12. Reston, Va.: National Council of Teachers of Mathematics, 1975.

National Assessment of Educational Progress. Math Fundamentals: Selected Results from the First National Assessment of Mathematics. Denver: Author, 1975.

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Passmore, D. L. An application of the Rasch one parameter logistic measurement model to the National League for Nursing Achievement Test in Normal Nutrition (Doctoral dissertation, University of Minnesota, 1974). Dissertation Abstracts International, 1974, 35, 963A. (University Microfilms No. 74-17,271)

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966a, 19, 49-57.

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.) Readings in Mathematical Social Science. Chicago: Science Research Associates, 1966b.

Rentz, R. R., & Bashaw, W. L. The national reference scale for reading: an application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work? A discussion for practitioners. Princeton: ERIC Clearinghouse on Tests, Measurement, and Evaluation, E. T. S., 1978. Also reprinted in Measurement in Education, National Council on Measurement in Education, 10, 2, 1979.

Robitaille, D. F., & Sherrill, J. M. The British Columbia Mathematics Assessment: Summary Report. Victoria, B.C.: Ministry of Education, Learning Assessment Branch, 1977a.

Robitaille, D. F., & Sherrill, J. M. The British Columbia Mathematics Assessment: Test Results. Victoria, B. C.: Ministry of Education, Learning Assessment Branch, 1977b.

Roderick, S. A. A comparative study of mathematics achievement by sixth graders and eighth graders, 1936 to 1973 (Doctoral dissertation, The University of Iowa, 1973). Dissertation Abstracts International, 1974, 35, 5601A-5602A. (University Microfilms No. 74-7423)

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

Rudman, H. C. The standardized test flap. Phi Delta Kappan, November 1977, 59, 179-185.

Russell, H. H., Robinson, F. G., Wolfe, C., & Dimond, C. Current Ontario Elementary School Mathematics Programs. Toronto: Minister of Education, 1975.

Slinde, J. A., & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Soriyan, M. A. Measurement of the goodness of fit of Rasch's probabilistic model of item analysis to objective achievement tests of the West African Certificate Examination (Doctoral dissertation, University of Pittsburgh, 1971). Dissertation Abstracts International, 1971, 32, 4433A. (University Microfilms No. 72-7895)

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.) Handbook of Experimental Psychology. New York: John Wiley and Sons, 1951.

Tinsley, H. E. A. An investigation of the Rasch simple logistic model for tests of intelligence or attainment (Doctoral dissertation, University of Minnesota, 1971). Dissertation Abstracts International, 1971, 32, 6629B. (University Microfilms No. 72-14,387)

Tinsley, H. E., & Dawis, R. V. An investigation of the Rasch simple logistic model: sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339. (ERIC Document Reproduction Service No. ED 069786)

Virgin, A. E., & Darby, L. M. 1974 Replication and Follow-up of a Survey of Mathematics and Reading Skills. Willowdale, Ont.: North York Board of Education, 1974. (ERIC Document Reproduction Service No. ED 128465)

Virgin, A. E., & Rowan, M. 1975 Replication of a Survey of Mathematics and Reading Skills. Willowdale, Ont.: North York Board of Education, 1975. (ERIC Document Reproduction Service No. ED 128466)

Waller, M. I. Estimating parameters in the Rasch model: removing the effects of random guessing. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975. (ERIC Document Reproduction Service No. ED 120261)

Whitely, S. E. Models, meanings and misunderstandings: some issues in applying Rasch's theory. Journal of Educational Measurement, 1977, 14, 227-235.

Whitely, S. E., & Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Whitely, S. E., & Dawis, R. V. The influence of test content on item difficulty. Educational and Psychological Measurement, 1976, 36, 329-337.

Willmott, A. S., & Fowles, D. E. The Objective Interpretation of Test Performance: The Rasch Model Applied. Windsor, Berks.: National Foundation for Educational Research, 1974.

Winne, P. Review of Achievement in Basic Skills: A Longitudinal Evaluation of Pupil Achievement in Language Arts and Mathematics by H. G. Hedges. Canadian Journal of Education, 1979, 4, 82-85.

Wright, B. D. Sample-free test calibration and person measurement. Paper presented at an Invitational Conference on Measurement, October 28, 1967. Princeton, N.J.: Educational Testing Service.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977a, 14, 97-116.

Wright, B. D. Misunderstanding the Rasch model. Journal of Educational Measurement, 1977b, 14, 219-225.

Wright, B. D. The Rasch model. Presentation at a Seminar on Item Calibration & Applications, Ontario Institute for Studies in Education, Toronto, July, 19, 1978.

Wright, B. D., & Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement, 1977a, 37, 47-60.

Wright, B. D., & Douglas, G. A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977b, 1, 281-295.

Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23A, Statistical Laboratory, Department of Education, University of Chicago, 1978.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

APPENDIX A

BRITISH COLUMBIA REPORT ON THE TESTING OF ARITHMETIC, GRADE VII, MARCH 1964 AND MAY 1970

OFFICE OF THE DIRECTOR
RESEARCH AND STANDARDS BRANCH
PARLIAMENT BUILDINGS

PROVINCE OF
BRITISH COLUMBIA

DEPARTMENT OF EDUCATION
VICTORIA, BRITISH COLUMBIA

Report on the Testing of Arithmetic

Grade VII, March 1964 and May 1970

Stanford Arithmetic Test, Advanced, Form L

The Stanford Arithmetic test was administered as part of a battery to 29,204 B.C. Grade VII students out of 29,533 enrolled in March, 1964. It consisted of 45 Reasoning and 44 Computation items. The sub-tests overlap to a certain extent: almost all 'reasoning' items require some skill in computation and many of the 'computation' items require a certain amount of problem-solving ability. Both sub-tests consist chiefly of applications to every-day life: the purchase of gasoline, renting a boat, reading a map, combining fractions, finding percentages.

The same test was reprinted and readministered in May, 1970. In the meantime, Grade VII enrolment had increased to 40,252 of whom 38,377 or 95% were tested. Of these, almost every pupil, including migrants from other provinces, would have had at least 6 years of "modern maths".

The purpose of the second administration was to determine the changes that had occurred in achievement in the ordinary arithmetic type of item. There had been rumours that while children were learning modern maths well in advance of their parents they were weak in the solution of the types of mathematical problems that were occurring at home. The changes in a great majority of the modern math items are, of course, impossible to determine because there is no previous basis of comparison.

It should be mentioned that in almost all surveys conducted in B.C. the average B.C. pupil has been away above the U.S. norm in Arithmetic Reasoning and slightly above the U.S. norm in Arithmetic Computation. That has not been surprising because B.C. also has been well above the U.S. norm in mental age, usually determined from verbal group tests which have a higher correlation with reasoning than with computation items.

That was true in 1964. In terms of the U.S. modal-age grade equivalents of that date, B.C. medians were 18 months or 1.8 school years ahead in Reasoning and 11 months or 1.1 school years ahead in Computation. When the same test was used in 1970, the B.C. students were found to have lost most of their advantage in Reasoning and more than that in Computation, as follows:

	Excess over U.S. Modal-Age Grade Norms (Yrs. Mo.)		Median, May 1970 in Terms of March, 1964 %iles	
	1964	1970	1964	1970
Arithmetic				
Reasoning	1.8	.8	38.3	50
Computation	1.1	-.1	25.4	50

What is of even greater concern are the differences obtained when individual schools are compared over the 6.2 year period. As an example we may take a school in which the physical facilities have been greatly improved with an open area and all kinds of modern equipment. As it has become an experimental school, the pupil/staff ratio has decreased and particular attention has been paid to "modern maths" demonstrations. The school is in an average area: Arithmetic stanines were 5.1 and 4.8 in 1964, and if anything, the neighbourhood seems to have improved since that time. Meanwhile, the Grade VII enrolment has increased from 47 to 73. Pupils are now on a level and continuous progress, rather than a grade system, with those called "Grade VII" being at the appropriate levels. There is no evidence of excessive promotion by age in the age-grade relationships, with the enrolments in Grades V and VI being 83 and 77 respectively. Here are the comparative results for the school:

	Pupils Writing	Mean (1964) Stanine	Equiv. L.G.	Equiv. Gr. Eq.	Gr. Eq. Years 1970-1964
Arith. Reasoning					
March, 1964	47	5.15	C	9.5	-1.4 -.2
May, 1970	73	4.16	C-	8.1	= -1.6
Arith. Computation					
March, 1964	47	4.79	C	9.1	-1.7 -.2
May, 1970	73	3.11	D	7.4	= -1.9

It must be pointed out that the grade-equivalent comparisons are in terms of pre-1964 U.S. norms. We have no new data but there are rumours that if new U.S. norms were prepared in 1970, they too would be considerably lower.

We do have the B.C. norms in terms of letter grades and percentiles for 1964, however, and a comparison of the percentiles for the two administrations produces the results given on page 1. The detail given on the norm sheet shows that in Reasoning the 99th percentile is exactly the same, but that the difference increases for average and weaker students. The latter is more pronounced in Computation, i.e. the weaker students are being left farther behind in comparison with 1964. This is a matter of considerable concern, because although results in Grades XI and XII show that we are doing a good job of producing a selected group of mathematicians, we still have to deal with members of a much larger group who will have to buy groceries, read a map scale, make time payments and pay taxes.

It is in the applied-mathematics type of item that the greatest increase in difficulty, i.e. in number of errors, is found. But first, was the test valid?

Validity

The validity of a test may be considered from several angles:

(a) Text-book validity - does the test measure what has been taught? This criterion is important in classroom tests.

(b) Curricular validity - does the test measure what a group of knowledgeable persons, or teachers in general have decided are reasonable outcomes in a particular field (say, Arithmetic) at a particular level (say, Grade VII)?

(c) Statistical validity - This is determined for the individual items and involves the assumption that the pupils who get the most correct answers are, in general, the best in Arithmetic, and those who obtain the lowest scores are the worst. If we compare success in individual items of the upper and lower thirds of the students, the most valid items are those where the differences are greatest, i.e. the items that do the best job of distinguishing the best pupils from the weakest ones.

Subjectively, the text-book validity of the Stanford test is much lower in 1970 than it was in 1964. It is still relatively high in (b) however when we consider the application of modern mathematics to everyday computations and problems as they are met. Mileages, interest and taxation have unfortunately not been abolished. And statistically, almost all of the items are more valid in 1970 than they were in 1964 (see the Table). That is largely due to the fact that the average difficulty of the items increased in relation to the average ability of the students. As the difficulty of an item approaches 50% the possible maximum validity rises and most of the items previously had difficulties of less than 50%. It should be noticed that while difficulty and validity are related they are not the same. A test may be invalid because it is entirely too difficult for the pupils, but it is not valid merely because it is easy. Statistically the Stanford items proved to have excellent validity in both the 1964 and the 1970 administrations.

The item-difficulties and item-validities for the two years are given in the Table on page 8. It will be noticed that there is a general increase in the number of errors with the exception of a few items. One of these: If $2m + 10 = 28$, $m =$, is definitely of a "new maths" emphasis. Another involved the setting up of an equation. Most of the remainder involve the meaning of mathematical terminology.

It is in the application of the terminology and in straight everyday computation that most of the old errors remain and, in fact, have increased. In at least seven of the 44 Computation items and one of the Reasoning items a "zero difficulty" is apparent, e.g. 400 X 201 or 3520 ÷ 5. In 1964, 13% and 23% of the pupils marked the respective wrong answers, in 1970 these became 17% and 45%. Operations involving fractions, decimals and percents also have become more difficult. Pupils have always found the following progressively harder:

$$\frac{1}{4} \text{ of } \frac{1}{4} \qquad \frac{1}{4} \times \frac{1}{4} \qquad \frac{1}{4} \div 4$$

Then they encounter: $\frac{1}{4} \div \frac{1}{4}$ The latter can be stated as, "How many quarters are there in one-quarter?" from which the step can be made to: "How many halves are there in one quarter?" $(\frac{1}{4} \div \frac{1}{2})$ and instead of the obvious answer, "None" they can be shown that there is $\frac{1}{2}$ of $\frac{1}{2}$ in $\frac{1}{4}$.

In the past, thousands of mediocre children learned by rote the "invert and multiply" method of dividing by fractions. Many of them never understood what they really were doing, but most of them got the correct answer. If we are now going to emphasize understanding we must see that it is complete, e.g. that $\frac{1}{4} \div \frac{1}{16}$ is really, "How many sixteenths are there in one quarter?" and a diagram shows that the answer can quite logically be larger than either of the original fractions.

Another obvious conclusion that one reaches when studying the items is that many pupils do not reason by analogy. For example, compare the difficulties of the following items in the Computation sub-test:

	1964	1970	
$\frac{2}{5} = \frac{4}{?}$	D = 5%	D = 8%	(The answer 10 was given.)
$\frac{1}{10} + \frac{1}{10} = ?$	D = 20%	D = 34%	(The answer $\frac{1}{5}$ was given but pupils chose "not given" or "2".)

Teachers may draw additional conclusions from the listings of the "more difficult" and "less difficult" items on pages 7 and 8. A brief indication of the process involved has been given for each one that is listed. The remainder are omitted, not because they are easy or hard, but because no significant change in difficulty has occurred since 1964.

