JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH 50

RITVA KOPONEN

AN ITEM ANALYSIS OF TESTS IN MATHEMATICS APPLYING LOGISTIC TEST MODELS

JYVÄSKYLÄN YLIOPISTO, JYVÄSKYLÄ 1983

JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH 50

RITVA KOPONEN

AN ITEM ANALYSIS OF TESTS IN MATHEMATICS APPLYING LOGISTIC TEST MODELS

esitetään Jyväskylän yliopiston kasvatustieteiden
tiedekuntaneuvoston suostumuksella julkisesti tarkastettavaksi
Blomstedtin salissa lokakuun 21. päivänä 1983 klo 12.

JYVÄSKYLÄN YLIOPISTO, JYVÄSKYLÄ 1983

JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH 50

RITVA KOPONEN

AN ITEM ANALYSIS OF TESTS IN MATHEMATICS APPLYING LOGISTIC TEST MODELS

JYVÄSKYLÄN YLIOPISTO, JYVÄSKYLÄ 1983

URN:ISBN:978-951-39-9938-4
ISBN 978-951-39-9938-4 (PDF)
ISSN 0075-4625

Jyväskylän yliopisto, 2024

ISBN 951-678-971-4
ISSN 0075-4625

Jyväskylän yliopiston monistuskeskus ja
Kirjapaino Kari Ky, Jyväskylä 1983

ABSTRACT

Koponen, Ritva
An item analysis of tests in Mathematics applying logistic test models /
Ritva Koponen. – Jyväskylä: Jyväskylän yliopisto, 1983. – 187 p. –
(Jyväskylä Studies in Education, Psychology and Social Research,
ISSN 0075-4625; 50)
ISBN 951-678-971-4
Matematiikan kokeiden osioanalyysi logistisia testimalleja käyttäen.

Diss.

The current study investigates learning difficulties in some areas of Mathematics from the point of view of the structure of the learning task. The study may be divided into three parts, each dealing with the different aspects of the tests. In the first part all errors made by the pupils are analysed, a flow chart being made of the most prelevant errors in test of equations and inequations in primary level. The second part concentrates on the application of the simple logistic model to the data. In addition, the results were compared to results obtained by traditional test theory. The third part used the linear logistic test model to examine the structure of the items. As a result the difficult operations were found. An estimate for the difficulty parameter was calculated for each operation.

Rasch model, simple logistic model, linear logistic model

PREFACE

The initial impulse for this study came through Associate Professor Paavo Malinen's trip to the Kiel IPN institute at the beginning of 1978. On his return, Professor Malinen told me of his interest in the Rasch models being used there. He brought along with him some research studies and journals, suggesting that I might study them as basic material for a doctoral thesis. The models looked interesting and I began studies in this field. By the spring of 1979 I had reached the stage of collecting empirical data. The correction of the tests of over 2000 pupils and the coding of the errors proved to be a time-consuming task. During their summer vacations, Mrs. Marja Autio and Mrs. Liisa Salminen offered to help me in this aspect of the thesis. I extend my heartfelt gratitude to them for their kindness. I should also like to thank all those, both teachers and students and pupils who made it possible for me to collect my data.

In 1980 I was given a grant by the Australian Government to study at the University of Western Australia in Perth under the guidance of Dr. David Andrich and Dr. Graham Douglas, both eminent scholars in the field of modern test theory. I should like to extend my sincere thanks to them for their excellent guidance at their department. Despite the fact that they had their own tuition as well as many administrative tasks, they were always at hand with time, kind words and encouragement, whenever I troubled them with questions

and my own solutions. I also owe a debt of gratitude to my friends within the university and also outside it for making my stay in a foreign country a pleasurable one. My family and friends in Finland, too, deserve my thanks for not forgetting me, even though the distance between us was long.

Valuable assistance was given to me in the computer processing of the results by Mr. Alan Lyne, Mr. Sakari Valkonen and Mr. Arto Ylipiha. Furthermore assistance was kindly given me by the staff of the Computing Centre at the University of Jyväskylä. My thanks are also due to the staff of the university library for prompt and expert help whenever it was required.

A travel grant from the University of Jyväskylä made it possible in the summer of 1979 to undertake a trip to Gothenburg, to see the work being done by Dr. Jan-Eric Gustafsson and to Kiel to discuss problems with my research program with Dr. Jürgen Rost. Both aforementioned gentlemen made available to me computer programs on latent trait theory, which were not available to our university at the time. My sincere thanks to all the aforementioned.

The last stages of the study were subsidised by the University of Jyväskylä in the form of the Ellen and Artturi Nyyssönen Fund grant given me in May, 1982. My thanks are due for linguistic corrections to Mr. David Wilson M.A. and Mr. Andres Perendi M.A. The preliminary examination of the thesis was done by Professor Raimo Konttinen and Associate Professor Jarkko Leino, to whom I am indebted for excellent suggestions for corrections. For constructive criticism I owe a vote of thanks to the leader of the research seminars, Professor Veikko Heinonen. For final typing of the thesis I am indebted to Miss Marja-Leena Vartiainen and to Mrs. Marja-Liisa Helimäki. I am further grateful to the University of Jyväskylä for including my study in their publications series.

There are still more people, who shall remain unmentioned, but who, in this period of five years, have in one way or

another, helped me to attain my goal. My sincere gratitude goes to them all.

Jyväskylä,
April, 1983 RITVA KOPONEN

CONTENTS

1. THE STRUCTURE OF THE RESEARCH

1.1. Introduction

During the last twenty years the developments in psychometrics have led to the use of new concepts as an alternative to traditional test theory. This has permitted the study of new kinds of issues associated with mental and achievement testing. In particular, components and structures of tasks found in test items can be analysed using probabilistic models.

Latent trait theory has been developed in different directions and the name of the Danish mathematician and statistician Georg Rasch is closely connected to one of these. His book (Rasch 1960, 1980) is still one of the basic textbooks in this area. In addition to the simple logistic model (SLM) more complex models, like the linear logistic test model (LLTM) have been used in research. Both of them belong to the family of Rasch models: the former is the basic model for measuring item difficulty and ability of persons, the latter can be used for analysing the structure of a task. In the present study, new test theoretical approaches were applied to learn more of the process in solving mathematical problems among 1st - 9th graders in the Finnish comprehensive school in the area of equations and inequations. The purpose was to find learning difficulties in each grade and to give suggestions for improving the Mathematics syllabus in this area, particularly in basic skills.

In the case of the SLM it is a question of only one item parameter: difficulty of an item. In the LLTM this single parameter is seen to be composed of a number of more elementary parameters. In different terms elementary parameters appear in different combinations. The SLM assumes that if a person can solve correctly an item i with a given probability he can solve the items which are easier than i with an even greater probability. A goodness of fit test has been constructed for checking that responses conform to the model.

In the present stydy, both the SLM and the LLTM were used. First, the SLM was applied to compare the relative difficulties of the items. If the model holds, the order of the relative difficulties should remain the same for different subsets of persons. The fact first having been checked that the items fit the SLM, then the LLTM was used for finding the structure of items. The LLTM was used to find out which parts of the item (operations) had an effect on the item difficulty. Elementary parameters have been selected after analysing and classifying all errors of pupils. For a test of about 30 items, 8 or 9 elementary parameters were used for finding out the special difficulties in solving items. For the items which do not fit the model an explanation for misfit has been presented. A new structure of elementary parameters would correct them and all items would fit the model. In our data there were only a few items which did not fit the model and that is why new computations have not been made even if each item which is outside the lines in the graphical test of fit has been analysed separately.

1.2. The tasks of the research

In Finland we have some important researches from the 1970's (e.g. Puro 1974, 1977a, 1977b) concerning learning difficulties and developing material for remedial instruction in Mathematics at the junior level of comprehensive

school. In this research linear logistic test model (LLTM) has been used for finding out learning difficulties. Simultaneously the whole junior level of comprehensive school can be studied and we can follow what kind of difficulties move from one grade to the next and what kind of difficult things pupil learn quite well before the next grade.

The tasks of the research are

- exploration of learning difficulties in Mathematics using error analysis
- following the development of skills using the simple logistic model
- exploration of errors using the linear logistic model.

A new point in the usage of the LLTM is that it is based on empirical error analysis. The common drawback in the use of the LLTM is the lack of the sound empirical of theoretical basis in the construction of the Q-matrix for finding estimates of elementary parameters (chapter 5).

This study is also a study of the Mathematics syllabus. It is focused to study the methods how to find points of difficulty in it. At first the tests are constructed according to Mathematics subject matter and the pupils are tested. The results are analysed using the SLM and the LLTM in finding out what are the topics which pupils have not mastered yet.

This research uses latent trait methods for evaluation of Mathematics curriculum. Their mode of action has been studied in the area of equations and inequations at the junior level of comprehensive school. No value judgement of the curriculum can be given.

A new point in this research is the longitudal method used in analysing tests according to the latent trait theory. Even if only the content area of equations is used, the main idea is to show how to study all contents of the Mathematics curriculum one at a time.