Stanford Arithmetic Tests, Advanced, Form L - Grade VII

Difficulty and Validity of Items - March 1964 vs. May 1970

REASONING

Item No.	% Difficulty Mar. 1964	% Difficulty May 1970	Validity (U-L%) Mar. 1964	Validity (U-L%) May 1970	Item No.	% Difficulty Mar. 1964	% Difficulty May 1970	Validity (U-L%) Mar. 1964	Validity (U-L%) May 1970
1	7	5	9	7	24	58	58	34	43
2	3	6	4	9	25	50	54	48	57
3	4	5	5	9	26	39	44	36	47
4	15	13	12	15	27	43	50	4	11
5	12	7	17	6	28	64	61	14	25
6	10	7	17	14	29	73	79	6	10
7	20	20	19	23	30	64	71	27	39
8	11	17	14	16	31	14	10	17	19
9	17	22	20	29	32	36	36	36	54
10	18	14	23	27	33	20	11	25	12
11	12	14	12	25	34	57	52	30	40
12	18	17	26	22	35	17	38	25	57
13	18	17	28	35	36	34	36	25	40
14	39	38	42	55	37	35	43	35	47
15	22	25	26	26	38	20	16	24	25
16	25	30	33	51	39	28	37	31	57
17	20	21	30	22	40	46	45	31	45
18	23	26	32	24	41	58	51	47	55
19	35	34	50	48	42	30	34	16	31
20	34	36	31	48	43	70	64	19	18
21	27	25	36	49	44	59	55	41	44
22	20	37	15	59	45	63	72	13	8
23	21	24	28	47					

Stanford Arithmetic Tests, Advanced, Form L - Grade VII

Difficulty and Validity of Items - March 1964 vs. May 1970

COMPUTATION

Item No.	% Difficulty Mar. 1964	% Difficulty May 1970	Validity (U-L%) Mar. 1964	Validity (U-L%) May 1970	Item No.	% Difficulty Mar. 1964	% Difficulty May 1970	Validity (U-L%) Mar. 1964	Validity (U-L%) May 1970
1	3	9	3	12	23	21	18	24	32
2	12	7	3	8	24	45	53	31	46
3	7	12	9	16	25	25	22	23	26
4	2	2	3	3	26	42	48	53	56
5	10	13	4	7	27	33	46	39	68
6	14	9	11	10	28	31	44	37	52
7	17	16	9	15	29	32	28	31	50
8	12	11	16	19	30	17	17	24	30
9	13	17	15	22	31	17	31	26	46
10	11	15	16	27	32	54	59	28	46
11	20	34	25	37	33	28	42	5	9
12	10	23	13	36	34	60	79	47	27
13	32	47	18	27	35	23	31	21	40
14	32	50	34	46	36	5	8	6	18
15	16	19	25	30	37	62	39	50	64
16	26	45	35	59	38	66	76	41	36
17	30	31	33	35	39	71	72	22	22
18	23	45	22	43	40	88	88	7	7
19	32	47	38	51	41	54	63	34	46
20	28	46	28	48	42	47	46	50	70
21	11	9	12	19	43	76	75	18	20
22	38	46	43	70	44	80	91	28	14

Content of Items Changing in Difficulty from 1964 to 1970

Reasoning

Less Difficult in 1970

Item

4. 14 X 18
5. subtract $ and ¢
6. (75¢/25¢ = 3) X 3

10. time at 40¢ per hour

28. reading gas meter (still very difficult)

31. identification of thousands position
33. lowest common denominator
    (cf. application in Computation items 10, 11)
34. smallest fraction in group of 4
38. setting up equation

41. 4^2 (cf. Computation item 44)

43. meaning of "dividends"
44. meaning of "quotient"

More Difficult in 1970

8. conversion of map scale
9. distracting data. Addition of mixed numbers with price immaterial.
15. 60¢ at 2 for 15¢

16. $4\frac{1}{2}$ X 8

18. (2 X 34¢) + (4 X 21¢)
22. average height (n.b. Computation 31)
23. conversion of map scale
25. %
26. + for speed and - for wind
27. radius, diameter, circumference
29. zero difficulty? $400/10,000 miles
30. instalment % (zero difficulty?)
35. decimal fraction = to $\frac{1}{5}$ (much more difficult)

37. rounding decimal to whole number
39. estimation of largest product: 888 X 101 vs. 888 X 90.9
42. 105%
45. areas

Content of Items Changing in Difficulty from 1964 to 1970

Computation

Less Difficult in 1970

Item

2. addition of $ and ¢ (but cf. subtraction in Computation 3)

6. 37 X 16

29. reading temperature graph

37. solution of equation

(D	1970/'64	39/62
V	1970/'64	64/50)

More Difficult in 1970

1.	205 X 7	zero difficulty
3.	$50.03 - 4.55	" "
5.	sum of 4 numbers	" "
9.	400 X 201	" "
10.	11/12 - 2/3	common denominator
11.	1/10 + 1/10	reduction of fraction
12.	3/5 X 7/10	
13.	1/4 ÷ 1/2	tricky: ÷ vs. 'of'
14.	30% of $10	decimal fraction in %'s, estimation of correct answer?
16.	200 X 2.5	estimation of correct answer or zero
18.	3520/5	zero difficulty
19.	20% of $500	zero difficulty
20.	.081/9	zero difficulty
22.	16 3/4 X 8	128 + 24/4 8 (a+b)
24.	1/4 ÷ 4	see # 13
26.	(100%) - 61%	("not given" in pie diagram but valid)
27.	addition and subtraction	reduction to lowest terms
28.	of lb. and oz.	
31.	average of 3 numbers	estimation of answer
32.	6.71/2.2	decimal fraction
33.	metres and centimetres	see 27 and 28
34.	1/2 X 15 X 18	(very difficult - new math?)
35.	If 25% = x, 100% = ?	%'s
38.	interest & taxation	%'s
41.	(really problems)	
44.	substitution in equation	meaning of r^2

B.C. Norms for STANFORD ARITHMETIC TEST, ADVANCED PARTIAL, FORM L

RAW SCORES, Grade VII, May 1970 vs. March 1964

Letter Grade	Num. Equiv.	Per-centile	Arithmetic Reasoning 1964	Arithmetic Reasoning 1970	Arithmetic Computation 1964	Arithmetic Computation 1970
A	9	99	43	43	42	41
		95	41	40	40	38
			40		39	
	8	90		39		36
	7		39		38	
B		85		37		35
			38		37	
		80		36		
			37		36	33
		75		35		
			36			32
		70			35	
C+	6			34		31
		65	35		34	
						30
		60		33		
			34		33	29
		55		32		
						28
C	5	50	33	31	32	
						27
		45	32	30	31	
						26
		40	31	29		
					30	25
		35	30	28		
C-	4				29	24
		30	29	27		23
		25	28	25	27	22
		20	27	24	26	21
D		15	26	22	25	20
	3					
	2	10	24	20	23	18
		5	21	17	21	15
E	1	1	15	11	16	11
	Means:		32.0	30.0	31.2	27.2

If class medians are used, they should be compared with the 50th percentile in the appropriate Table.

N (1964) 29,204/29,533
N (1970) 38,377/40,252

Research and Standards Branch

B.C. Norms for STANFORD ARITHMETIC TEST, ADVANCED PARTIAL, FORM L

Modal-Age Grade Equivalents
at the Grade VII-6 Level, March 1964, and VII-9 Level, May 1970

Letter Grade	Num. Equiv.	Per-centile	Arithmetic Reasoning 1964	Arithmetic Reasoning 1970	Arithmetic Computation 1964	Arithmetic Computation 1970
A	9	99	12.3	12.3	12.1	11.8
					11.4	
		95	11.8	11.5		10.7
			11.5		11.1	
	8	90		11.2		10.1
	7		11.2		10.7	
B		85		10.7		9.7
			11.0		10.4	9.3
		80		10.4		
			10.7		10.1	9.0
		75		10.1		
			10.4			8.7
		70			9.7	
C+	6			9.7		8.5
		65	10.1		9.3	
						8.2
		60		9.4		
			9.7		9.0	8.0
		55		9.1		
						7.9
C	5	50	9.4	8.7	8.7	
						7.7
		45	9.1	8.5	8.5	
		40	8.7	8.1		
					8.2	7.3
		35	8.5	7.9		
C-	4				8.0	7.1
		30	8.1	7.6		7.0
		25	7.9	7.2	7.7	6.9
		20	7.6	7.0	7.5	6.7
D		15	7.4	6.6	7.3	6.6
	3					
	2	10	7.0	6.3	7.0	6.4
		5	6.4	5.9	6.7	5.7
E	1	1	5.5	4.8	5.9	4.9
Equiv. of Mean Scores:			9.1	8.5	8.5	7.7

APPENDIX B

STANFORD ACHIEVEMENT TESTS:

ARITHMETIC REASONING AND ARITHMETIC COMPUTATION

Advanced Battery
ARITHMETIC TESTS
FORM L

STANFORD ACHIEVEMENT TEST

TRUMAN L. KELLEY • RICHARD MADDEN • ERIC F. GARDNER • LEWIS M. TERMAN • GILES M. RUCH

BRITISH COLUMBIA EDITION

Name _____
Surname Given Name

School District Number _____ School Name _____

Date of Test _____
Day Month Year

Reasoning ☐

TEST 5 *Arithmetic Reasoning* PART I

◀11

DIRECTIONS: Work an example, and then compare your answer with the answers which follow it. If your answer is one of those given, mark the answer space that has the same letter as your answer. Sometimes the correct answer is not given. If you do not find the correct answer, mark the space under the letter for **not given**.

SAMPLES: 51 How many balls are 3 balls and 4 balls? *a* **3** *b* **4** *c* **7** *d* **12** *e* **not given** ..... 51 — *c*

52 How many books are 3 books and 2 books? *f* **2** *g* **3** *h* **4** *i* **6** *j* **not given** 52

---

1 Alice has done 14 problems and Ruth has done 8. How many more problems must Ruth do to equal Alice? *a* **6** *b* **8** *c* **14** *d* **22** *e* **not given** 1

2 The teacher has 27 sheets of paper. How many children will get paper if she gives each child 3 sheets? *f* **3** *g* **9** *h* **24** *i* **30** *j* **not given** 2

3 Mother bought groceries for $1.19. She gave the clerk two half dollars and a quarter. How much change should she receive? *a* **16¢** *b* **19¢** *c* **25¢** *d* **$1.25** *e* **not given** 3

4 Dot's mother is going to buy tomato plants to set out. There are to be 14 rows with 18 plants in each row. How many plants will be needed? *f* **252** *g* **262** *h* **352** *i* **362** *j* **not given** 4

5 Father spent $37.25 last month for gasoline and oil. The gasoline alone cost him $34.67. What did he spend for oil? *a* **$2.58** *b* **$2.62** *c* **$3.52** *d* **$3.62** *e* **not given** 5

6 Onions cost 25¢ for 3 bunches. How many bunches can be bought for 75¢? *f* **9** *g* **10** *h* **12** *i* **25** *j* **not given** 6

7 On an average day, Jane's hens lay a dozen eggs, which will sell for 80¢. How much would that amount to in 7 days? *a* **56¢** *b* **80¢** *c* **87¢** *d* **$5.60** *e* **not given** 7

8 The scale of a map reads that 1 inch = 80 miles. How many inches long must a line on the map be to show a distance of 60 miles? *f* $\frac{3}{4}$ *g* $1\frac{1}{8}$ *h* **9** *i* **48** *j* **not given** 8

9 Thelma wants to buy $1\frac{1}{3}$ yards of ribbon at 12¢ per yard, and $2\frac{1}{3}$ yards at 30¢ per yard. How many yards of ribbon does she want to buy? *a* $3\frac{1}{3}$ *b* $3\frac{2}{3}$ *c* **16** *d* **110** *e* **not given** 9

10 Dick and his father are going to rent a boat for fishing. If they leave at 10 A.M. and return at 2 P.M., how much must they pay for renting the boat at 40¢ an hour? *f* **40¢** *g* **80¢** *h* **$1.60** *i* **$2.00** *j* **not given** 10

11 It is 90 miles to Cloverdale. The scheduled time for the mail train is $3\frac{1}{4}$ hours and that for the streamliner is 2 hours. How many more hours does the mail train take? *a* **1** *b* **2** *c* $3\frac{1}{4}$ *d* $5\frac{1}{4}$ *e* **not given** 11

12 You know how much money you had at the start and at the finish of an automobile trip. To find out how much money you spent on the trip, you would — *f* **add** *g* **multiply** *h* **subtract** *i* **divide** *j* **not given** 12

13 On three days, it rained $\frac{3}{4}$ inch, $\frac{1}{2}$ inch, and $\frac{3}{4}$ inch. How much did it rain during all of these days? *a* $1\frac{1}{4}''$ *b* **2″** *c* $2\frac{1}{2}''$ *d* **3″** *e* **not given** 13

14 On May 1, $640 was deposited in a checking account. Since then there has been a deposit of $360, a withdrawal of $70, and another withdrawal of $110. How much is the balance now? *f* **$100** *g* **$720** *h* **$820** *i* **$1180** *j* **not given** 14

Go on to the next page.

TEST 5 *Arithmetic Reasoning* (Continued)

◀12

15 When ice-cream bars are 2 for 15¢, how many can be bought for 60¢?
a 4 *b* 8 *c* 9 *d* 30 *e* not given ... 15

16 At 8 miles an hour, how many miles can a skater go in $4\frac{1}{2}$ hours?
f 28 *g* 32 *h* 34 *i* 36 *j* not given ... 16

17 A car's mileage read 4185.4 miles at the beginning of a trip. At the end it read 4211.6 miles. How long was the trip?
a 26 mi. *b* 26.2 mi. *c* 27.2 mi. *d* 73.8 mi. *e* not given 17

18 Frank wants 2 balls at 34¢ each and 4 toy cars at 21¢ each. How much will they cost all together?
f 55¢ *g* $1.10 *h* $1.42 *i* $1.52 *j* not given ... 18

19 Bill worked $2\frac{1}{4}$ hours. Fred worked $1\frac{3}{4}$ hours. Ned worked 5 hours. How many hours longer did Ned work than Bill?
a $\frac{1}{2}$ *b* $2\frac{1}{4}$ *c* $2\frac{3}{4}$ *d* $3\frac{3}{4}$ *e* not given ... 19

20 Father bought a radio. The price of the radio plus the carrying charge was $52.50. He paid $20 in cash and agreed to pay the rest in 5 equal monthly payments. How much will each monthly payment be?
f $4 *g* $6.50 *h* $10.50 *i* $14.50 *j* not given ... 20

21 How many miles can a man walk in an hour at the rate of $\frac{3}{4}$ mile in 15 minutes?
a 2 *b* $2\frac{1}{4}$ *c* $3\frac{3}{4}$ *d* 4 *e* not given ... 21

22 The heights of the 5 boys on a basketball team are: 64 inches, 60 inches, 65 inches, 57 inches, and 59 inches. What is the average height of the players in inches?
f 59 *g* 60 *h* 61 *i* 64 *j* not given ... 22

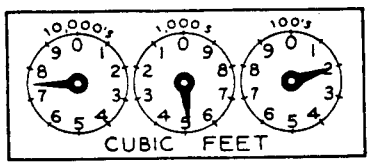
23 How many miles apart are two towns that are $3\frac{1}{2}$ inches apart on the map, if the map scale reads 1 inch = 20 miles?
a $3\frac{1}{2}$ *b* 7 *c* $23\frac{1}{2}$ *d* 70 *e* not given ... 23

24 Ben slept from 9:20 P.M. until 6:35 the next morning. How many hours, to the nearest quarter hour, did he sleep that night?
f $8\frac{3}{4}$ *g* $9\frac{1}{4}$ *h* $9\frac{3}{4}$ *i* $15\frac{3}{4}$ *j* not given ... 24

25 Ruth budgets her yearly allowance this way: clothes, $80; lunches, $50; shows, $20; carfare, $20; miscellaneous, $30. What per cent of her allowance does she spend for clothes?
a 25 *b* $33\frac{1}{3}$ *c* 40 *d* 80 *e* not given ... 25

26 If +250 is the miles per hour which an airplane would travel if there were no wind, and −40 represents the loss of speed in miles per hour, due to a cross wind of 60 miles an hour, how many miles of forward progress does the plane make in an hour?
f 150 mi. *g* 190 mi. *h* 210 mi. *i* 270 mi. *j* not given 26

27 If the radius of a circle is doubled, the circumference will be increased how many times?
a 2 *b* $3\frac{1}{7}$ *c* 4 *d* $6\frac{2}{7}$ *e* not given ... 27



28 What is the reading of the gas meter shown at the left, in cubic feet?
f 762 *g* 25,700 *h* 75,200
i 76,200 *j* not given ... 28

29 Mr. Jones bought a car for $2000. At the end of the year he sold it for $1600. The difference is called *depreciation*. If he drove the car 10,000 miles, how much was the cost per mile for depreciation?
a 1.6¢ *b* 2¢ *c* 3.6¢ *d* 4¢ *e* not given ... 29

30 Furniture which sells for $500 cash costs on an installment plan $90 down and 10 equal payments of $45 each. By what per cent is the installment-purchase cost greater than the cash price?
f 5% *g* 6% *h* 9% *i* 18% *j* not given ... 30

Go on to the next page.

TEST 5 *Arithmetic Reasoning* PART II

◀13

DIRECTIONS: The answer to each of these examples can be thought out without doing any figuring on paper. You are to think out the answer and mark the answer space that is lettered the same as your choice.

31 In which number is the 9 in the thousands position? *a b c d*
a 1,988 *b* 11,911 *c* 19,111 *d* 88,911 31

32 If all the odd-numbered houses are on the same side of the street, which of the following would be on the same side as No. 2437? *e f g h*
e No. 2432 *f* No. 2524 *g* No. 2645 *h* No. 3724 32

33 What is the lowest common denominator for $\frac{3}{4}$, $\frac{3}{8}$, and $\frac{1}{2}$? *a b c d*
a 8 *b* 16 *c* 32 *d* 64 33

34 Which is the smallest fraction? *e f g h*
e $\frac{1}{9}$ *f* $\frac{1}{12}$ *g* $\frac{1}{18}$ *h* $\frac{2}{21}$ 34

35 $\frac{1}{5}$ = *a b c d*
a .20 *b* .05 *c* .01 *d* .00$\frac{1}{5}$ 35

36 The amount left from a sale after costs and expenses are taken out is called — *e f g h*
e rent *f* profit *g* wholesale *h* commission 36

37 How much is 46.735 rounded off to a whole number? *a b c d*
a 46 *b* 46.7 *c* 46.8 *d* 47 37

38 Dorothy earned b cents and spent d cents. How many cents did she have left? *e f g h*
e bd *f* $\frac{b}{d}$ *g* $b - d$ *h* $\frac{d}{b}$ 38

39 By estimation, choose the example which will have the largest product. *a b c d*
a 888 × 90.9 *b* 888 × 9.09 *c* 888 × 10.1 *d* 888 × 101 39

40 Which is the same as "18 more than a number = 44"? *e f g h*
e $18N = 14$ *f* $\frac{18}{N} = 44$ *g* $N + 18 = 44$
h $N = 18 + 44$ 40

41 4^2 = *a b c d*
a 2 *b* 4 *c* 8 *d* 16 41

42 How would 105% of a number compare in size with the number? *e f g h*
e more than twice *f* slightly larger *g* slightly smaller
h less than half 42

43 For the use of money paid for a share of its stock, a company pays — *a b c d*
a dividends *b* bonds *c* a premium *d* a mortgage 43

44 By estimation, choose the quotient which will be larger than 1. *e f g h*
e $136 \div 135\frac{3}{4}$ *f* $125 \div 125\frac{1}{4}$ *g* $148 \div 148$ *h* $152 \div 153$ 44

45 When the dimensions of a square are doubled, its area becomes how many times as large? *a b c d*
a 2 *b* 4 *c* 6 *d* 8 45

Stop.

No. RIGHT	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	36	37	38	39	40	41	42	43	44	45
Gr. score	24	28	31	33	36	38	40	43	45	47	48	50	52	54	55	57	59	60	61	63	64	66	68	70	72	74	76	79	81	85	87	91	94	97	101	104	107	110	112	115	118	121	123	126	128

Advanced Battery ARITHMETIC TESTS — FORM L

STANFORD ACHIEVEMENT TEST

TRUMAN L. KELLEY • RICHARD MADDEN • ERIC F. GARDNER • LEWIS M. TERMAN • GILES M. RUCH

BRITISH COLUMBIA EDITION

Name _____
Surname Given Name

School District Number _____ School Name _____

Date of Test _____
Day Month Year

Computation ☐

TEST 6 *Arithmetic Computation*

◀ 14

DIRECTIONS: Work each example. Then compare your answer with the answers given at the right of the example. If your answer is one of those given, mark the answer space that has the same letter as your answer. Sometimes the correct answer is not given. If the correct answer is not given, mark the answer space under the letter for **not given**. Look carefully at each example to see what it tells you to do. If you need to do any figuring, use a separate sheet of paper.

1	Multiply $\begin{array}{r} 205 \\ \times\ 7 \end{array}$	*a* 1235 *b* 1505 *c* 1615 *d* 1705 *e* not given	a b c d e
2	Add $\begin{array}{r} \$8.70 \\ 5.65 \end{array}$	*f* \$13.35 *g* \$13.45 *h* \$14.35 *i* \$15.35 *j* not given	f g h i j
3	Subtract $\begin{array}{r} \$5.03 \\ 4.55 \end{array}$	*a* \$.48 *b* \$.58 *c* \$1.48 *d* \$1.52 *e* not given	a b c d e
4	$34\overline{)68}$	*f* 1 *g* 2 *h* 3 *i* 20 *j* not given	f g h i j
5	Add $\begin{array}{r} 638 \\ 67 \\ 56 \\ 334 \end{array}$	*a* 985 *b* 995 *c* 1085 *d* 1195 *e* not given	a b c d e
6	Multiply $\begin{array}{r} 37 \\ \times\ 16 \end{array}$	*f* 582 *g* 592 *h* 602 *i* 692 *j* not given	f g h i j
7	Subtract $\begin{array}{r} 211.355 \\ 174.879 \end{array}$	*a* 36,476 *b* 37,576 *c* 46,576 *d* 47,476 *e* not given	a b c d e
8	$42\overline{)1428}$	*f* 33 *g* $36\frac{6}{21}$ *h* 304 *i* 340 *j* not given	f g h i j
9	Multiply $\begin{array}{r} 400 \\ \times\ 201 \end{array}$	*a* 8400 *b* 80,400 *c* 82,600 *d* 84,400 *e* not given	a b c d e
10	Subtract $\begin{array}{r} 1\frac{1}{2} \\ \frac{2}{3} \end{array}$	*f* $\frac{1}{6}$ *g* $\frac{5}{12}$ *h* $\frac{3}{4}$ *i* $\frac{9}{9}$ *j* not given	f g h i j
11	Add $\begin{array}{r} \frac{1}{10} \\ \frac{1}{10} \end{array}$	*a* $\frac{1}{10}$ *b* $\frac{1}{5}$ *c* 2 *d* 11 *e* not given	a b c d e
12	$\frac{3}{5} \times \frac{7}{10} =$	*f* $\frac{21}{50}$ *g* $\frac{2}{3}$ *h* $\frac{6}{7}$ *i* $2\frac{1}{10}$ *j* not given	f g h i j
13	$\frac{1}{4} \div \frac{1}{2} =$	*a* 2 *b* $\frac{1}{2}$ *c* $\frac{1}{4}$ *d* $\frac{1}{8}$ *e* not given	a b c d e
14	30% of \$40 =	*f* $\$\frac{30}{100}$ *g* $\$1\frac{1}{3}$ *h* \$7.50 *i* \$12 *j* not given	f g h i j
15	Add $\begin{array}{r} 2\frac{1}{4} \\ \frac{1}{4} \\ 1\frac{1}{2} \end{array}$	*a* 5 *b* 4 *c* $3\frac{3}{4}$ *d* 3 *e* not given	a b c d e

Go on to the next page.

TEST 6 *Arithmetic Computation* (Continued)

◀ 15

16 $200 \times 2.5 =$ — *f* .50 *g* 5 *h* 5.00 *i* 500 *j* not given

17 Add 4667.55, 786.68, 99.64, 6547.78 — *a* 12,001.65 *b* 12,100.65 *c* 12,101.65 *d* 12,110.65 *e* not given

18 $5\overline{)3520}$ — *f* 74 *g* 704 *h* 724 *i* 740 *j* not given

19 Selling Price = $500; Rate of Commission = 20%; Commission = ? — *a* $5.20 *b* $40 *c* $100 *d* $400 *e* not given

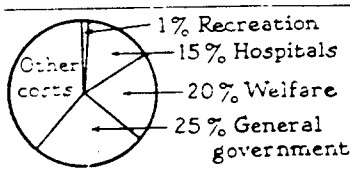
20 $9\overline{).081}$ — *f* .009 *g* .09 *h* .9 *i* 9 *j* not given

21 Subtract $2\frac{3}{4}$ from $\frac{3}{4}$ — *a* 1 *b* $1\frac{1}{2}$ *c* 2 *d* $3\frac{1}{2}$ *e* not given

22 Multiply $16\frac{3}{4}$ by 8 — *f* $24\frac{3}{4}$ *g* 128 *h* $128\frac{3}{4}$ *i* 134 *j* not given

23 If $8y = 56$, $y =$ — *a* 7 *b* 8 *c* 49 *d* 64 *e* not given

24 $\frac{1}{4} \div 4 =$ — *f* $\frac{1}{16}$ *g* 1 *h* 4 *i* 16 *j* not given



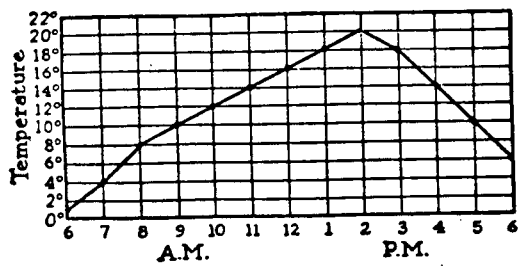
25 What per cent of the taxes was spent for hospitals and welfare? — *a* 15% *b* 20% *c* 35% *d* 60% *e* not given

26 What per cent was spent for "other costs"? — *f* 39% *g* 40% *h* 60% *i* 61% *j* not given

27 Subtract 9 lb. 8 oz. − 6 lb. 12 oz. — *a* 16 lb. 8 oz. *b* 16 lb. 4 oz. *c* 2 lb. 12 oz. *d* 2 lb. 2 oz. *e* not given

28 Add 6 lb. 9 oz., 12 lb. 13 oz., 7 lb. 8 oz. — *f* 28 lb. *g* 27 lb. 14 oz. *h* 27 lb. 6 oz. *i* 26 lb. 14 oz. *j* not given

TEMPERATURE CHART



29 How many degrees warmer was it at 3 P.M. than it was at 9 A.M.? — *a* 4° *b* 8° *c* 10° *d* 18° *e* not given

30 How much did the temperature fall from 2 P.M. to 6 P.M.? — *f* 5° *g* 10° *h* 18° *i* 20° *j* not given

31 Find the average of 16 ft., 9 ft., 11 ft. — *a* 9 ft. *b* 11 ft. *c* 12 ft. *d* 36 ft. *e* not given

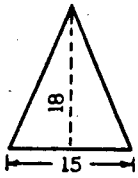
Go on to the next page.