All the errors of the junior level of comprehensive school in the equation tests were analysed. Chapter 2 gives the results of the error analysis. They have been used in

constructing a set of elementary parameters for the use of the LLTM (chapter 5). Linking of elementary parameters makes it possible to compare all primary grades simultaneously.

The latent trait theory has been presented shortly emphasizing the most essential concepts. This has been done in chapter 3. The next step is to analyse the empirical test data using the SLM and traditional test theory (chapter 4). We can get the answers for example for the questions:

- Does the SLM hold in the data of items and persons?
- What are the misfitting items and persons and what are the reasons for misfit?
- What are the criteria for goodness of items in the traditional test theory?
- What are the differences in ways of measuring reliability in traditional and latent trait test theories?

2. EXPLORATION OF LEARNING DIFFICULTIES IN MATHEMATICS USING ERROR ANALYSIS

2.1. Research problems concerning errors

An important aspect of the teacher's job is to consider how children have been thinking when they have failed in computing or problem solving. Error analysis is one way of doing so. Combined with interviews it may give useful feedback from learning. This kind of analysis is needed particularly for

(1) diagnosing learning difficulties,
(2) finding individual instructions for teaching,
(3) developing the curriculum.

Since the 1920's research work has been done in analysing errors in school Mathematics (Radatz 1979). Studies are basically national and it is not easy to compare the results. Differences in teaching methods and curricula set limitations to finding a common base for the classification of errors. Only some studies can be found in this area in Finland. The first one was published thirty years ago (Lahti 1949) and it was one of the first doctoral dissertations in education. Some error analysis has been done later (Kaila 1971). Lahti collected his data in the period 1920 - 1940 when he taught Mathematics at school. Kaila has got her data from Malinen's doctoral dissertation (1969). The inability to use a computer has restricted Lahti's methods. This hindrance ceased to exist and it is worth studying errors again by means of some modern test models.

Error analysis has attained ever greater significance because it is very likely that the "new" Mathematics has brought with it some new types of errors. Instructions usually include more figures (models of sets etc.) and other kinds of ways for which visual shapes are needed. Traditionally these kind of things have basically been taught in geometry. However, geometry is not as important a topic in the new Mathematics as it was in tradiotinal courses. Analysing errors is a very troublesome task if no multi-choice items are used. It is worth doing because it may reveal some very common mistakes that can be avoided if the teacher knows in advance what will be the most likely mistakes in this particular topic.

Among others Newman (1977) and Casey (1978) have analysed errors in mathematical tasks in primary and junior secondary levels in Australia. Both of them has developed their own category of errors. Clements (1979) has compared their categories and developed two strategies of problem solving. Newman's hierarchy is constructed for one-step verbal mathematical problems and it is quite simple. Casey's approach emphasizes identifying and solving an appropriate set of subproblems. The main difference between the hierarchies of Newman & Casey and the category of errors in the present study (figure 3) is the problem-solving process of the former. The latter does not consist only of written mathematical tasks. That is why many errors are errors in computations.

The present study is an attempt to identify the characteristics of equations that are associated with differences in difficulty of solution. For this purpose it is necessary at first to find out the most common learning difficulties that pupils have in solving equations and applying their skills in problems. The first task in this study will be to answer the research problems:

1. What are the most common types of errors in each item?
2. What kind of structure of errors do they form?
3. Are the same errors common also in the United States and in Australia?

2.2. Methods used in empirical error analysis

Subjects are pupils of Finnish comprehensive school. The error analysis is based on the data of primary level. However, all subjects and tests are mentioned here because they are needed in chapter 4.

2.2.1. Subjects and their schools

Subjects came from different parts of Finland. No random sampling has been done because models used are sample-free and it is obvious that if we choose some schools or classes from different parts of Finland they will cover different ability levels of pupils because we are dealing with the comprehensive school and classes are heterogenous.

Primary schools were:

Jyväskylä, Huhtasuo	466
Keitele, kirkonkylä	151
Loimaa, maalaiskunta	344
Nilsiä, kirkonkylä	286
Rauma, normaalikoulu	20
	1267 subjects

Junior secondary schools were: Loimaan maalaiskunta, Opintien yläaste; Jyväskylä, Syrjälä; Inari, Pohjois-Inarin yläaste: Vammala, Sylvään yläaste: Nilsiä; Keitele: Vehkalahti: Ylihärmä: Kannus. The total number of subjects is 1310.

Pupils came from schools of different sizes (Table 1). Size of school is missing in the sheets of 25 pupils.

Table 1. Size of school

Number of pupils in school	Number of subjects	Percentage of subjects
- 100	423	16.5
100 - 200	500	19.7
200 - 300	-	-
300 - 400	306	12.0
400 - 500	763	29.9
500 - 600	-	-
600 - 700	466	18.2
700 - 800	94	3.7

According to official statistics of Finnish schools distribution of pupils in schools of different sizes was in the Finnish-speaking area (Table 2):

Table 2. Number of pupils in schools of different size in Finnish speaking Finland in 1978 - 1979

	Number of pupils				
	-109	110-199	200-499	500-699	700-
primary level 395 761	107 613	53 913	150 946	45 984	37 305
junior secondary 220 889	598	11 833	138 274	48 801	21 387
sum 616 650	17.5 %	10.7 %	46.9 %	15.4 %	9.5 %
Corresponding percentage in the present study	16.6 %	19.6 %	41.9 %	18.2 %	3.7 %

It can be seen that small schools (less than 100 pupils) are represented in empirical data to the same extent shown by the statistic for the whole country.

The number of schools in this study is 21. Location of schools can be seen in the map in Figure 1.

Compared to the number of inhabitants, Southern Finland is not very well represented in the data. The reason for this is that in spring 1979 the district of Helsinki had not yet adopted the comprehensive school curriculum. Schools from districts such as the latter have been avoided.

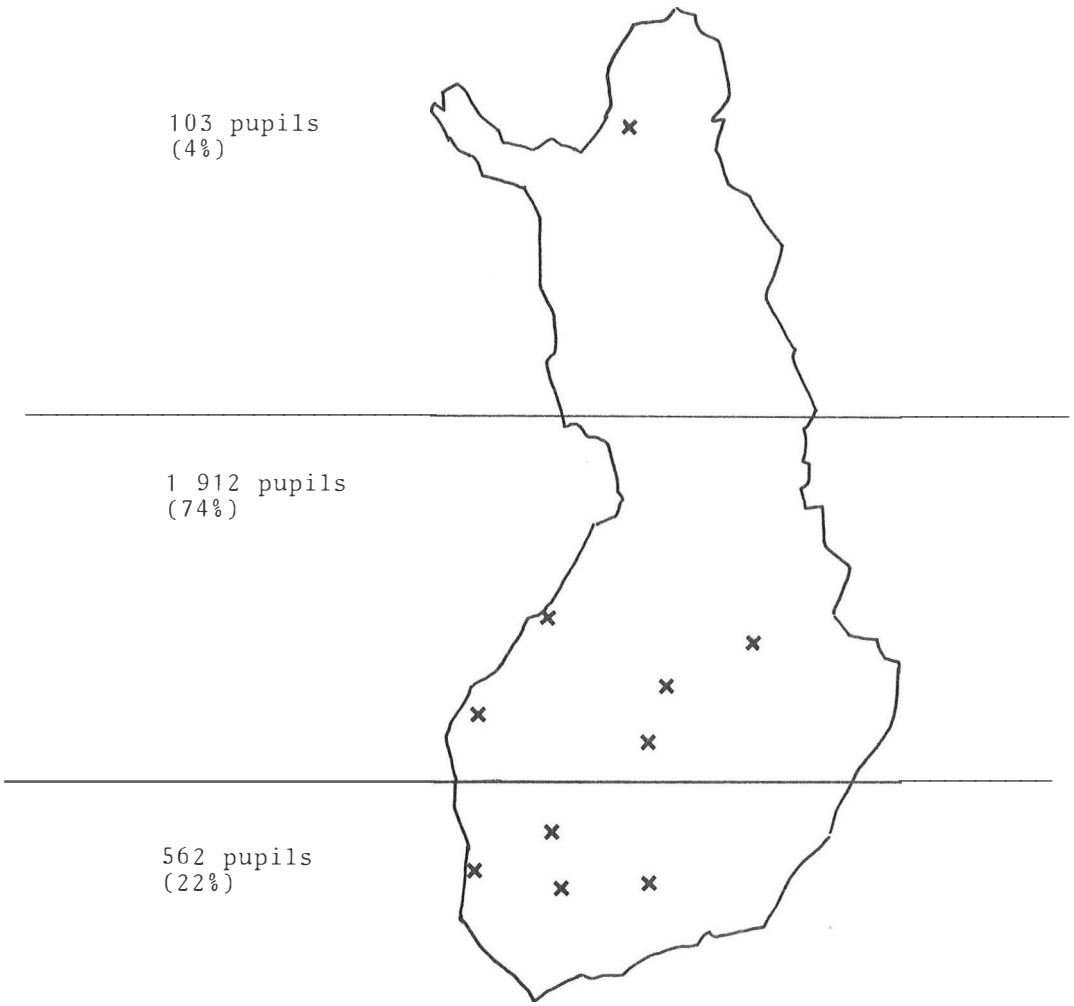


Figure 1. Location of the schools

The number of girls (47.7 %) and boys (52.3 %) in the data are approximately equal. In the junior secondary school some significant differences exist in the number of pupils in the level groups. Boys tend to choose lower level groups than girls, three quarters of the pupils at the lowest level group are boys. This may be the case in the whole country. Tests were administered in March, April or May 1979. They were sent to schools on 7th March. Marks in Mathematics were distributed as follows:

Table 3. Distribution of marks in Mathematics

Marks	Frequency
4	21
5	148
6	351
7	513
8	575
9	458
10	26
Missing	485
Mean	7.41
Variance	1.64

2.2.2. Tests

All the items are equations, inequations or verbal applications of them. This area has been chosen because it seems to be traditional enough: teachers similarly emphasize it (Koponen 1976). The items are of a free-response style in every case except items no. 13 -22 in lower level groups at the 7th, 8th and 9th grades. The number of items and subjects can be seen in Table 4.

The data contains 15 tests for different grades of the Finnish comprehensive school. For primary level there is one test for each grade and for the junior secondary level one test for each level group. For the 7th grade there are two

similar tests for the upper group, the only difference between tests is that items are administered in opposite order. Tests were administered in the spring term in 1979. They are presented in Appendix of this study.

Table 4. Number of items and subjects

Grade	Test	Number of items	Number of subjects	
		30	202	primary level 1267 pupils
2	2	30	210	
3	3	30	207	
4	4	30	149	
5	5	30	238	
6	6	30	261	
7, lower group	7	22	109	junior secondary level 1310 pupils
7, upper group	8	22	197	
7, upper group	9	22	110	
8, lower group	10	22	98	
8, middle group	11	22	173	
8, upper group	12	24	168	
9, lower group	13	22	48	
9, middle group	14	22	191	
9, upper group	15	24	216	
			2577	

Most of the distributions ot test scores were normal (Table 5).

Coefficients of reliability are quite high. At the primary level the split half method gives reliabilities from 0.77 to 0.88. Reliabilities in the junior secondary school are not as high. One reason for this is that three tests contain multi-choice items and give the possibility of guessing. Most items are the same for the 7th. 8th and 9th grades. They seem to be most suitable for the 9th grade because reliability at that grade is highest (Table 22).

Table 5. Distributions of total scores, Kolmogorov-Smirnov test

Test	Mean score	sd	D_{K-S}	p	Distribution
	17.6	6.1	.0708	.263	normal
2	17.8	6.4	.0820	.119	normal
3	16.0	6.4	.0658	.332	normal
4	15.7	7.3	.0801	.294	normal
5	16.0	6.1	.0482	.637	normal
6	15.8	6.8	.0643	.230	normal
7	6.5	2.8	.1128	.125	normal
8	10.1	3.9	.0971	.049	
9	10.0	3.6	.0936	.290	normal
10	5.1	2.5	.1297	.074	
11	6.1	2.5	.1331	.004	
12	12.4	4.0	.0846	.180	normal
13	5.1	1.9	.1799	.089	
14	8.2	4.2	.1122	.016	
15	14.2	4.5	.0930	.048	

2.2.3. Methods of analysis

Answer sheets of pupils were examined and corrected every error being given its own coding number. If the answer was correct its code was 01, incorrect answers were coded with numbers from 02 to 99, and if the answer was missing the code was 00. HYLPS-programmes were used for computing percentages of each error (IT-programme). OSANA-programmes were used for analysing the most common errors (Konttinen & Kortelainen 1979). In the second version of OSANA it is possible to use classifications of ten classes. In this case classes were: 1 = correct answer, 2, 3,...,9 = 8 most common incorrect answers and 10 = other incorrect answers. Every test has been analysed separately. For this analysis pupils at every grade were divided into two groups using their raw score as a criterion. Groups are approximately of the same size. These groups

were called group (50 % higher scores) and group 2 (50 % lower scores). In the comparision of errors in each group the following points have been taken into consideration:

(i) How many most common errors are needed in order to take 100 % (or nearly 100 %) of errors in the group in question into account?
(ii) Are the most common errors the same in each group?
(iii) What kind of errors are easy to get rid from?
(iv) What kind of faulty computations have caused errors?

2.3. Results of the empirical analysis of errors

2.3.1. Common item to grades 1 - 6

There is one item which is the same in every test, and some items (about five) which are common at two consecutive grades. In some cases items are common at three or four grades.

If we look at the item □ - 4 = 16 (which is common in every test) at 1st grade as an example of coding and presenting different types of errors, the summary of frequences after the OSANA-run in this item is in Table 6.

Table 6. Errors in item □ - 4 = 16 for the first grade

Answer	Correct 20	Incorrect answers 10	0	12	19	11	6	4	8	other errors	sum of errors
group	80	3	0	8	1	1	0	0	0	1	14
group 2	37	6	3	19	3	1	5	5	3	13	58
Σ	117	9	3	27	4	2	5	5	3	14	72

Using these frequencies a table of percentages can be constructed (cumulative percentage for different errors in each group).

Table 7. Most frequent errors at 1st grade in item □ - 4 = 16

Error of group 1	Group 1 (%)	Error of group 2	Group 2 (%)
12	57	12	33
10	79	10	43
19	86	4	52
11	93	6	60
93 % of errors in higher score group can be explained using only 4 most frequent incorrect responses		19	66
		8	71
		0	76
		11	78

In the group of 50 % higher scores only some types of errors exist. In the lower group a greater variety of incorrect responses can be found. The wrong answers 12 and 10 are the most frequent in both groups.

The next in order in group 1 is the incorrect answer 19. It can be seen that in group 2 this incorrect answer becomes more popular in the 2nd grade (the second most frequent) and at the same time it has completely disappeared in group 1 after the first grade.

It can be seen that group 1 has only a few types of errors. Also, in group 2, the number of different errors tends to decrease through the years. The only exception appears in the 6th grade. The reason for new errors in the 6th grade is likely to be that pupils have just learned some new computations in the domain of the whole numbers (Z) and they cannot apply correctly the rules for negative whole numbers. That is why they make mistakes in signs also in case of the common item.

The way of getting the most common incorrect answer is to subtract 16 - 4 12, which shows that the concept of the equation has not been understood. It might be better to use the vertical form in subtraction as Cox (1975) has recommended in his study of systematic errors in addition and subtraction computations.

In the 3rd and 4th grades a new error can be found in the answers: this is the incorrect answer 4. It becomes the most frequent or the next most frequent error type. The new mistake may be the result of practising multiplying skills in these grades. Pupils have assumed that this item will be a multiplication too, i.e. x·4 = 16 and so they have got the answer 4. At grades 5 and 6 this error can be seen in group 2 but it is no longer very common.

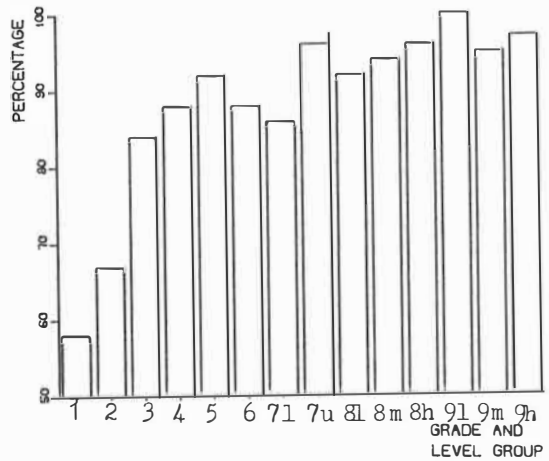


Figure 2. The percentage of correct answers to the item x-4 = 16

2.3.2. Items common to three consecutive grades at the primary level

The tests include three items which are the same for the three grades. The item: "You have 20 litres of juice. How many half litre bottles are needed for bottling?" was administered to the 3rd, 4th and 5th grades. Two typical incorrect answers in this item are 10 and 4. The error 10 becomes more common in group 1 than in group 2. Instead of multiplying by two pupils have divided by two. They have not thought critically after getting the answer. If they had compared the response to 20 litres they would have noticed their error. Lack of critical thinking is a typical feature of a child's thinking at the level of concrete operations. Primary pupils do not hesitate to give any answer, whether it seems to be of the right class of magnitude or not. For example to the item for second graders: "Mother baked 54 buns. One baking tray takes 20 buns. How many trays did she need?" many pupils presented the answer 74, which is more than the amount of buns. They have added the figures in the item and have not considered whether their answer has within the bounds of possibility or not. The most common mistake in this item was 34 which is the difference of 54 and 20. It is question of poor reading skills and repeating the computations that have been practised just before the test. (The latter item mentioned in this paragraph was not common for consecutive grades: it is mentioned only because of the similar errors to the former item.) Adding all numbers in an item and using the opposite operation are very common mistakes in word problems (Arter & Clinton 1974). Also the syntax of sentences and vocabulary used in items are essential features when the difficulty of an item is in question (Rosenthal & Resnick 1974; Linville 1976). One essential feature of an item is whether real objects are used in it or not (Caldwell & Coldin 1979).