TEST 6 *Arithmetic Computation* (Continued)

◀16

32 $2.2\overline{)6.71}$ — *f* .305 *g* 3.05 *h* 30.5 *i* 305 *j* not given ..... 32 (f g h i j)

33 Add 15 m. 8 cm. / 4 m. 5 cm. — *a* 19 m. 13 cm. *b* 19 m. 3 cm. *c* 20 m. 3 cm. *d* 21 m. 3 cm. *e* not given ..... 33 (a b c d e)



34 If $A = \frac{1}{2}bh$, what is the area of the triangle shown at the left?
f $16\frac{1}{2}$ *g* 33 *h* 135 *i* 270 *j* not given ..... 34 (f g h i j)

35 If 25% of an amount is $1.25, what is the amount?
a $20 *b* $6.00 *c* $.3125 *d* $.0125 *e* not given ..... 35 (a b c d e)

36 $\frac{2}{5} = \frac{4}{?}$ — *f* 5 *g* 10 *h* 20 *i* 40 *j* not given ..... 36 (f g h i j)

37 If $2m + 10 = 28$, $m =$ — *a* 9 *b* 12 *c* 16 *d* 18 *e* not given ..... 37 (a b c d e)

38 Assessed Valuation = $4000
Tax Rate per $100 = $2.50
Amount of Tax = ?
f $42.50 *g* $100 *h* $4250 *i* $10,000 *j* not given ..... 38 (f g h i j)

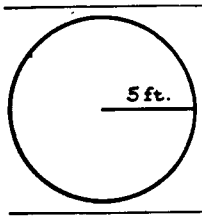
39 $\frac{-21}{+3} =$ — *a* 18 *b* 7 *c* −21 *d* −24 *e* not given ..... 39 (a b c d e)

40 Multiply $+4a$ / -3 — *f* $12a$ *g* -12 *h* 12 *i* $1a$ *j* not given ..... 40 (f g h i j)

41 Principal = $600
Annual Interest = $30
Rate of Interest = ?
a .5% *b* 2% *c* 5% *d* 50% *e* not given ..... 41 (a b c d e)

42 If $\frac{n}{3} = 18$, $n =$ — *f* 6 *g* 15 *h* 21 *i* 54 *j* not given ..... 42 (f g h i j)

43 Principal = $800
Rate = 3%
Time = 6 mo.
Interest = ?
a $4 *b* $12 *c* $24 *d* $144 *e* not given ..... 43 (a b c d e)



44 If $A = \pi r^2$, what is the area of the circle shown at the left? ($\pi = 3.14$)
f 78.50 sq. ft. *g* 77.50 sq. ft. *h* 31.40 sq. ft.
i 15.70 sq. ft. *j* not given ..... 44 (f g h i j)

Stop.

No. right	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	36	37	38	39	40	41	42	43	44
Gr. score	21	28	31	33	35	38	40	43	45	47	49	51	53	55	57	59	61	62	64	66	67	69	70	71	73	75	77	79	80	82	85	87	90	93	97	101	104	107	111	114	118	121	124	127

APPENDIX C

THE COMPUTER PROGRAM BICAL

The computer program used to perform the analysis of data was BICAL (Wright & Mead, 1978). The description in this appendix relies heavily on the documentation given for the program. The program consists of three major sections: input, estimation, and fit.

The input portion reads the control cards, stores the data for each person, and computes the frequency of raw scores and the proportion of correct responses on each item (marginals). As part of this section a subroutine eliminates zero and perfect scores from item and person files. The user has the option of specifying minimum and maximum scores to be included in the calibration sample. This feature helps to alleviate the guessing problem by eliminating from the calibration process those scores less that that due to chance alone on multiple-choice questions. There is also a facility by which items may be removed from the analysis thereby allowing recalibration without changing other control cards. There are several forms in which data may be input; in the present study each item was coded 0,1,2,3,4, or 5 according to response selected, and a scoring key was provided.

The estimation section calculates the estimates of ability and difficulty from the marginal person and item score distributions. There are two estimation options available. The first (PROX) is an approximate method using Cohen's procedure (Wright & Douglas, 1977a) which may be used for long tests with symmetrical score distributions. The second (UCON) is the corrected unconditional maximum likelihood procedure

(Wright & Douglas, 1977b) suggested for use with shorter tests and skewed distributions. Given the final difficulty estimates the program computes the corresponding ability estimates for all raw scores. The origin for both persons and items is at the centre of estimated item difficulties.

The fit section computes a mean square test of fit for each item and organizes the results into a summary table. The sample is divided into a number of subgroups (maximum of six) stratified by total scores. The observed successes on each item in each score subgroup are compared with those predicted for that subgroup from the total sample estimate. The model suggests that the difficulty estimates are independent of the ability of the sample, hence there should be close agreement between observed and predicted successes on each item. Finally, BICAL calculates a residual index of the slope of each item characteristic curve after the model has been fitted. This statistic may be interpreted as an index of item discrimination.

An example of the output of BICAL is included in this appendix. The following detailed description of the printout on each page should give some insight into the data analysis.

Page 1 lists the control cards. It also lists the item reponses for the first subject and gives the total number of items and subjects to ensure that the data file was read correctly.

Page 2 tabulates the responses to each item. The letters indicate that the items were from the reasoning test

and the computation test. The figure in the "unknown" column is the number of times the respondents omitted the item or responded unacceptably, for example, by marking several responses for the item.

Page 3 summarizes the editing process. Four subjects whose scores were less than the minimum of 15 were deleted from the analysis. No one achieved a perfect score.

Page 4 shows the frequencies of raw scores and the corresponding histogram. An inspection of the histogram shows a preponderance of high scores.

Page 5 shows the number of correct responses for each item on the test and the corresponding histogram. Again, casual inspection shows the items to be generally easy.

Page 6 lists the item difficulty estimates and the associated estimates of the standard error of calibration. The mean estimate of ability is 1.23. Items having difficulties close to this value show the least standard error. These items are best matched to the ability of the subjects and are best estimated. Items of least difficulty are least best estimated as shown by the higher standard error of calibration. The scale factors at the top of the page are related to the PROX (Cohen) procedure of approximate estimation.

The unnumbered page following page six shows the relationship between raw scores and estimates of person abilities, along with the standard error of measurement related to each score. The mean ability at the bottom of the page is the summary statistic for the abilities of all persons

in the sample. The curve is the function relating raw scores on the vertical axis and ability on the horizontal axis, with the typical logistic shape.

Pages 7 and 8 contain information on which to judge the degree of fit to the model for each item. The subjects are divided into, in this case, six groups ranging from low ability to high ability. The score range, number of subjects, and mean ability for each group are shown at the bottom of page 7. The figures in the six columns under "item characteristic curve" are the fraction of each group correctly answering a given item. On item C19, for example, 33% of the 45 persons in the lowest ability group gave correct answers. Moving across the groups for item C19 it will be noted that the item is well-behaved; as ability increases the proportion of correct answers increases. This trend is not evident for item C39. The six columns under "departure from expected ICC" show how the obtained values differ from values predicted from the theoretical item characteristic curve based on the mean group ability and the estimated difficulty of the item. For example, 16% of the first group correctly answered item C39, and this is 8 percentage points more than expected.

Under the section labelled "fit mean square" the "within group" column indicates the variance remaining in the groups after removing the effect of differences in the shapes of characteristic curves. If the correct proportion of the group succeeded but the wrong people in the group were successful the within group value will be large relative to the between group variance. The "between group" variance

serves to evaluate the agreement between the observed item characteristic curve and the theoretical curve. These mean squares have expected values of unity. A non-significant value indicates that statistically equivalent estimates of item difficulty are produced regardless of which scoring group was used in the calibration.

The basic fit statistic for each item is the one given under the heading "total". Its expected value again is unity. The value will be large when the observed trend does not follow the predicted trend, for example, if too many higher ability persons fail or vice versa. It is an indicator of disagreement between the ability called for on the item and that defined by the aggregate of items.

The discrimination index is related to the pattern of "departures from expected ICC". Its model value is unity. If the pattern of departures runs from negative to positive across the six groups, the discrimination index will be greater than unity, for example, item C19. If the trend is positive to negative the index will be less than unity, for example, item C40. The index is a measure of the linear residual trend across score groups.

The point biserial values are the customary correlations between a subject's success on each item and his or her estimated ability score from the test.

On page 8, the items are arranged in three different ways in order to facilitate retrieval of information. The fit order on the right hand side is the most useful in selecting non-fitting items.

Page 9 shows the relationship between probability of success for a score group on a given item and the mean square for the group. The latter figure is obtained by standardizing and squaring the figures in the centre panel on page 7. This plot is useful on multiple-choice questions where the presence of guessing is indicated by large values of mean squares located to the left of the chance probability level for the test. In this instance item 40 would be a good candidate for guessing.

Pages 10, 11, and 12 contain two-way plots of item difficulty, residual discrimination, and total fit mean square. They might be useful in determining by inspection any particularly interesting trends.

1964 computation test calibration

CONTROL PARAMETERS

NITEM	NGROP	MINSC	MAXSC	LREC	KCA	B SCORE	1	2	3	4	5	6	7	8	9	10	11
44	0	14	43	160	2	0	0	0	0	0	0	0	0	0	0	0	0

COLUMNS SELECTED

```
         1         2         3         4         5         6         7         8
1********0*********0*********0*********0*********0*********0*********0*********0

           1111111111 1111111111 1111111111  1111111111 1111
```

KEY

KEY

```
           5312521525 2124243231 3411313425  3213521255 3421
```

FIRST SUBJECT

```
 0020000000 1  1241115124 2353152425 4325314432  3113431131 3443300000
```

FIRST SUBJECT

```
 0020000000 2  5312352524 5142334235 3122434545  5515154131 5155000000
```

NUMBER OF ITEMS 44
NUMBER OF SUBJT 296

ALTERNATIVE RESPONSE FREQUENCIES

SEQ NUM	ITEM NAME		1	2	3	4	5	UNKN		KEY
1	C1	I	0	4	1	1	290	0	I	5
2	C2	I	11	3	267	8	6	1	I	3
3	C3	I	276	5	7	0	8	0	I	1
4	C4	I	0	291	1	2	2	0	I	2
5	C5	I	0	3	11	14	268	0	I	5
6	C6	I	6	258	2	10	19	1	I	2
7	C7	I	255	4	5	1	28	3	I	1
8	C8	I	6	11	3	5	269	2	I	5
9	C9	I	24	260	0	0	10	2	I	2
10	C10	I	7	11	3	2	272	1	I	5
11	C11	I	9	248	4	0	34	1	I	2
12	C12	I	269	1	5	3	15	3	I	1
13	C13	I	29	210	12	31	13	1	I	2
14	C14	I	6	14	15	211	43	7	I	4
15	C15	I	8	255	11	10	11	1	I	2
16	C16	I	2	5	8	225	53	3	I	4
17	C17	I	13	12	202	3	63	3	I	3
18	C18	I	60	228	0	4	3	1	I	2
19	C19	I	19	9	213	12	39	4	I	3
20	C20	I	224	52	8	2	5	5	I	1
21	C21	I	0	1	269	20	4	2	I	3
22	C22	I	6	11	30	203	44	2	I	4
23	C23	I	224	17	4	10	33	8	I	1
24	C24	I	159	94	4	19	17	3	I	1
25	C25	I	35	11	223	3	20	4	I	3
26	C26	I	169	5	7	14	98	3	I	1
27	C27	I	2	12	215	10	50	7	I	3
28	C28	I	6	5	7	209	67	2	I	4
29	C29	I	25	214	22	11	21	3	I	2
30	C30	I	13	8	4	18	250	3	I	5
31	C31	I	9	2	249	16	19	1	I	3
32	C32	I	13	150	48	5	78	2	I	2
33	C33	I	207	2	65	3	15	4	I	1
34	C34	I	19	26	115	65	61	10	I	3
35	C35	I	13	20	28	3	231	1	I	5
36	C36	I	5	284	4	0	2	1	I	2
37	C37	I	108	13	59	57	47	12	I	1
38	C38	I	36	101	28	32	78	21	I	2
39	C39	I	38	67	18	77	78	18	I	5
40	C40	I	101	29	7	98	40	21	I	5
41	C41	I	27	42	130	17	62	18	I	3
42	C42	I	81	19	15	152	17	12	I	4
43	C43	I	15	68	94	51	44	24	I	2
44	C44	I	45	15	55	79	72	30	I	1