A common item at grades 4, 5 and 6 was: 8700 - 3888 = x. In this item, group 1, as well, made many different mistakes. Good computing skills are needed for correct solution of this

item. The percentage of correct answers decreases from the 5th to the 6th grade (71 % at 4th, 91 % at 5th and 79 % at 6th grade). A typical incorrect answer in this item is 4912. The digit of hundreds has not been subtracted by one, although one hundred had been borrowed. According to the Finnish mode of calculating subtraction in the example 8700 - 3888 4812 after borrowing from e.g. the tens to subtract units, the loan is deducted from the tens of the number being subtracted from. A similar case can be found in the study of Pincus (1975) and it is called "difficulty with multiple exchanges". Another incorrect answer 5812 is quite analogous in the case of digits of thousands. The wrong answer 4811 is not common after the 4th grade. That is a sign of learning multiple exchanges at the 5th grade. The error 5188 suggests subtracting from the bigger number the smaller one. Pupils subtract in each column in the direction which offers the least difficulty (Roberts 1968; Cox 1975; Pincus 1975).

The third common item was $2 \cdot x - 5 = 15$ (common for 4th, 5th and 6th grades). The most typical error in both groups was 20. Pupils who have got this answer have solved $2 \cdot x$ instead of x. They have forgotten about the second part of this two-stage problem. This is a typical error in elementary algebra-problems (Malinen 1969). As early as the 4th grade pupils "know how to make" the error 20. At this grade level the classification of errors is easiest. All errors of group 1 are similar, that is to say, 20 is the only error in this group. At the same time, group 2 has 7 different errors. For getting the correct response to this item, skills in solving equations are needed. Solution procedures are taught at the 7th grade. At the primary level all equations are solved by means of four fundamental processes of calculation. There is no point in practising complicated equations with many stages at the primary level when in every case pupils must think of the qualities of addition, subtraction, multiplication or division for arriving at the solution.

2.3.3. Items common to two consecutive grades

Because of linking items in the Rasch simple logistic model it is necessary to have some common items at consecutive grades. Common items are also important for error analysis. It is interesting to compare the types of errors in the same item at different grades. By means of the items which are unique for each grade, it is possible to check what kind of errors are typical at that grade and to get more information for constructing elementary parameters for the linear logistic model.

First and second graders got 5 common items, one of them being addition, two subtractions, one multiplication and one a word problem. Mistakes in addition procedure were caused by missing a figure carried over or marking it in the middle of the sum (Cox 1975; Pincus 1975). The former error remains permanent, the latter error has nearly disappeared by the second grade. In subtraction the most common wrong answer suggests that addition has been applied instead of subtraction (Roberts 1968; Lankford 1974; Cox 1975). The next common wrong response, 24, for first graders in group 1 is the result of borrowing when it was not needed. The same error exists in group 2 at the second grade when they also have learned something about borrowing. The most frequent mistakes in group 1 in the other subtraction item at the first grade are the typical errors of group 2 at the second grade.

The common multiplication item was: "There are 8 pieces in every orange. How many pieces are there in 3 oranges?" This multiplication item in the first grade was made easier by drawing a picture that may have helped in getting the correct answer. In the picture there are only two oranges in which pieces can be counted one by one, the third orange is a whole one. No picture was presented to second graders. The incorrect answer 16 in the first grade suggests that the pieces of two oranges had been counted in the picture. This answer is no longer common in the second grade when the picture is missing. Pupils in group 2 tend to add the numbers 8 and 3.

Second graders have incorrect responses 23 and 25 which are likely to be caused by an error in computation. The answer 32 may be found by adding up the pieces of two oranges and taking it twice. Only by interviewing students might it be possible to get information about their thinking processes. However examples of the above-mentioned errors can be found in many other studies (Roberts 1968; Lankford 1974; Cox 1975; Newman 1977; Hollander 1978; Clements 1979; Hendrickson 1979).

In verbal problem number 30, at the first grade, the common mistakes are 78 and 50 which can be got straight from the problem without any calculations. The incorrect answers that are typical for second graders (28, 32 and 38) suggest that pupils have got quite far in the solution but they have made a mistake in their computations because of carelessness and because of the difficult numbers used in this item.

Tests include 4 common items for second and third graders. One of them is subtraction with borrowing in it: $\begin{array}{r} 102 \\ -\ 81 \end{array}$ The most common incorrect answer for second grader in both groups and for third graders in group 2 was 181. This response can be found by subtracting the smaller number from the bigger one in spite of their position in the algorithm. Borrowing without any reason (Cox 1975; Pincus 1975) gives the incorrect answer 11.

The rest of the common items for second and third graders are multiplications. The simpliest one 3 = 24 shows that difference between groups 1 and 2 decreases from the second to third grade when an easy routine calculation task is concerned. The multiplication tables from 1 to 5 have usually been taught in the second grade and revised in the third grade at the same time when the tables from 6 to 10 have been taught. In the case of a more difficult multiplication task 7 · 8 = the classification of error cannot be found so easily. Nearly the same errors can be seen in group 1 and group 2. The multiplication 3 · 124 was intended to be solved using addition at the second grade and multiplication at the third grade. That is why the mistakes are quite different in this item. Analysis of erroneous answers gives cause to suppose that every

number has been multiplied by 3 separately but the figure carried up has been forgotten or it has been marked at an incorrect place in the middle of the answer.

Tests include 2 common items for third and fourth graders. In the item x : 4 = 7 the typical errors 1, 3, 2, 24 and 11 at the third grade seem to exist in group 2 at the fourth grade. Only 7 % of pupils in group 1 have some error at the fourth grade. The corresponding percentage in group 2 is 43 %. The common word problem is quite complicated (item 3.28). Also the language in it is very difficult. The most common mistake at the third grade in group 1 will be the most common mistake in group 2 at the fourth grade: at the same time most of the fourth graders in group 1 have learned to avoid it. In both grades this item has caused the greatest number of errors of all items. Third graders made 65 and fourth graders 50 different errors in this item.

The incorrect resposes to the common word item of grades 4 and 5 cannot be precisely classified at the fourth grade. At the 5th grade the typical errors seem to be 13 and 13.33, which suggests errors in division processes. The words "third part" has been interpreted too literally "what is the third part" (item 4.28). In the common item x : $^{-}4$ = 8 at the 5th and 6th grades mistakes in sign of x are typical.

2.4. Summary of errors at the primary level

Errors can be classified into some basic types. The structure of error types is clearer in the case of easy fundamental items than in case of word items with many operations. Certain types or errors can be found in the group of high raw scores at earlier grades than can be found in the group of lower total scores in this test.

The most frequent errors are basically caused by oversimplification of the problem. The response has been formed as easily as possible from the numbers of the problem. The idea of equality has not been taken into account. Only the calcu-

lation operation which first comes to mind has been performed in order to get the response. The same feature can be seen in word problems, too. If a problem gives a hint of some operation this operation is used whether it is correct or not (e.g. "If temperature rose 4^{o}, it would be $^{-}20^{o}$. What is the temperature now?" This has been solved as: $^{-}20^{o} + 4^{o} = ^{-}16^{o}$ because "rise" gives a hint of adding the temperatures). Another reason for faulty responses are unfinished items. The answer has been taken straight from the problem or from the figure illustrating the problem or some computations have been performed but not all that were needed. Most textbooks in Mathematics are basically workbooks with ready structures of problems. Pupils need only fill in some empty squares. This gives reason to suspect that pupils are not used to constructing solving procedures from the beginning to the end themselves.

The flow chart in figure 3 contains the summary of all common errors. Every error which more than 10 % of pupils at the same grade have made can be found in this flow chart. The same diagram would be suitable with certain modifications for analysing errors at the 7th, 8th and 9th grades.