NUMBER OF ZERO SCORES 0
NUMBER OF PERFECT SCORES 0

NUMBER OF ITEMS SELECTED 44
NUMBER OF ITEMS NAMED 44

SUBJECTS BELOW	14	3
SUBJECTS ABOVE	43	0
SUBJECTS IN CALIB.		293
TOTAL SUBJECTS		296

REJECTED ITEMS

ITEM NUMBER	ITEM NAME	ANSWERED CORRECTLY
NONE		

SUBJECTS DELETED = 0
SUBJECTS REMAINING = 293

ITEMS DELETED = 0
POSSIBLE SCORE = 44

MINIMUM SCORE = 14
MAXIMUM SCORE = 43

1964 computation test calibration
SCORE DISTRIBUTION OF ABILITY

	COUNT	PROPORTION	2 4 6 8 10
			I*********0*********0*********0*********0*********0I
1	0	0.0	I I
2	0	0.0	I I
3	0	0.0	I I
4	0	0.0	I I
5	0	0.0	I I
6	0	0.0	I I
7	0	0.0	I I
8	0	0.0	I I
9	0	0.0	I I
10	0	0.0	I I
11	0	0.0	I I
12	1	0.00	I I
13	2	0.01	IXX I
14	1	0.00	I I
15	1	0.00	I I
16	3	0.01	IXXX I
17	0	0.0	I I
18	1	0.00	I I
19	3	0.01	IXXX I
20	5	0.02	IXXXXX I
21	6	0.02	IXXXXXX I
22	7	0.02	IXXXXXXX I
23	9	0.03	IXXXXXXXXX I
24	9	0.03	IXXXXXXXXX I
25	10	0.03	IXXXXXXXXXX I
26	11	0.04	IXXXXXXXXXXX I
27	21	0.07	IXXXXXXXXXXXXXXXXXXXX I
28	15	0.05	IXXXXXXXXXXXXXX I
29	14	0.05	IXXXXXXXXXXXXX I
30	15	0.05	IXXXXXXXXXXXXXX I
31	12	0.04	IXXXXXXXXXXX I
32	14	0.05	IXXXXXXXXXXXXX I
33	17	0.06	IXXXXXXXXXXXXXXXX I
34	31	0.11	IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
35	18	0.06	IXXXXXXXXXXXXXXXXX I
36	23	0.08	IXXXXXXXXXXXXXXXXXXXXXX I
37	16	0.05	IXXXXXXXXXXXXXXX I
38	11	0.04	IXXXXXXXXXXX I
39	5	0.02	IXXXXX I
40	5	0.02	IXXXXX I
41	6	0.02	IXXXXXX I
42	4	0.01	IXXXX I
43	0	0.0	I I
44	0	0.0	I I

EACH X = 0.34 PERCENT

```
1964 computation test calibration
ITEM        DISTRIBUTION OF EASINESS
------------------------------------------------------------------------------------
       COUNT   PROPORTION       2         4         6         8        10
                          I*********0*********0*********0*********0*********0I
  1     287     0.98      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  2     265     0.90      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  3     274     0.94      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  4     289     0.99      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  5     266     0.91      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  6     258     0.88      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  7     253     0.86      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  8     268     0.91      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
  9     258     0.88      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 10     271     0.92      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 11     247     0.84      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 12     267     0.91      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 13     210     0.72      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 14     211     0.72      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 15     254     0.87      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 16     225     0.77      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 17     202     0.69      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 18     226     0.77      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 19     212     0.72      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 20     224     0.76      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 21     267     0.91      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 22     203     0.69      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 23     224     0.76      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 24     159     0.54      IXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 25     222     0.76      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 26     169     0.58      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 27     215     0.73      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 28     209     0.71      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 29     214     0.73      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 30     248     0.85      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 31     248     0.85      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 32     149     0.51      IXXXXXXXXXXXXXXXXXXXXXXXXX I
 33     204     0.70      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 34     115     0.39      IXXXXXXXXXXXXXXXXXXXX I
 35     230     0.78      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 36     282     0.96      IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX I
 37     108     0.37      IXXXXXXXXXXXXXXXXXX I
 38     101     0.34      IXXXXXXXXXXXXXXXXX I
 39      78     0.27      IXXXXXXXXXXXXX I
 40      40     0.14      IXXXXXXX I
 41     130     0.44      IXXXXXXXXXXXXXXXXXXXXXX I
 42     152     0.52      IXXXXXXXXXXXXXXXXXXXXXXXXXX I
 43      67     0.23      IXXXXXXXXXXX I
 44      45     0.15      IXXXXXXXX I
```

1964 computation test calibration

PROCEDURE USED UCON

DIFFICULTY SCALE FACTOR 1.148 ABILITY SCALE FACTOR 1.35
NUMBER OF ITERATIONS = 2

SEQUENCE NUMBER	ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR	LAST DIFF CHANGE	PROX DIFF	FIRST CYCLE	
1	C1	-2.922	0.413	-0.018	-3.153	-2.905	II
2	C2	-1.269	0.205	-0.012	-1.293	-1.257	II
3	C3	-1.708	0.242	-0.014	-1.777	-1.695	II
4	C4	-3.332	0.502	-0.018	-3.626	-3.315	II
5	C5	-1.311	0.208	-0.013	-1.339	-1.299	II
6	C6	-1.004	0.188	-0.010	-1.007	-0.994	II
7	C7	-0.840	0.178	-0.009	-0.831	-0.831	II
8	C8	-1.400	0.215	-0.013	-1.436	-1.387	II
9	C9	-1.004	0.188	-0.010	-1.007	-0.994	II
10	C10	-1.545	0.227	-0.013	-1.596	-1.532	II
11	C11	-0.662	0.169	-0.008	-0.643	-0.654	II
12	C12	-1.355	0.212	-0.013	-1.387	-1.342	II
13	C13	0.177	0.140	-0.000	0.221	0.177	II
14	C14	0.158	0.140	-0.000	0.201	0.158	II
15	C15	-0.871	0.180	-0.009	-0.864	-0.862	II
16	C16	-0.127	0.148	-0.003	-0.087	-0.123	II
17	C17	0.327	0.137	0.002	0.371	0.325	II
18	C18	-0.148	0.149	-0.003	-0.109	-0.145	II
19	C19	0.139	0.141	-0.000	0.182	0.139	II
20	C20	-0.105	0.147	-0.003	-0.065	-0.102	II
21	C21	-1.355	0.212	-0.013	-1.387	-1.342	II
22	C22	0.308	0.137	0.001	0.353	0.307	II
23	C23	-0.105	0.147	-0.003	-0.065	-0.102	II
24	C24	1.055	0.129	0.010	1.090	1.045	II
25	C25	-0.063	0.146	-0.003	-0.022	-0.060	II
26	C26	0.893	0.129	0.008	0.931	0.885	II
27	C27	0.080	0.142	-0.001	0.123	0.081	II
28	C28	0.196	0.140	0.000	0.240	0.196	II
29	C29	0.099	0.142	-0.001	0.143	0.100	II
30	C30	-0.690	0.170	-0.008	-0.673	-0.683	II
31	C31	-0.690	0.170	-0.008	-0.673	-0.683	II
32	C32	1.216	0.128	0.012	1.247	1.204	II
33	C33	0.290	0.138	0.001	0.334	0.289	II
34	C34	1.771	0.131	0.019	1.788	1.752	II
35	C35	-0.237	0.152	-0.004	-0.200	-0.233	II
36	C36	-2.296	0.310	-0.016	-2.437	-2.281	II
37	C37	1.890	0.132	0.020	1.904	1.870	II
38	C38	2.012	0.134	0.021	2.024	1.991	II
39	C39	2.444	0.143	0.024	2.450	2.420	II
40	C40	3.386	0.180	0.028	3.403	3.359	II
41	C41	1.523	0.129	0.016	1.546	1.507	II
42	C42	1.168	0.128	0.012	1.200	1.157	II
43	C43	2.676	0.150	0.025	2.682	2.651	II
44	C44	3.233	0.172	0.028	3.245	3.206	II

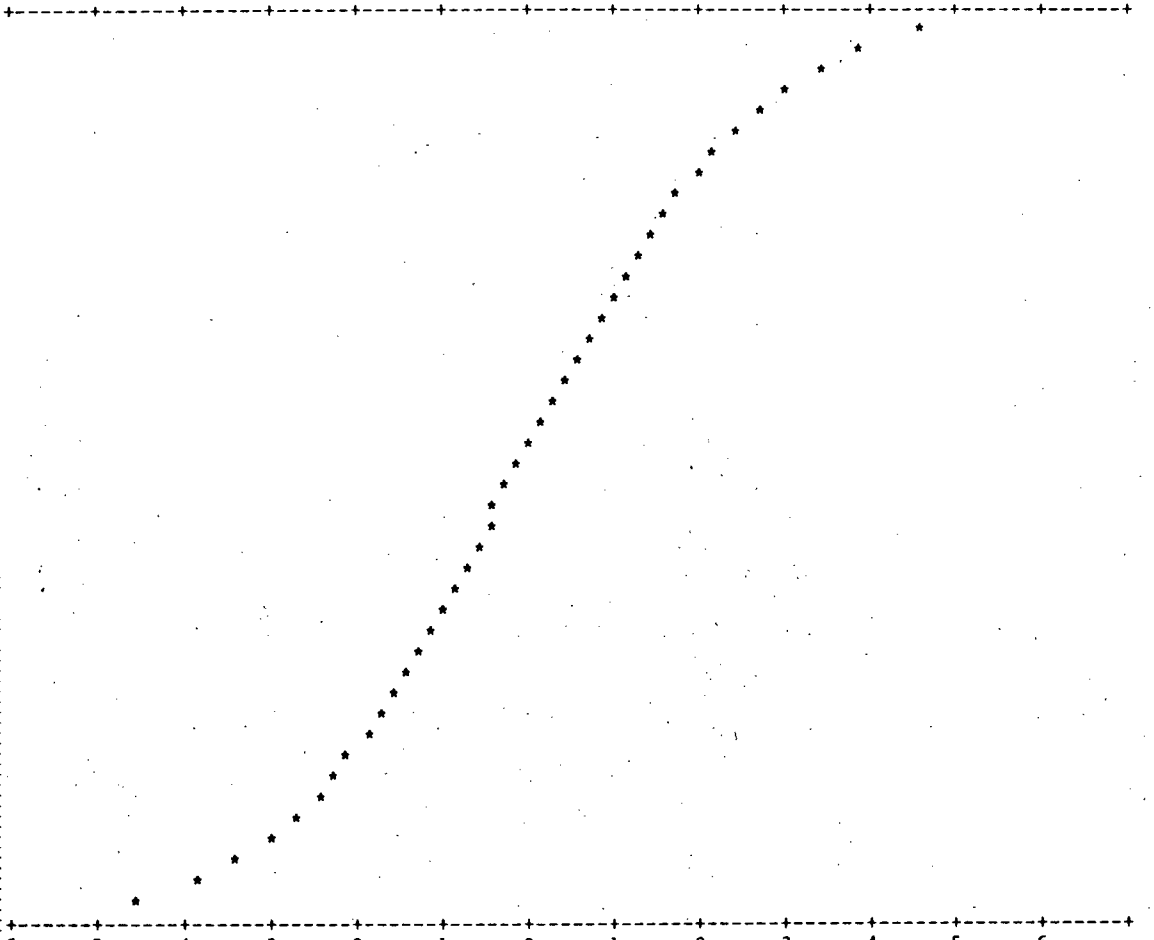
ROOT MEAN SQUARE = 0.013

MEAN ABILITY = 1.23

COMPLETE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	LOG ABILITY	STANDARD ERRORS
43	0	4.71	1.06
42	4	3.93	0.78
41	6	3.44	0.66
40	5	3.06	0.59
39	5	2.75	0.55
38	11	2.48	0.51
37	16	2.24	0.48
36	23	2.02	0.46
35	18	1.83	0.44
34	31	1.64	0.43
33	17	1.47	0.41
32	14	1.31	0.40
31	12	1.15	0.39
30	15	1.00	0.39
29	14	0.86	0.38
28	15	0.72	0.37
27	21	0.59	0.37
26	11	0.46	0.36
25	10	0.33	0.36
24	9	0.20	0.36
23	9	0.08	0.36
22	7	-0.05	0.36
21	6	-0.17	0.35
20	5	-0.29	0.36
19	3	-0.42	0.36
18	1	-0.54	0.36
17	0	-0.67	0.36
16	3	-0.79	0.36
15	1	-0.92	0.37
14	1	-1.06	0.37
13	2	-1.20	0.38
12	1	-1.34	0.39
11	0	-1.49	0.39
10	0	-1.65	0.41
9	0	-1.82	0.42
8	0	-2.00	0.44
7	0	-2.20	0.46
6	0	-2.42	0.49
5	0	-2.67	0.53
4	0	-2.96	0.57
3	0	-3.32	0.64
2	0	-3.80	0.76
1	0	-4.57	1.04

TEST CHARACTERSTC CURVE

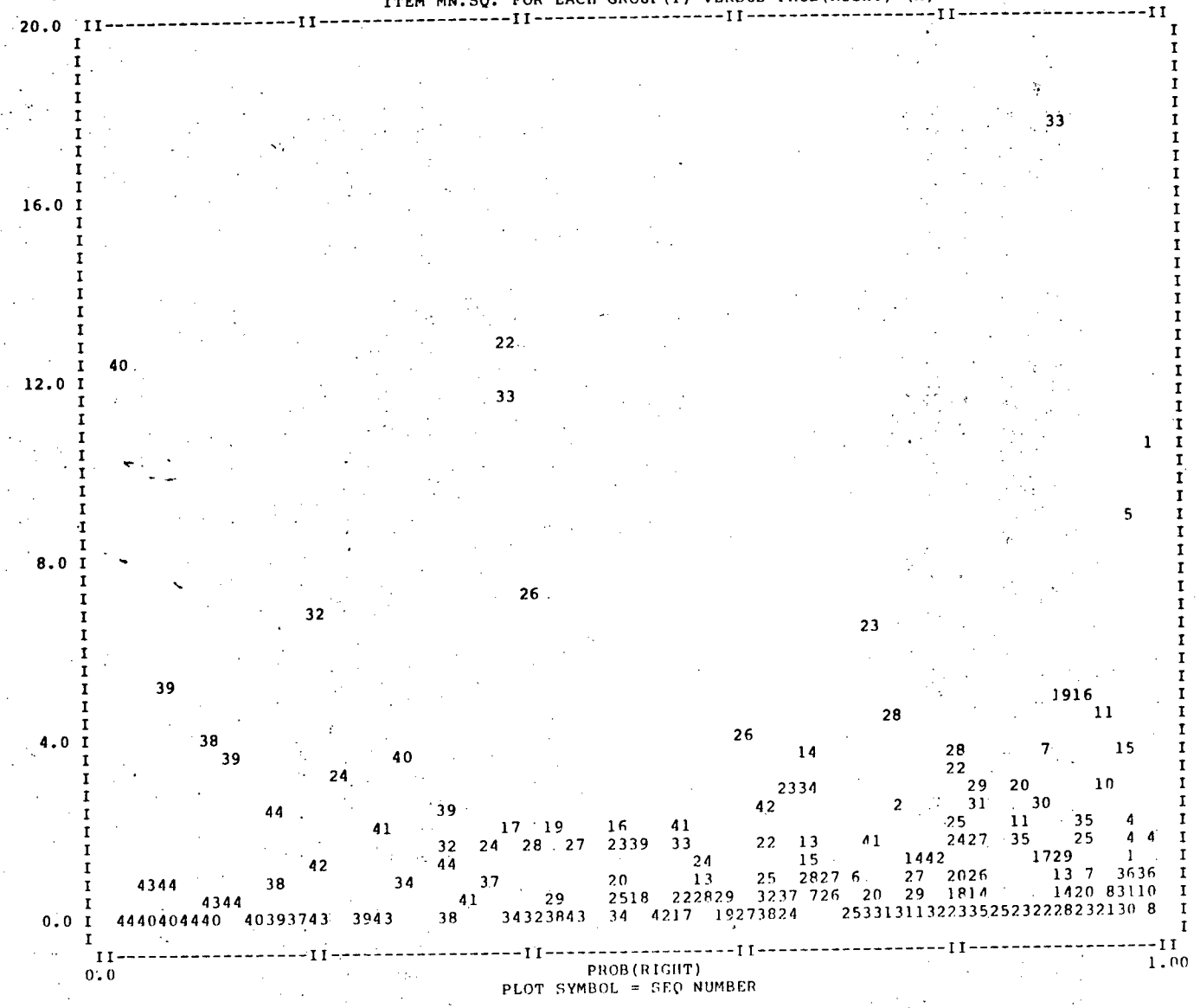


MEAN ABILITY = 1.23
SD OF ABILITY= 0.88

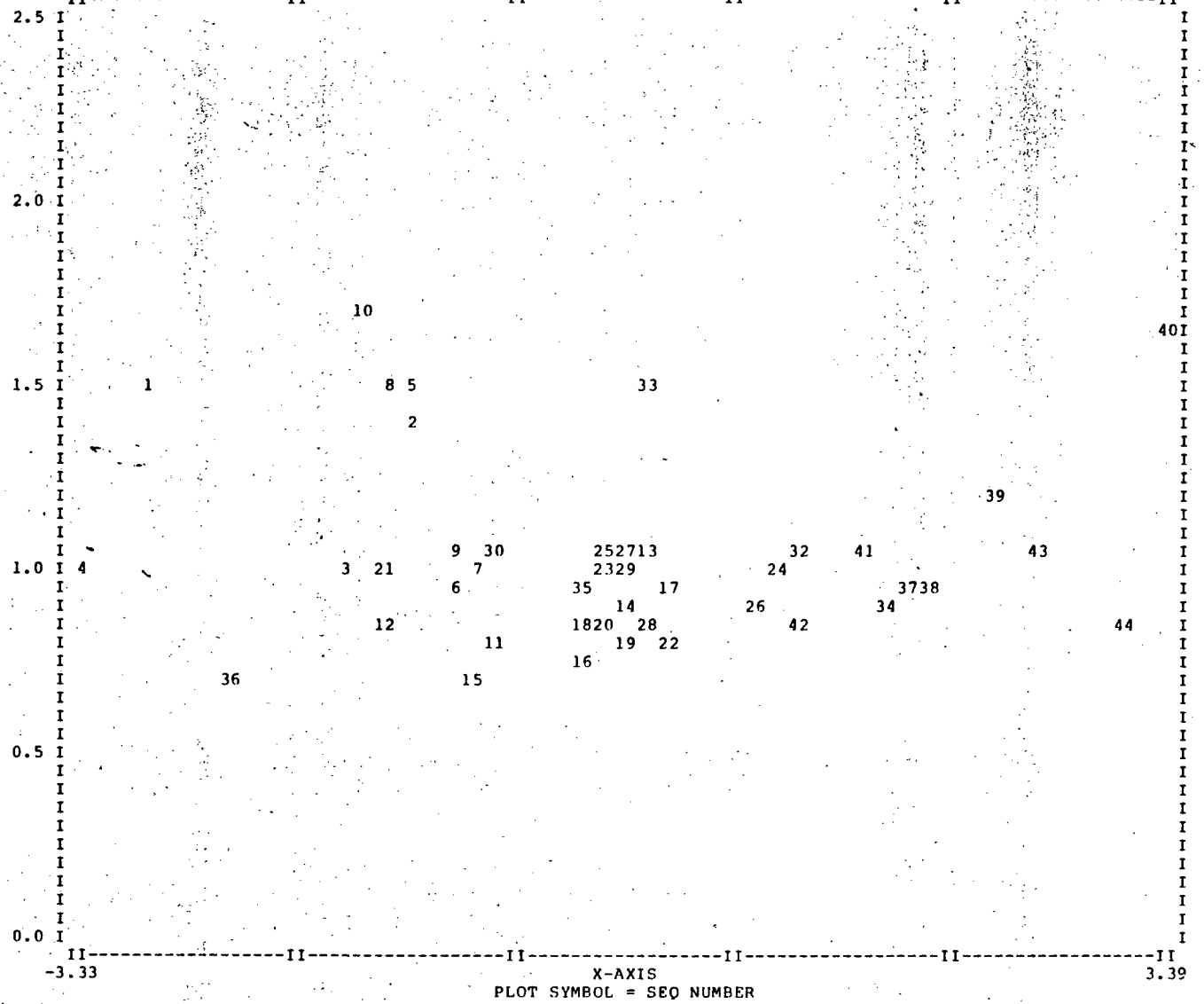
SEQ NUM	ITEM NAME	ITEM CHARACTERISTIC CURVE 1ST GROUP	2ND GROUP	3RD GROUP	4TH GROUP	5TH GROUP	6TH GROUP	DEPARTURE FROM EXPECTED ICC 1ST GROUP	2ND GROUP	3RD GROUP	4TH GROUP	5TH GROUP	6TH GROUP	FIT MEAN SQUARE WITHN GROUP	BETWN GROUP	TOTAL	DISC INDX	POINT BISER
1	C1	0.93	1.00	1.00	0.98	1.00	0.97	-0.01	0.03	0.02	-0.01	0.01	-0.02	1.49	2.41	1.51	0.69	0.09
2	C2	0.84	0.83	0.89	0.93	0.96	0.94	0.09	-0.02	-0.01	0.00	0.01	-0.03	1.45	1.13	1.44	0.66	0.15
3	C3	0.91	0.88	0.91	0.93	0.96	0.99	0.09	-0.02	-0.02	-0.02	-0.01	0.00	1.05	0.72	1.05	0.73	0.13
4	C4	1.00	0.95	0.98	1.00	0.98	1.00	0.04	-0.03	-0.01	0.01	-0.01	0.00	1.04	1.14	1.04	1.97	0.04
5	C5	0.82	0.86	0.98	0.91	0.94	0.93	0.06	-0.00	0.08	-0.03	-0.01	-0.05	1.48	2.44	1.50	0.56	0.14
6	C6	0.76	0.81	0.84	0.93	0.92	0.97	0.06	-0.01	-0.02	0.02	-0.02	0.00	1.01	0.30	0.99	0.89	0.23
7	C7	0.71	0.81	0.86	0.81	0.96	0.96	0.05	0.02	0.02	-0.08	0.03	-0.00	1.00	0.93	1.00	0.87	0.26
8	C8	0.76	0.88	0.91	0.95	0.98	0.97	-0.02	0.01	0.00	0.02	0.02	-0.01	1.55	0.22	1.52	1.03	0.26
9	C9	0.73	0.83	0.86	0.93	0.92	0.96	0.03	0.02	-0.00	0.02	-0.02	-0.01	1.07	0.22	1.06	0.85	0.26
10	C10	0.82	0.86	0.91	1.00	0.96	0.97	0.02	-0.03	-0.01	0.05	-0.00	-0.01	1.77	0.67	1.75	0.93	0.20
11	C11	0.56	0.79	0.86	0.81	0.94	1.00	-0.07	0.03	0.04	-0.07	0.02	0.05	0.80	1.50	0.82	1.12	0.36
12	C12	0.71	0.93	0.91	0.95	0.94	0.99	-0.05	0.07	0.01	0.02	-0.02	0.01	0.87	0.63	0.86	1.08	0.28
13	C13	0.40	0.64	0.75	0.77	0.78	0.87	-0.02	0.06	0.09	0.01	-0.05	-0.03	1.10	0.78	1.09	0.85	0.31
14	C14	0.47	0.55	0.55	0.84	0.86	0.93	0.04	-0.03	-0.12	0.07	0.03	0.03	0.92	1.16	0.93	1.10	0.40
15	C15	0.60	0.81	0.86	0.91	0.94	1.00	-0.07	0.01	0.01	0.01	0.01	0.04	0.74	0.86	0.74	1.16	0.40
16	C16	0.40	0.64	0.73	0.84	0.88	0.99	-0.10	-0.01	-0.00	0.03	0.02	0.06	0.78	1.25	0.79	1.32	0.46
17	C17	0.49	0.50	0.55	0.72	0.80	0.93	0.10	-0.04	-0.08	-0.01	-0.00	0.04	0.99	1.00	0.99	1.00	0.37
18	C18	0.44	0.60	0.73	0.86	0.94	0.94	-0.06	-0.06	-0.01	0.05	0.07	0.02	0.88	0.91	0.88	1.24	0.41
19	C19	0.33	0.57	0.61	0.81	0.88	0.97	-0.10	-0.02	-0.06	0.05	0.05	0.07	0.82	1.62	0.83	1.38	0.48
20	C20	0.42	0.62	0.68	0.86	0.94	0.94	-0.07	-0.03	-0.04	0.05	0.08	0.02	0.85	1.04	0.85	1.26	0.44
21	C21	0.76	0.93	0.93	0.86	0.96	0.99	-0.01	0.07	0.03	-0.08	0.00	0.01	1.00	1.27	1.00	0.94	0.24
22	C22	0.16	0.60	0.73	0.74	0.90	0.90	-0.24	0.05	0.09	0.01	0.10	0.01	0.80	3.20	0.85	1.37	0.52
23	C23	0.58	0.76	0.57	0.79	0.86	0.93	0.09	0.12	-0.16	-0.02	-0.00	0.00	1.03	1.87	1.05	0.83	0.27
24	C24	0.13	0.45	0.48	0.65	0.67	0.74	-0.10	0.09	0.03	0.08	0.02	-0.05	0.99	1.39	1.00	0.98	0.41
25	C25	0.53	0.57	0.75	0.88	0.84	0.89	0.05	-0.06	0.03	0.08	-0.02	-0.04	1.07	0.97	1.07	0.87	0.32
26	C26	0.29	0.21	0.52	0.74	0.65	0.86	0.02	-0.19	0.03	0.14	-0.04	0.04	0.92	2.19	0.95	1.16	0.46
27	C27	0.36	0.60	0.75	0.72	0.90	0.94	-0.09	-0.01	0.06	-0.06	0.06	0.03	1.05	1.13	1.05	1.23	0.44
28	C28	0.33	0.62	0.73	0.63	0.92	0.91	-0.09	0.05	0.07	-0.13	0.10	0.01	0.88	1.99	0.90	1.15	0.44
29	C29	0.49	0.55	0.73	0.72	0.92	0.87	0.05	-0.05	0.05	-0.05	0.09	-0.04	1.01	1.17	1.01	0.91	0.34
30	C30	0.64	0.71	0.84	0.95	0.90	0.96	0.01	-0.05	0.02	0.07	-0.02	0.00	1.10	0.61	1.09	1.01	0.33
31	C31	0.58	0.71	0.91	0.88	0.94	0.97	-0.05	-0.05	0.08	0.00	0.02	0.02	0.83	0.82	0.83	1.17	0.39
32	C32	0.36	0.24	0.39	0.56	0.57	0.77	0.15	-0.09	-0.03	0.03	-0.05	0.00	1.07	1.65	1.08	0.84	0.36
33	C33	0.62	0.64	0.66	0.72	0.73	0.76	0.23	0.09	0.02	-0.02	-0.07	-0.13	1.44	5.59	1.53	0.23	0.09
34	C34	0.04	0.19	0.23	0.42	0.51	0.74	-0.09	-0.03	-0.06	0.03	0.03	0.08	0.91	1.35	0.92	1.37	0.48
35	C35	0.49	0.74	0.77	0.81	0.82	0.97	-0.03	0.06	0.02	-0.01	-0.06	0.04	0.95	0.93	0.95	1.05	0.36
36	C36	0.89	0.95	0.95	0.95	1.00	1.00	-0.00	0.01	-0.00	-0.02	0.02	0.01	0.72	0.49	0.71	0.99	0.21
37	C37	0.13	0.19	0.25	0.30	0.47	0.67	0.01	-0.01	-0.01	-0.06	0.01	0.03	1.00	0.26	0.98	1.07	0.42
38	C38	0.02	0.24	0.27	0.33	0.41	0.63	-0.09	0.06	0.03	-0.01	-0.02	0.02	0.95	0.97	0.96	1.10	0.42
39	C39	0.16	0.21	0.18	0.28	0.22	0.44	0.08	0.09	0.01	0.03	-0.10	-0.07	1.22	2.35	1.24	0.56	0.23
40	C40	0.11	0.07	0.09	0.09	0.18	0.21	0.08	0.02	0.02	-0.02	0.03	-0.09	1.63	2.88	1.65	0.51	0.12
41	C41	0.18	0.36	0.30	0.42	0.45	0.77	0.01	0.09	-0.05	-0.03	-0.10	0.06	1.09	1.17	1.10	0.99	0.37
42	C42	0.16	0.26	0.39	0.53	0.73	0.83	-0.06	-0.08	-0.04	-0.00	0.10	0.05	0.88	1.17	0.88	1.32	0.51
43	C43	0.09	0.07	0.14	0.21	0.27	0.46	0.03	-0.03	-0.00	0.00	-0.01	0.00	1.08	0.24	1.06	0.97	0.33
44	C44	0.02	0.02	0.09	0.16	0.10	0.39	-0.01	-0.04	0.01	0.03	-0.08	0.05	0.85	0.95	0.85	1.17	0.35
SCORE RANGE		14-24	25-27	28-30	31-33	34-35	36-43	N= 45	42	44	43	49	70	287	6	293	DEG OF FROM	
MEAN ABILITY		-0.14	0.49	0.86	1.33	1.71	2.50	PLUS=TOO MANY RIGHT MINUS=TOO MANY WRONG						0.08	0.58	0.08	STD ERROR	
GROUP MN SQ		2.2	0.9	0.7	0.9	1.0	2.0	SE = 0.2										
SD(MN SQ)		3.2	1.3	1.3	1.3	1.1	3.3	EXPECT 1.4										