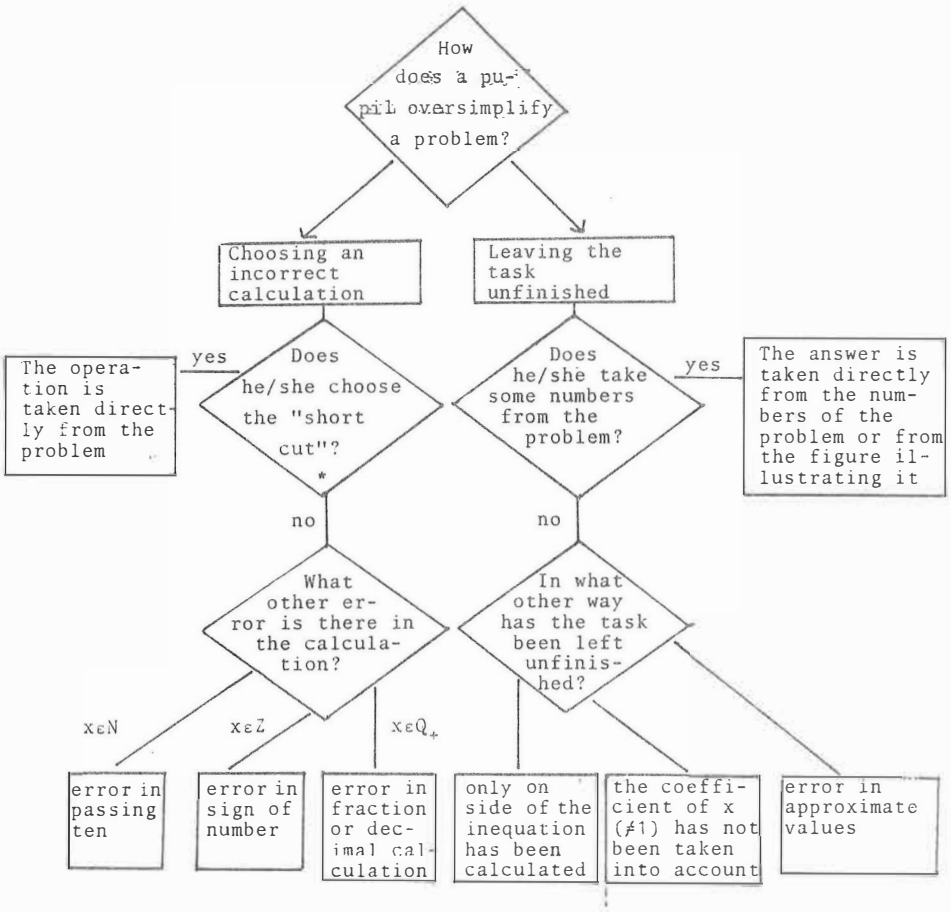
Types of errors can be studied more closely after this general view in the form of flow chart.

(1) Errors that are caused by using a "short cut" are very common for 1st, 2nd and 3rd graders if an item creates the possibility of making this kind of error. Table 8 gives a summary of these errors and their percentages.

(2) After this other erroneous choices of computation processes can be taken into consideration. The computation (x) has been incorrectly changed to another computation (y). This can be seen in Table 9 (x –> y).

In the error classification of many researchers some classification of wrong operations can be found (examples of newest studies: Roberts 1968; Arter & Clinton 1974; Pincus 1975; Newman 1977). Hollander (1978) takes the study of Ricer as an example of the finding that an incorrect operation had been chosen very frequently already in the 1920's in grades three through eight. It was the most common error in all

grades. A fifth of the subjects, interviewed individually, could give no reason why they had proceeded in their problem solving attempts as they had.



* We mean the simplest way of handling numbers. For example, when a first grade pupil sees the problem $\square - 4 = 3$, he thinks that it is subtraction and the numbers for operating are 4 and 3, and that the answer must be $4 - 3 = 1$.

Figure 3. Summary of error analysis at primary level

Table 8. Errors which may be caused by choosing a "short cut"

Item Grade. Number	Error	Percentage of pupils obtaining this incorrect answer
1.9	1	31.7
1.10	12	13.4
1.12	7	14.4
2.9	46	13.8
2.10	64	10.5
3.11	2	32.9
3.16	2000	27.1
3.19	8	11.6
3.25	-16^{o}	30.9
3.27	100	10.1
4.21	22	15.4
5.30	45 and 20	40.5 and 13.5
6.8	7 and $^{-}7$	11.9 and 23.4

Table 9. Wrong computation processes at primary level

Item	Error (computation)	% of pupils
2.6	64 - → +	11.4
2.27	34 : → -	13.3
3.5	5 + → ·	13.5
3.8	3 · → -	12.1
3.11	5 · → -	15.5
3.13	120 : → -	11.6
3.29	10 · → :	16.5
4.25	10 . → :	12.1
5.12	2 + → -	29.8
6.7	12 power →·	12.6
6.25	4 percentage → :	13.1

Lindvall and Ibarra have obtained results about different error types in open addition and subtraction sentences that support our results (Lindvall & Ibarra 1980). They collected data of 101 first and second graders from five elementary schools in the USA. Their research was an exploratory investigation of the various incorrect solution procedures used by students in attempting to solve a variety of forms of open sentences. It was found that pupils did not have a true understanding of the equals sign and of equations until they had shown some mastery of open sentences. Four different testing conditions were used: a written test, reading open sentences, providing the answer to oral story problems, demonstrating with blocks what the sentence means. The sentences of the type a + ☐ = c and ☐ + b = c showed basically that most students had little difficulty with this kind of item except in its application to story problems. There was some tendency to add the two numbers given or to merely respond by reporting one of these given numbers as the needed answer. The most common incorrect procedure in sentences of the form ☐ - b = c was that the smaller had been subtracted from the larger. Lindvall and Ibarra explain this as follows: "For many students a number sentence is not an expression that states a relationship and that must be read in the sequential order given but only a listing of two numbers and an operation sign, and the operation is to be applied to these numbers in the possible, or the most convenient, manner."

(3) The remaining errors in our data are characteristic of some domain of numbers. Passing ten is a difficult task for first and second graders. Errors seem to be of three types:

(i) In the vertical subtraction form the smaller number has been subtracted from the bigger one,

(ii) or the figure carried-up has been forgotten or

(iii) borrowing has been done incorrectly.

(4) Errors in sign are very common in the 5th and 6th grades. Many rules have been learned by heart and confusion of different computations can be seen. The most typical errors are mentioned in Table 11.

Table 10. Errors in passing ten

Item	Error	% of pupils	
1.29	21	11.7	type (i)
2.7	21	11.9	
2.18	181	19.2	
1.23	32	16.3	type (ii)
1.23	312	14.9	
1.28	3 Fmk 50 p	13.9	type (iii)
2.28	32	13.9	
2.29	32	11.0	

Table 11. Sign errors in some computations or in answer

Item	Error	% of pupils
5.3	$^{-}5$	24.4
5.4	$^{-}4$	22.3
5.6	$^{-}8$	30.3
5.7	$^{-}12$	28.2
5.8	32	17.6
5.12	8	14.3
5.27	$^{-}7\frac{1}{2}^{o}$	10.9
5.27	$8\frac{1}{2}^{o}$	16.8
6.2	7	55.6
6.12	32	20.7
6.15	$^{-}15$	21.5
6.20	3	15.7
6.22	55	10.0
6.29	$^{-}5.4$	10.3
6.29	3	10.3

The large number of errors in sign suggests that there is something wrong in the way negative numbers are taught or that they are taught too early.

(5) Errors in fraction or decimal calculations are mentioned in Table 12.

Table 12. Errors in fraction or decimal calculations

Item	Error	% of pupils
2.30	4	16.2
5.13	8	26.1
5.15	6	11.3
5.16	7	13.0
5.29	2	14.3
6.5	40 500	15.7

Similar error types to (3), (4) and (5) can be found in research reports. Passing ten was mentioned in the connection of common items for consecutive grades. Sign, fraction and decimal errors are analysed e.g. by Lankford (1974) and Knifong & Holtan (1976). At every grade the interpretation of the symbol " = " is a problem for some pupils. Hendrickson (1979) states that children at the first grade use a frequent interpretation of " = " as the symbol that precedes the answer rather than as the symbol that relates two sets or situations in a certain way.

(6) The problem is unfinished, some number from the problem has been supposed to be the answer. Table 13 illustrates this case.

Table 13. The incorrect answer has been taken from the initial problem

Item	Error	% of pupils
1.30	50	17.8
2.26	4, remains 6	13.8
4.27	4.00	12.8
5.17	$\frac{3}{4}$	30.3
5.22	$2\frac{3}{4}$	17.2
5.23	0.6	38.2
5.24	$3\frac{3}{8}$	18.9

Other errors of unfinished tasks are collected in Table 14. In some cases (i) only one side of the inequation has been solved or (ii) instead of x only n . x (n ε N) has been solved or (iii) there is some error in approximate quantities.

Table 14. Unfinished computations

Item	Error	% of pupils	
1.16	2	25.2	type (i)
1.17	4	30.7	
2.15	26	11.9	
2.17	7	10.5	
3.24	4	17.4	
4.4	9	12.8	
4.16	9	11.4	
3.12	6	12.1	type (ii)
3.27	16.40	17.9	
4.14	8	12.8	
4.22	20	14.1	
4.26	6 Fmk	10.8	
2.27	2	14.3	type (iii)
5.5	20	14.7	

Every error, which is so common, that 10 % or more of primay pupils in the data have made it, has been classified. Although random sampling has not been used in this research it can be considered that errors which exist in the case of 10 % pupils are worth of studying. It may be possible to improve teaching methods or the curriculum in Mathematics after careful studies of frequent errors. Mathematics cannot be taught without pupils making errors. However, it is good to follow a classification of errors. If no certain types of error can be found, all errors seeming to be random, it may be that nothing has been learned. Learning seems to result in a decrease of different types of errors. Concentrating on certain types of errors seems to be sign of learning. Analysis of errors is not an operation for eliminating all the

errors, but for finding development of children's ways of thinking and correcting their wrong thinking procedures.

The process of solving verbal problems in Mathematics has often been described as consisting of two general stages - translation and computation (Caldwell & Coldin 1979). During the translation stage, the problem solver sets up the verbal problem statement as a system of mathematical expressions or equations. During the computation stage, the problem solver performs the algebraic operation necessary to obtain the solution. Only some items in our data are verbal. This solution has been adopted because it is only by interviewing pupils that their thinking processes in verbal problems can be discovered. In this study it was not possible to interview anybody: that is why the various kinds of difficulties in problem solving have not been separated. There is only one elementary parameter for problems in the linear logistic model.

The linear logistic model assumes that every person uses the same strategy in getting the correct answer. Some studies hint at the great number of memorized facts in fundamental calculations (Cifarelli & Wheatley 1979). In many cases children develop and employ strategies for finding sums and differences regardless of the instructional treatment (Steffe 1979; Rathmell 1979). In many cases the strategies of first graders are based on the counting of physical objects (Jenks & Peck 1976). The correct answer can be found by means of different strategies. It must be the same with incorrect answers. Error analysis procedures in this research are based on classifying the most common errors and trying to follow the thinking processes of children for each common error. It might be a good topic for further research to interview students who have made a certain type of mistake and to attempt to discover the psychological reasons for their errors.

It was very interesting to see that basically the same structure that was found at the primary level suits error analysis at the junior secondary level. Some improvements have been made because

(i) the domain of numbers is larger,
(ii) solving equations becomes more routine and
(iii) one number is no longer enough for the solution of inequations.

Error analysis at the junior secondary level could be the topic of another study. This, however, falls outside the scope of the current report.

Choosing the system of elementary parameters in the linear logistic test model has been criticized by Scheiblechner (1975) for lack of sufficient background with the argument that some parameters are elementary parameters some others are not. This research suggests one way of finding elementary parameters by means of error analysis.

Another way of studying thinking processes of pupils in mathematical tasks could be that used by Krutetskii (1976) and Leino (1978). It is based on interviews and its result is the knowledge about abilities of the pupils and individual differences in their thinking processes.

3. MATHEMATICAL MODELS FOR STUDYING LEARNING DIFFICULTIES

3.1. Model of achievement tests

A characteristic feature of statistical research is the attempt at finding a mathematical model for describing the phenomenon in question as clearly and simply as possible while still preserving the analogy with the real situation. Common features of different situations can be found by using mathematical models and their test of fit. Mathematical models must not simply remain abstract and useless games; they must be utilized in solving theoretical and practical questions.

Learning models have been adopted in educational research for the following reasons (Pelka 1975):

(i) When the learning situation includes some parameters it is possible to follow changes in these parameters from one learning situation to another.

(ii) Using statistical models it is possible to analyse the similarity and difference between observations.

(iii) The character of the learning process can be studied by simulating learning conditions.

Theories can be considered to be either probabilistic or deterministic. Using probabilistic theories it is possible

(i) to test the structure of a theory,

(ii) to estimate parameters, and

(iii) to compare parameters in a methodologically sensible way.

The limitations of deterministic theories are their assumptions. They make it difficult to find any empirical data for them, the situation often having been made unreal. That is why it is often difficult to apply deterministic theories in teaching and learning research. Both teaching and learning are complex processes and the assumptions of the theories are rarely valid. In all events, theoretical decisions must be reduced for studying learning. In probabilistic models there are features of probabilistic theories even if the assumptions are not as straightforward as they are in the case of theories.

Finnish researchers who deal with the usage of probabilistic and deterministic models (Koskenniemi 1968; Malinen 1974) prefer the former ones in the case where variables are not easily measurable and when it is the question of a survey research without any prior theoretical foudation. Probabilistic models are in these circumstances instruments needed for further research.

3.2. Some characteristic features of latent trait models

Traditional methods of measuring achievements may be good enough for some purposes. However the coefficients of reliability and validity are not sample-free (Rasch 1960; Lumsden 1976). This is related to the specific objectivity, which is one of the major differences between traditional test theory and latent trait theory.

3.2.1. Specific objectivity

The Danish Yearbook of Philosophy (1977) illustrates the concept of specific objectivity from viewpoint of different sciences. Gerholm (1977) begins with the semantics of the words "subjective" and "objective". "Subject" derives from the Latin subjectus, sub means under or below, and jacere means to throw. In an analogous way "object" comes from objectus,

the prefix ob meaning to or toward. In other words: something (the object) is thrown towards and under something else (the subject). Kant made a distinction between two different kinds of subject: the psychological subject - the "real subject" and the epistemological subject - the "formal subject". Apparently what Kant means by "psychologically subjective" comes close to what is now meant by subjective, whereas his "epistemologically subjective" is about the same as our present "objective".

In behavioral sciences experiments are conducted to elicit information which will help to locate the person (the object of the experiment) on the variable. Typically the person is "prodded" with an agent (otherwise known as a stimulus or an item) and the reaction of the person is observed. If the agent is well chosen, the outcome of the object-agent interaction should give a hint of the person's position on the variable. In general, the more relevant the agents are, and the more the number of responses observed, the more precise will be the estimate of the person's position on the variable.

Rasch (1977) defines "specific objectivity" as follows: He takes two collections of elements: objects $\mathcal{O}$ and agents $\mathcal{A}$. The corresponding single elements are O_v and A_i; the collections can be finite or infinite. Starting with the Poisson model in reading errors and general gas laws, Rasch (1977) gives a theoretical interpretation to the concept of general objectivity. He defines the frame of reference of objects, agents and outcomes within which the concept of comparison has been defined. Specific objectivity is not an absolute concept; it is related to the specific frame of reference.

In the present study the objects are pupils, the agents are tests of equations (or items in them), and the reactions are different responses. A possible comparing function could be, for example, "pupil A can solve division items better than pupil B" or "pupil A has mastered operation multiplication for 4th graders better than pupil B". Objective comparison (in the sense of specific objectivity) could be made by using total scores of those items and the structure matrix of the tests (Q-matrix).

"Generally each object O may enter into a well-defined contact C with every agent A, and every such contact has an "outcome" R. The collection of possible contacts is $\mathcal{C}$ and the collection of possible outcomes is denoted $\mathcal{R}$. The specification of the three collections of elements gives the frame of reference

$$\mathcal{F} = [\mathcal{O}, \mathcal{A}, \mathcal{R}]$$

within which we shall now define the concept of comparison." (Rasch 1977, p. 75)

Object O_v can be characterized by a scalar parameter ξ_v and analogously agent A_i by parameter δ_i (Fischer 1974),

(3.1)
$$P\{X_{vi} = x_{vi} \mid O_v, A_i\} = f\{x_{vi} \mid \xi_v, \delta_i\}.$$

That is to say that the reaction (outcome) is fixed after fixing parameters ξ_v and δ_i. On the other hand different combinations of parameters may give the equivalent reaction. In the case of the SLM the difference $\xi_v - \delta_i$ is essential, not only the values of parameters.

Rasch gives in his first main theorem the necessary and sufficient condition for the existence of specific objectivity:

"Let objects and agents in the bifactorial determinate frame of reference $\mathcal{F}$ be characterizable by scalar parameters ω and α, and reactions by a scalar reaction function of "convenient" mathematical properties

$$\xi = \varphi(\omega, \alpha).$$

Then the existence of three (strictly) monotonic functions

$$\omega' = \psi(\omega), \quad \alpha' = \Psi(\alpha), \quad \xi' = \chi(\xi),$$

transforming the parametric reaction function into a purely additive relation

$$\xi' = \omega' + \alpha'$$

will be a necessary and sufficient condition for specifically objective comparability of objects as well as agents. If such functions exist they are unique apart from trivial linear transformations. The specifically objective comparison of two objects O_λ and O_v and of two agents A_i and A_j is based on the respective differences

$$\xi'_{\lambda i} - \xi'_{vi} = \omega_\lambda - \omega_v$$

and

$$\xi'_{vi} - \xi'_{vj} = \alpha'_i - \alpha_j$$

the right hand expressions of which are denoted the elementary comparators." (Rasch 1977, pp. 79-80)

Rasch gives some examples of using his first main theorem in physics. Spada and Fischer (1973a, 1973b) develop methods, which have the property of specific objectivity, for evaluating responses in projective tests of psychology in the case of more than two response categories. Andrich (1978) gives an iterpretation of category coefficients and scoring functions for polychtomous Rasch models in terms of thresholds and discriminations at the thresholds. Douglas (1980) shows that specific objectivity holds also in the generic Rasch model.

3.2.2. Sufficiency

Two essential requirements of good measurement can be expressed as follows:

> "At the very least, a good measurement model should require that a valid test satisfy the following conditions:
> 1. A more able person always has a better chance of success on an item than does a less able person.
> 2. Any person has a better chance of success on an easy item than on a difficult one.
> It follows from these conditions that the likelihood of a person succeeding on an item is the consequence of the person's position on the variable (his ability) and the item's position on the same variable (its difficulty) and that no other variables influence the outcome. This implies that the difficulty of an item is an inherent property of that item which adheres to it under all relevant circumstances without reference to any particular population of persons to whom the item might be administered." (Wright & Mead 1977, p. 6)

These requirements lead to the conclusion that total score is a sufficient statistic of ability. For item difficulty, the sufficient statistic is the number of persons who got the item correct.