SERIAL ORDER						DIFFICULTY ORDER						FIT ORDER					
SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	POINT BISER	I
1	C1	-2.92	0.69	1.51	I	4	C4	-3.33	1.97	1.04	I	36	C36	-2.30	0.99	0.71	0.21	I
2	C2	-1.27	0.66	1.43	I	1	C1	-2.92	0.69	1.51	I	15	C15	-0.87	1.16	0.73	0.40	I
3	C3	-1.71	0.73	1.04	I	36	C36	-2.30	0.99	0.71	I	16	C16	-0.13	1.32	0.79	0.46	I
4	C4	-3.33	1.97	1.04	I	3	C3	-1.71	0.73	1.04	I	11	C11	-0.66	1.12	0.81	0.36	I
5	C5	-1.31	0.56	1.50	I	10	C10	-1.55	0.93	1.74	I	31	C31	-0.69	1.17	0.82	0.39	I
6	C6	-1.00	0.89	0.99	I	8	C8	-1.40	1.03	1.51	I	19	C19	0.14	1.38	0.83	0.48	I
7	C7	-0.84	0.87	1.00	I	21	C21	-1.35	0.94	1.00	I	22	C22	0.31	1.37	0.84	0.52	I
8	C8	-1.40	1.03	1.51	I	12	C12	-1.35	1.08	0.86	I	20	C20	-0.11	1.26	0.85	0.44	I
9	C9	-1.00	0.85	1.05	I	5	C5	-1.31	0.56	1.50	I	44	C44	3.23	1.17	0.85	0.35	I
10	C10	-1.55	0.93	1.74	I	2	C2	-1.27	0.66	1.43	I	12	C12	-1.35	1.08	0.86	0.28	I
11	C11	-0.66	1.12	0.81	I	6	C6	-1.00	0.89	0.99	I	42	C42	1.17	1.32	0.88	0.51	I
12	C12	-1.35	1.08	0.86	I	9	C9	-1.00	0.85	1.05	I	18	C18	-0.15	1.24	0.88	0.41	I
13	C13	0.18	0.85	1.09	I	15	C15	-0.87	1.16	0.73	I	28	C28	0.20	1.15	0.89	0.44	I
14	C14	0.16	1.10	0.92	I	7	C7	-0.84	0.87	1.00	I	34	C34	1.77	1.37	0.92	0.48	I
15	C15	-0.87	1.16	0.73	I	31	C31	-0.69	1.17	0.82	I	14	C14	0.16	1.10	0.92	0.40	I
16	C16	-0.13	1.32	0.79	I	30	C30	-0.69	1.01	1.08	I	26	C26	0.89	1.16	0.94	0.46	I
17	C17	0.33	1.00	0.99	I	11	C11	-0.66	1.12	0.81	I	35	C35	-0.24	1.05	0.95	0.36	I
18	C18	-0.15	1.24	0.88	I	35	C35	-0.24	1.05	0.95	I	38	C38	2.01	1.10	0.95	0.42	I
19	C19	0.14	1.38	0.83	I	18	C18	-0.15	1.24	0.88	I	37	C37	1.89	1.07	0.98	0.42	I
20	C20	-0.11	1.26	0.85	I	16	C16	-0.13	1.32	0.79	I	6	C6	-1.00	0.89	0.99	0.23	I
21	C21	-1.35	0.94	1.00	I	20	C20	-0.11	1.26	0.85	I	17	C17	0.33	1.00	0.99	0.37	I
22	C22	0.31	1.37	0.84	I	23	C23	-0.11	0.83	1.04	I	24	C24	1.06	0.98	1.00	0.41	I
23	C23	-0.11	0.83	1.04	I	25	C25	-0.06	0.87	1.07	I	21	C21	-1.35	0.94	1.00	0.24	I
24	C24	1.06	0.98	1.00	I	27	C27	0.08	1.23	1.05	I	7	C7	-0.84	0.87	1.00	0.26	I
25	C25	-0.06	0.87	1.07	I	29	C29	0.10	0.91	1.00	I	29	C29	0.10	0.91	1.00	0.34	I
26	C26	0.89	1.16	0.94	I	19	C19	0.14	1.38	0.83	I	4	C4	-3.33	1.97	1.04	0.04	I
27	C27	0.08	1.23	1.05	I	14	C14	0.16	1.10	0.92	I	3	C3	-1.71	0.73	1.04	0.13	I
28	C28	0.20	1.15	0.89	I	13	C13	0.18	0.85	1.09	I	23	C23	-0.11	0.83	1.04	0.27	I
29	C29	0.10	0.91	1.00	I	28	C28	0.20	1.15	0.89	I	9	C9	-1.00	0.85	1.05	0.26	I
30	C30	-0.69	1.01	1.08	I	33	C33	0.29	0.23	1.52	I	43	C43	2.68	0.97	1.05	0.33	I
31	C31	-0.69	1.17	0.82	I	22	C22	0.31	1.37	0.84	I	27	C27	0.08	1.23	1.05	0.44	I
32	C32	1.22	0.84	1.08	I	17	C17	0.33	1.00	0.99	I	25	C25	-0.06	0.87	1.07	0.32	I
33	C33	0.29	0.23	1.52	I	26	C26	0.89	1.16	0.94	I	30	C30	-0.69	1.01	1.08	0.33	I
34	C34	1.77	1.37	0.92	I	24	C24	1.06	0.98	1.00	I	32	C32	1.22	0.84	1.08	0.36	I
35	C35	-0.24	1.05	0.95	I	42	C42	1.17	1.32	0.88	I	41	C41	1.52	0.99	1.09	0.37	I
36	C36	-2.30	0.99	0.71	I	32	C32	1.22	0.84	1.08	I	13	C13	0.18	0.85	1.09	0.31	I
37	C37	1.89	1.07	0.98	I	41	C41	1.52	0.99	1.09	I	39	C39	2.44	0.56	1.24	0.23	I
38	C38	2.01	1.10	0.95	I	34	C34	1.77	1.37	0.92	I	2	C2	-1.27	0.66	1.43	0.15	I
39	C39	2.44	0.56	1.24	I	37	C37	1.89	1.07	0.98	I	5	C5	-1.31	0.56	1.50	0.14	I
40	C40	3.39	0.51	1.65	I	38	C38	2.01	1.10	0.95	I	8	C8	-1.40	1.03	1.51	0.26	I
41	C41	1.52	0.99	1.09	I	39	C39	2.44	0.56	1.24	I	1	C1	-2.92	0.69	1.51	0.09	I
42	C42	1.17	1.32	0.88	I	43	C43	2.68	0.97	1.05	I	33	C33	0.29	0.23	1.52	0.09	I
43	C43	2.68	0.97	1.05	I	44	C44	3.23	1.17	0.85	I	40	C40	3.39	0.51	1.65	0.12	I
44	C44	3.23	1.17	0.85	I	40	C40	3.39	0.51	1.65	I	10	C10	-1.55	0.93	1.74	0.20	I
	MEAN	0.00	1.01	1.05														
	S.D.	1.51	0.29	0.25		CORRELATION		DIFF*DISC= -0.10		DIFF*MNSQ= -0.07				DISC*MNSQ= -0.62				