The starting point of the Rasch models is Poisson's law in analysing reading errors (Rasch 1960). Probability of errors is a function of a person's ability β_v and the difficulty of the text δ_i. It can be presented as a product

$$(3.2) \qquad \theta_{vi} = \frac{\delta_i}{\beta_v}$$

It is a question of relative difficulty and relative ability parameters. Similarly Rasch has used reading speed for which the corresponding result is

$$(3.3) \quad \lambda_{vi} = \frac{\beta_v}{\delta_i}$$

which means that reading ability of a person v has been expressed using the ability parameter of a person and the difficulty parameter of an item. For finding the probability function for the SLM Rasch takes the ratio $\frac{\beta}{\delta}$ as a variable and gives the definition using the easiest possible function which has values from 0 to 1 when $\frac{\beta}{\delta}$ has values from 0 to ∞. This function gives the probability of getting the item i correct in the case of person v:

$$(3.4) \quad P(+|\beta_v, \delta_i) = \frac{\beta_v/\delta_i}{1 + \beta_v/\delta_i} = \frac{\beta_v}{\beta_v + \delta_i}$$

Rasch gives definitions for "degree of ability" and "degree of difficulty" when he is dealing with the structural model for items of a test.

> "If we maintain that the abilities of two persons have certain values and the difficulties of some test items certain other values, it must be possible to check whether this is true or not. For given values of β and δ it must, therefore, be possible to find an indicator of how easily the person solves the problem and the indicators must be comparable to each other." (Rasch 1980, p. 73)

Usually the probability of a correct answer, which is the basic equation of the SLM has been expressed in the form:

$$(3.5) \quad P(+|\beta_v, \delta_i) = \frac{\exp(\beta_v - \delta_i)}{1 + \exp(\beta_v - \delta_i)}$$

The main advantage of the logarithmic metric is that parametric relationships become additive, from which it follows that the measurements, which are parameter estimates, are at an interval level (Andrich 1975). On the assumption that the answers of different persons to a set of items are independent, dichotomouos stochastic variables and that the probabilities

of the two possible answers 1 and 0 of a person to an item depend only on two scalar positive parameters characterized by the person β_v and the item δ_i respectively then (3.5) but for trivial transformations is a necessary condition for the parameters always to be separable (Rasch 1968).

In the SLM persons with the same total score are grouped together. We need not take into account which particular items a person has got correct. A person's total score is a sufficient statistic for his ability parameter. This property is also called "item-free person measurement". Analogously, it should not matter what sample group of persons is used - the item parameter estimates should be the same. For the SLM the equation of probability can be expressed in the form

$$(3.6) \qquad P\{X_{vi} = x_{vi} | \beta_v, \delta_i\} = \frac{\exp x_{vi}(\beta_v - \delta_i)}{1 + \exp (\beta_v - \delta_i)}$$

if we want to give the probabilities of the correct and wrong responses, $x_{vi} = 1$ if the response is correct and 0 if it is wrong. Sufficient statistics for persons and items are respectively:

$$(3.7) \qquad r_v = \sum_{i=1}^{k} x_{vi} \qquad (v = 1,\ldots,N)$$

where k is the number of items and r_v is the total score of person v;

$$(3.8) \qquad S_i = \sum_{v=1}^{N} x_{vi} \qquad (i = 1,\ldots,k)$$

where N is number of persons and S_i number of persons getting item i correct.

The corresponding expected values are

$$(3.9) \qquad E(r_v) = \sum_{i=1}^{k} P_{vi}$$

and

$$(3.10) \qquad E(S_i) = \sum_{v=1}^{N} P_{vi} = \sum_{r=1}^{k-1} n_r P_{ri}.$$

Using these properties and second derivates of likelihood we can derive standard errors:

$$(3.11)\qquad SE(\hat{\beta}_r) = \frac{1}{\sqrt{\sum_{i=1}^{k} \hat{P}_{ri}(1-\hat{P}_{ri})}}$$

and

$$(3.12)\qquad SE(\hat{\delta}_i) = \frac{1}{\sqrt{\sum_{r=1}^{k-1} n_r \hat{P}_{ri}(1-\hat{P}_{ri})}}$$

where

$$(3.13)\qquad \hat{P}_{ri} = \frac{\exp(\hat{\beta}_r - \hat{\delta}_i)}{1 + \exp(\hat{\beta}_r - \hat{\delta}_i)} \quad .$$

The precision of the estimates, measured by their standard errors, will vary depending on which persons are chosen for the calibration of items or which items are chosen for the measurement of persons. A more detailed study of good items can be found in the book by Wright and Stone (1979). Also, in the LLTM the number of correct answers is a sufficient statistic for β_v given the item perameters δ_i (Fischer 1977a). Estimates of elementary parameters $\hat{\eta}_j$ (j = 1,...,m) are sample-free with respect to sample of items as long as the Q-matrix is of full rank.

Sufficient statistic is a function of observed data and it completely summarizes the information pertaining to a single ability of a subject or to the difficulty of an item. Minimal sufficient statistic for a parameter is the simplest form of sufficient statistic. In a more accurate statistical sense sufficiency is combined with conditional probabilities and factorization of likelihood function (Cox & Hinkley 1974; Hogg & Craig 1978). Andersen (1977) establishes that if a sufficient statistic exists for the measurement of subject ability, then this statistic is the unweighted number of correct responses. It follows that the model must be of the Rasch type with logistic item characteristic curves and equal item-discriminating powers (Lord & Novick 1968). The result

can be extended for equidistant scoring in multiple choice questionnaires (Andrich 1978).

Separability of parameters is connected with objectivity of measurement. Comparisons between objects (persons) and agents (items) can be done separately if the model holds. Otherwise specifically objective statements cannot be derived from the data. Firstly, the failing of specific objectivity means that the conclusion about any set of person parameters, for example, will depend on which other persons are also compared. Secondly, the conclusions about the persons would depend on just which items were chosen for the comparison.

We will return later to the concepts of sufficiency and objectivity in comparing traditional test theory and Rasch latent trait theory (chapter 4).

3.2.3. Dimensionality

In the SLM there is only one parameter for each item. It means that only one quality - the difficulty of items - can be estimated. All other item parameters must be assumed to be controlled: items must measure the same dimension of latent space with approximately the same discrimination power and without influence of guessing. Items are supposed to be unidimensional if we want the SLM to hold. That is the case also in person space. Unidimensionality must also hold there (Andrich & Koponen 1980).

Dimensionality can be taken into account in designing the test and in analysing the results. It can be studied, e.g. using factor analytic techniques for finding homogeneous clusters of items for latent trait analysis (Lumsden 1957; Hambleton & Cook 1977; Mc Donald 1980). Factoral methods give some possibilities of misinterpretation:

> "If only a single trait is measured, but the variance of this trait in the sample is small, then simply because of this small variance, many factors which are unstable from sample to sample, will emerge. In the limit, as the subject abilities tend to equality, then as many factors as items will be found. On the other hand, if

> the traits measured are slightly correlated, and the subject variances on the traits are large, then the first factor will tend to smother all other factors. This feature of factor analysis does not seem to have been considered when interprations of factors have been made." (Andrich & Godfrey 1977, pp. 4-5)

Another way of checking unidimensionality is to use the SLM and its test of fit statistics. For example RATE-program (and its version NEWRATE which has been used in this research) gives not only item fit but also person fit statistics (Andrich & Sheridan 1980). The third general notion seems to be expressed in terms of homogeneity and its near-synonym internal consistency. Cronbach's alpha is one of the most useful measures of these properties.

3.2.4. Benefits of Rasch latent trait theory

Traditional test theory and latent trait theory are not completely different perspectives on testing. They give different methods by which we can get different types of information from tests. This may help in making better conclusions and improving tests (Konttinen 1979, 1981). One of the basic assumptions in the former theory is that observed score is the sum of true score and error score. Some assumptions of distributions are also essential. In the latter basic principles are specific objectivity and related concepts. Fischer mentions in his critique of traditional test theory the concepts of reliability, validity and homogeneity which are not independent of the sample used (Fischer 1974, p. 133). Some scaling problems of traditional test theory can be avoided using latent trait models. For the SLM and the LLTM qualitative data can be used (correct - incorrect answer) instead of the interval scale which must be used in the case of traditional methods.

In traditional theory some connections with latent trait theory can be seen: for example Guttman's scale and his radex-principle. Guttman's concept "radex" (radial expansion of complexity) is also useful in describing the character of tests (Lazarsfeld 1954).