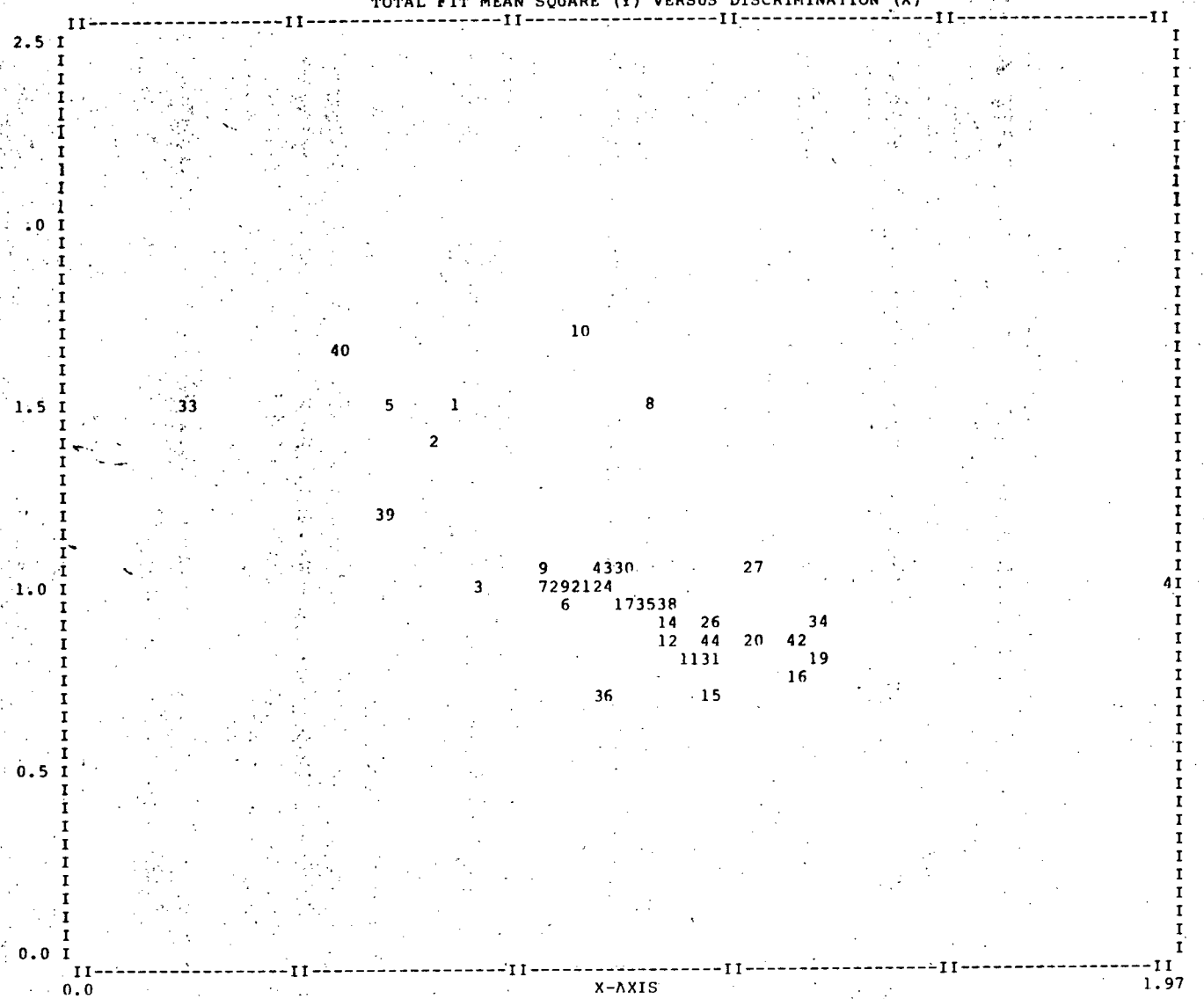
1964 computation test calibration

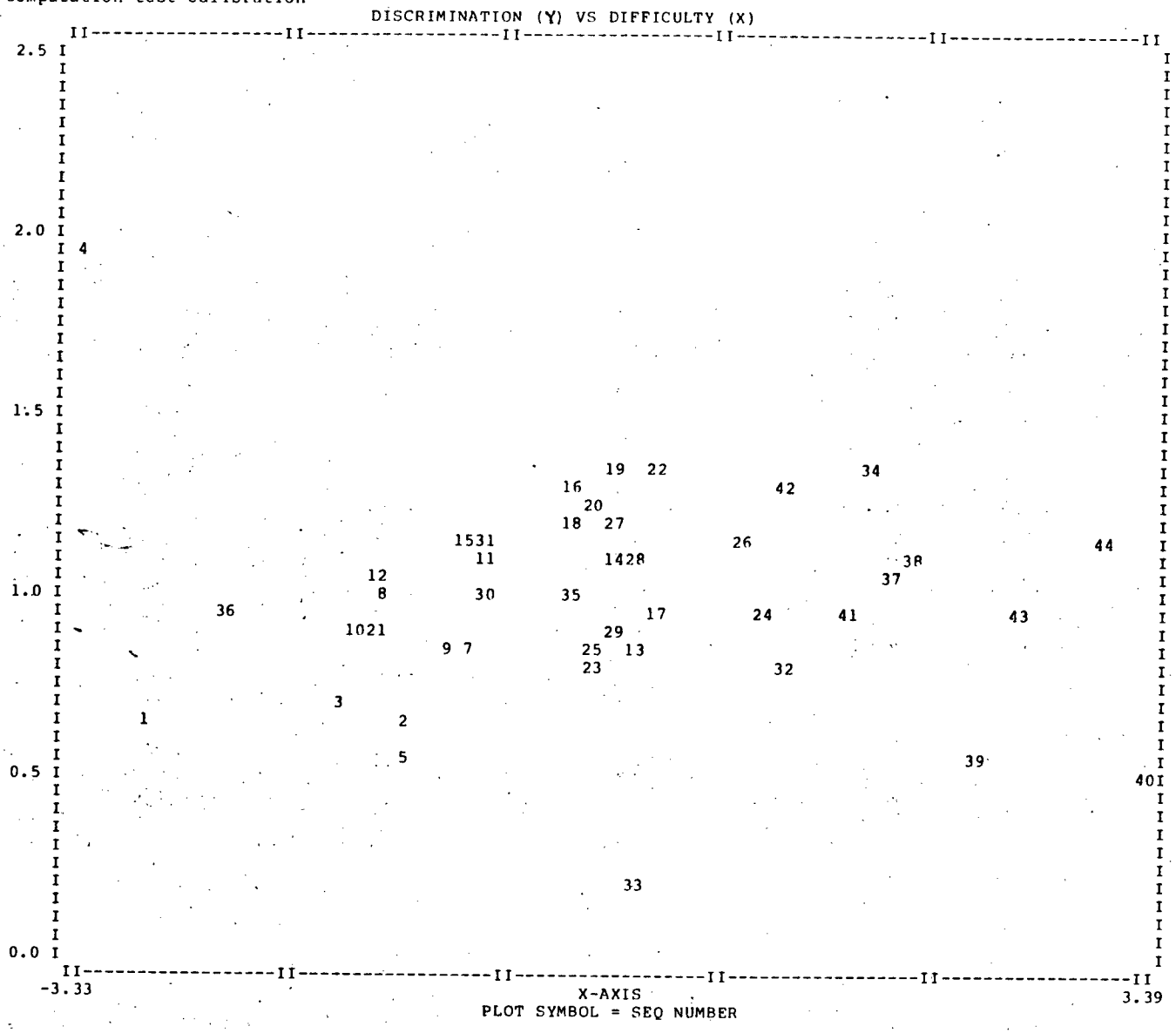


TOTAL FIT MEAN SQUARE (Y) VERSUS DIFFICULTY (X)



TOTAL FIT MEAN SQUARE (Y) VERSUS DISCRIMINATION (X)




DISCRIMINATION (Y) VS DIFFICULTY (X)
2.5
2.0
1.5
1.0
0.5
0.0
-3.33
3.39
X-AXIS
PLOT SYMBOL = SEQ NUMBER

APPENDIX D

CORRESPONDENCE

THE UNIVERSITY OF BRITISH COLUMBIA
2075 WESBROOK MALL
VANCOUVER, B.C., CANADA
V6T 1W5

FACULTY OF EDUCATION

March 26, 1979.

Mr. D. R. Sutherland,
Superintendent,
School District No. 77,
Box 339,
Summerland, B. C.,
V0H 1Z0.

Dear Mr. Sutherland:

I am writing to enlist your support for an important research project which Mr. Thomas O'Shea, my research assistant, and I are conducting this spring. The study is designed to obtain information concerning the changes in achievement levels in Grade 7 mathematics from 1964 to the present.

In 1964 and 1970, the Ministry of Education administered standardized achievement tests in mathematics to Grade 7 students throughout the province, and these data have been made available to us. Preliminary analysis indicates that, although some decline occurred, changes appear to be confined to specific content areas within the curriculum. We propose to administer these same tests to a sample of present Grade 7 mathematics classes and to compare and contrast the resultant achievement patterns. The sample has been constructed in such a way as to minimize the chances of a given class being asked to participate in any Ministry-sponsored projects this spring. The list of schools from your district whose participation is requested is attached.

Administration of the tests requires two forty-five minute class periods for each classroom selected, preferably on consecutive days. Detailed instructions, administrative directions, and test materials will be mailed directly to the principals of the schools involved. We hope to have the teachers administer the tests in the week of April 23-27. Strict confidentiality with respect to students, schools, and districts will be observed. The study will result in comparisons in performance across time on a province-wide basis only. A summary of the findings will be sent to you before the commencement of the 1979-1980 school year.

Permission is granted for Dr. David Robitaille and Mr. Thomas O'Shea of the Faculty of Education, University of British Columbia, to contact the following schools with regard to the administration of standardized tests in arithmetic to Grade Seven students:

_____

_____

_____

_____

_____

_____ _____

Superintendent Date

_____

School District

THE UNIVERSITY OF BRITISH COLUMBIA
2075 WESBROOK MALL
VANCOUVER, B.C., CANADA
V6T 1W5

FACULTY OF EDUCATION

April 18, 1979.

Principal,
T. M. Roberts School,
10 Wattsville St.,
Cranbrook, B.C.,
V1C 2A2.

Dear Sir/Madam:

The superintendent of your district has given me permission to contact you in order to enlist your help in carrying out an important research project which Mr. Thomas O'Shea, a doctoral student at U. B. C., and I are conducting this spring. The project, which has been approved by our Behavioural Sciences Screening Committee for Research Involving Human Subjects, is being undertaken as a doctoral dissertation in the Department of Mathematics Education at U. B. C. Financial support has been provided through a grant from the Educational Research Institute of British Columbia.

In 1964 and 1970, the Ministry of Education administered standardized tests in mathematics to Grade 7 students throughout the province, and these data have been made available to us. Preliminary analysis, using a new statistical model, indicates that, although some decline in performance occurred, changes appear to be confined to specific content areas within the mathematics curriculum, for example, operations on common fractions. We propose to administer these same tests to a sample of present Grade 7 mathematics classes and to compare and contrast the resultant achievement patterns. Your school has been selected as part of a stratified random sample, based on geographic region and school size, of over 60 schools in more than 30 districts throughout the province. The sample has been constructed in such a way as to minimize the chance that your school will be asked by the Ministry of Education to participate in any projects this spring.

We believe that the results will be of interest to you and your teachers by helping to identify continuing strengths or potential weaknesses within the elementary mathematics curriculum. Strict confidentiality with respect to students, schools, and districts will be observed. The study will result in comparisons in performance across time on a province-wide basis only. A summary of the findings will be sent to your district superintendent before the beginning of the 1979-80 school year.

THE UNIVERSITY OF BRITISH COLUMBIA
2075 WESBROOK MALL
VANCOUVER, B.C., CANADA
V6T 1W5

FACULTY OF EDUCATION

April 18, 1979.

To the Teacher/Test Administrator:

The district superintendent, and your principal, have given us permission to ask for your help in conducting an important research project in British Columbia schools. A letter to your principal contains information on the background and purpose of the project. Briefly, the study is designed to yield information concerning changes in Grade 7 mathematics achievement from 1964 to the present. Claims of a general decline in performance have not been substantiated by our preliminary analysis of data from 1964 to 1970. However, some decline in specific content areas within the mathematics curriculum seems to be indicated. We hope to be able to identify particular topics on which the performance of present-day students is different from that of students in 1964 or 1970.

Your Grade 7 class has been selected as part of a random sample of over 60 classes across the province. The study is designed so that conclusions can be drawn regarding the provincial Grade 7 population only. No comparisons are possible of individuals, classes, schools, or districts. Student names and school district numbers are necessary for clerical purposes only. Once the data have been transferred from the test papers, no identifying codes will be retained. The names of students are required on the test papers only to ensure that the two parts of the test which each student writes may be matched. If you prefer to use some other means of identifying papers which will accomplish the same purpose, please feel free to do so.

The tests to be administered are identical to those used in the 1964 and 1970 testing programs. They are the Arithmetic Reasoning and Arithmetic Computation tests from the Stanford Achievement Test. Specific instructions regarding administration procedures are contained in the document Directions for Administration which is enclosed. Please follow these closely since they are based on the original directions for giving the tests. The administration of each test requires about one class period of 45 minutes. It would be preferable to give the tests on two consecutive days,

ARITHMETIC ACHIEVEMENT TESTS

Directions for Administration

The teacher should become thoroughly familiar with all of the following directions before giving the tests.

General Directions

1. Before beginning each test, see that the desks are cleared and that each pupil has an eraser and one or two sharpened pencils, preferably with very soft leads. Pens should not be used. A supply of extra pencils should be at hand. Scratch paper should be provided.

2. A natural classroom situation should be retained as far as possible. Provision should be made to ensure quiet and freedom from interruptions of any kind.

3. The teacher should take pains to ensure that the pupils understand what they are to do in each test and how they are to record their answers. This can be done best by reading the directions verbatim and supplementing with explanations as questions from the pupils indicate need. When doing this, the teacher should not give help on specific test questions, but may fully clarify the directions.

4. After a test has been started, the teacher should circulate about the room to see that instructions are being followed. When they are not, clarify for the individual pupil but do not disturb the entire class.

5. Adhere to the time limits. A watch with a second-hand should be used in order to guarantee uniformity of time.

6. Following is the schedule for the arithmetic tests:

FIRST SITTING

Distributing booklets, reading directions, etc. 5 min.
Test 1: Arithmetic Reasoning ..... Work time 35 min.
Total ..... 40 min.

SECOND SITTING

Distributing booklets, reading directions, etc. 5 min.
Test 2: Arithmetic Computation ..... Work time 35 min.
Total ..... 40 min.

If all pupils finish a test before the recommended time has elapsed, time may be called.

7. Under no conditions should a test be started unless sufficient time is available to complete it.

Specific Directions

To administer each test, say to the pupils:

> "This is a test to show how much you have learned in arithmetic. When you get your test booklet, do not write on it or open it until I tell you to." (Be sure pupils do not open booklets.)

Pass out the test booklets. Then say:

> "Now look at the front page where it says 'Name'. (Point to the proper place.) Write your first and last names here. Be sure to write plainly. (Pause.) In the second line, write your school district number, and the name of your school. (Pause.) In the third line, write the date."

After the blanks have been filled in, continue:

> "Now listen carefully. You must do your best, but I do not expect you to be able to answer all the questions. Do not start until I say 'BEGIN' and when I say 'STOP' put your pencil right down. If you break your pencil, hold up your hand and I will give you another. After we have begun you must not ask questions." (Continue with the directions for the first test, given below.)

First Sitting - Arithmetic Reasoning

> "Now open your booklet, Arithmetic Reasoning. Fold the page back, like this, so that only the first page of questions is showing." (Demonstrate.)
>
> "Look at the top of the page, where it says 'Directions'. (Hold up a booklet and point to the proper place.)
>
> "They say: 'Work an example, and then compare your answer with the answers which follow it. If your answer is one of those given, mark the answer space that has the same letter as your answer. Sometimes the correct answer is not given. If you do not find the correct answer, mark the space under the letter for 'not given'. Now look at the samples." (Hold up a booklet and point to the sample exercises.)
>
> "The first sample says: 'How many are 3 balls and 4 balls? 3 4 7 12 not given'. Which is the correct answer?" (Wait for the class to answer.)
>
> "Yes, the answer is '7'. The letter beside the '7' is 'c', so the answer space under the letter 'c' has been filled in. Now study the second sample. What is the answer?" (Pause for reply.)

"Yes, '5' is the correct answer to this problem, but it is not listed among the choices. Hence, the correct answer for this example is the answer 'not given', so you fill in the space under the letter 'j'."

"For each example on this page and on the next page, decide which is the correct answer, and fill in the answer space below the letter which represents the answer you have chosen. Use the scratch paper you were given to figure on."

"Begin with Question No. 1 and answer as many questions as you can. When you finish the first two pages, go right on to Part Two, on the last page. When you finish the last page, go back and check your answers. READY. BEGIN!" (Record the starting time. Add twenty-five minutes and ten minutes.)

After twenty-five minutes, say:

"If you have not already started work on Part Two on the last page, do so now." (Make sure the pupils do this.) Then say: "Go on working."

After an additional ten minutes - i.e., at the end of thirty-five minutes - say:

"STOP! Put your pencil down."

Collect the test booklets immediately. (The first sitting ends here.)

Second Sitting - Arithmetic Computation

Distribute the test booklets.

Have the pupils complete the title page as in the first sitting (see 'Specific Directions').

Continue with:

"Now open your booklet, Arithmetic Computation. Fold the booklet back, like this, so that only the first page of questions is showing. (See that all do this correctly.)

"Look at the top of the page, where it says 'Directions'. (Hold up a booklet and point to the proper place.)

"They say: 'Work each example. Then compare your answer with the answers given at the right of the example. If your answer is one of those given, mark the answer space that has the same letter as your answer. Sometimes the correct answer is not given. If the correct answer is not given, mark the answer space under the letter for 'not

given'. Look carefully at each example to see what it tells you to do. If you need to do any figuring, use a separate sheet of paper.'"

"Now begin with Question No. 1 and answer as many questions on this page and the next two pages as you can. When you finish the last page, go back and check your work. Are there any questions about what you are to do? (Pause.) READY. BEGIN!" (Record the starting time and add thirty-five minutes.)

After thirty-five minutes, say:

"STOP! Close your booklet and put your pencil down."

Collect the test booklets immediately. (The second sitting ends here.)