> "Two distinct notions are involved in a radex. One is of a difference in kind between tests, and the other is of a difference in degree.
>
> Within all tests of the same kind, say a numerical ability, differences will be in degree. We shall see that addition, subtraction, multiplications and division differ among themselves, largely in the degree of their complexity. Such a set of variables will be called a simplex. It possesses a simple order of complexity. The tests can be arranged in a simple rank order from least complex to most complex.
>
> Correspondingly, all tests of the same degree of complexity will differ among themselves only in the kind of ability they define. We shall postulate a law of order here too, but one which is not from "least" to "most" in any sense. It is an order which has no beginning and no end, namely, a circular order. A set of variables obeying such a law will be called a circumplex, to designate a "circular order of complexity"." (Guttman 1954, p. 260)

In this research the radex-principle can be seen in the structure of tests.

(i) Difference in degree for tests can be seen when new labels have been given to elementary parameters when the domain of numbers has become more extensive. All the time it is a question of the same type of calculation (for example multiplication in the 3rd grade, multiplication in the 4th grade).

(ii) Difference in kind for tests can be seen in each test in its elementary parameters as a circular order of complexity. None of the parameters can be considered to be the easiest or hardest.

The main difference between traditional and latent trait theories is the probabilistic nature of the latter. Only that property makes it possible to use the concepts of sufficiency and information which are basics in this theory. Latent trait theories give new opportunities for analysing tests, particularly in studying the structure behind items and tests. For this kind of use the LLTM has been used (e.g. Scheiblechner 1972; Fischer 1973; Kempf 1975; Spada & Kempf 1977). Before using the LLTM it must be made sure that the data first fits the SLM.

One aim in using Rasch latent trait models is to test how good items are in the sense of objective measurement.

Having rejected poor items, the items left can be used for item banking. In classroom testing a small number of items is enough for getting reliable results if items are good from the point of view of latent trait theory. Also different ability groups in the classroom may get different samples of items (Izard & White 1980). The quality of items has been studied in this research from the point of view of traditional and latent trait theories. Also for the other research problem - a structural view of the items and tests - the latent trait models are used.

One of the research problems in this study is to find a common structure to the tests of solving equations. The problem cannot be solved by means of traditional test theory as effectively as by means of latent trait models. The linear logistic test model used in this research is an extraordinary tool for describing the data. The main principle of statistics, i.e. to express the characteristic features of the whole data by means of only a few parameters proves to be true in the case on the LLTM. Learning difficulties in the LLTM can easily be diagnosed when the tests have been constructed taking into account the common elementary parameters for linking. Test of person and item fit shows that items fit the SLM extremely well and also the LLTM. Even if they are constructed only in a very narrow area of mathematics, it can be argued that other consistent areas could also be used, and the test would still fit the model if the basic principles of difficulty order of items have been taken into account. This might be a useful way of rapidly getting information about learning difficulties in order to help pupils overcome them as effectively as possible. This kind of information would also be valuable for developing curriculum as well as discovering the thinking processes of pupils in different areas of Mathematics.

3.3. Criteria for choosing a test model

If the model is complicated and contains many parameters it will be difficult to find any empirical data fitting the model. In constructing the model, it is important to bear in mind some recommendations for checking that the model will fit the empirical situations for which it has been constructed (Cox & Hinkley 1974):

(i) The model must have connections with earlier theories.
(ii) The model must take the limits of its parameters into account, e.g. its assymptotes.
(iii) Every parameter must have its equivalent in reality.
(iv) The model should preferably have only a few parameters because the model has been constructed to express the essential features of the data in as compact a form as possible.
(v) It is desirable that the statistical theory of the model should be as easy as possible.

The test models used in this research fit the criteria mentioned above quite nicely. Logistic test models (SLM and LLTM) make it possible to describe empirical data using only some parameters for subjects and items. Logistic models are connected with a strong statistical theory which has been developed keeping in mind statistical qualifications of estimates e.g. sufficiency. By means of logistic models it is possible to deal with problems which are impossible in traditional test theory, such as sample-free item difficulty (Fischer 1976). The number of parameters becomes minimal in the case of the SLM, which has only one parameter for items: item difficulty. This model is one of the commonly used logistic test models bacause of its simplicity and effectiveness. Only the SLM has the quality that a person's raw score is a sufficient statistic for his ability on the latent trait the test has been constructed to measure. The last criterion for choosing test model concerns simplicity of the theory. In the case of logistic models it is hard to say how simple theories are: simplicity is always a relative concept. Statistical

methods for estimating parameters have been developed for conditional and unconditional estimates. One of the criteria in development has been to find reliable estimates using as little computing time as possible.

3.4. Simple logistic model

If it is supposed that the discrimination power of each item is the same and the possibility of guessing is eliminated, we get

(3.14)
$$P_i(\beta) = F(\beta - \delta_i) = \frac{\exp(\beta - \delta_i)}{1 + \exp(\beta - \delta_i)}$$

which is the basic equation of the SLM. It is a specific case of logistic models. There is only one item parameter: item difficulty δ_i. That must be taken into consideration when constructing items for a test. The difficulty level of items may - and it usually must - vary. Discrimination power must be approximately the same. In the SLM the discrimination parameter is standardized to be one: the model does not hold if it is different for different items. Essential benefits of the SLM are (Gustafsson 1977):

(i) comparing achievements of persons is item-free,
(ii) comparing the difficulties of items is person-free,
(iii) the score of a person is a sufficient statistic of his attainment parameter.

Reactional parameter λ_{vi} is defined (Fischer 1974) as the ratio of probabilities to success and to failure

(3.15)
$$\lambda_{vi} = \frac{P_{vi}}{1 - P_{vi}} = \exp(\beta_v - \delta_i)$$

In the ratio of λ's for two persons u and v the item parameter δ_i is missing from the expression. That is to say, comparing persons is item-free. Only attainment parameters are essential:

$$(3.16) \qquad \frac{\lambda_{vi}}{\lambda_{ui}} = \frac{\exp(\beta_v - \delta_i)}{\exp(\beta_u - \delta_i)} = \exp(\beta_v - \beta_u)$$

Analogously we can see that comparing items is not dependent on persons:

$$(3.17) \qquad \frac{\Lambda_{vi}}{\Lambda_{vj}} = \frac{\exp(\beta_v - \delta_i)}{\exp(\beta_v - \delta_j)} = \exp(\delta_j - \delta_i)$$

These results are essential when the objectivity of a test model is concerned. Because $0 < P_{vi} < 1$ $\log P_{vi}$ and λ_{vi} also exist. Both item and person parameters of the SLM are on the same scale. If we assume that teaching has improved learning results, we can say that items have become easier for these persons or alternatively a person's ability to solve items correctly has increased. Studying changes is one of the most important objects of research in education. For this purpose not only the SLM but also the LLTM has been commonly used.

3.5. Linear logistic model

In the simple logistic model (SLM) there is only one parameter for each person and for each item. In the linear logistic test model (LLTM) either the person parameter or the item parameter can be divided into components. Basically two different methods can be separated when the LLTM is used. The model can be used

(i) for two points of time

$$(3.18) \qquad P(+|\beta_v, \delta_i, t_1) = \frac{\exp(\beta_v - \delta_i)}{1 + \exp(\beta_v - \delta_i)}$$

$$(3.19) \qquad P(+|\beta_v, \delta_i, t_2) = \frac{\exp(\beta_v + \Delta_v - \delta_i)}{1 + \exp(\beta_v + \Delta_v - \delta_i)} .$$

The latter formula can be interpreted so that in the time interval (t_1, t_2) learning has happened and person parameter has become bigger. In the case on mass media (Fischer 1972) the learning effect in the time interval (t_1, t_2) is

$$(3.20) \qquad \Delta_v = \sum_{j=1}^{m} q_{vj}\, \eta_j + \tau$$

where

η_j's are different ways of learning (newspapers, radio, tv, etc.),

q_{vj} is the amount of time which person v has devoted to mass medium j,

τ is the trend which is independent of η's.

On the other hand it can be considered that an item i has become easier for a person and the probability of the correct answer in time t_2 is

$$(3.21) \qquad P(+|\beta_v, \delta_i, t_2) = \frac{\exp(\beta_v + \Delta_i - \delta_i)}{1 + \exp(\beta_v + \Delta_i - \delta_i)} = \frac{\exp(\beta_v - \delta^*_i)}{1 + \exp(\beta_v - \delta^*_i)}$$

where

$$(3.22) \qquad \delta^*_i = \sum_{j=1}^{m} q_{ij}\, \eta_j + \tau$$

η_j's are basic operations in item i ($j = 1,\ldots,m$; $i = 1,\ldots,k$),

q_{ij} is frequency of basic operation j in item i,

τ is constant for scaling (if needed) for making $\sum_{i=1}^{k} \hat{\delta}^* = 0$.

(ii) The same kind of basic operations can also be used without two different measurements. If all items in the test can be apportioned to the same components and the frequency matrix $((q_{ij}))$ is of rank m and in addition to this, the number of operations is smaller than that of items ($m < k$) it is not only possible to estimate item difficulties but also

difficulties of each basic operation even on an absolute scale. The linear logistic model is a Rasch model

$$(3.23) \qquad P(+|\beta_v, \delta_i) = \frac{\exp\{\beta_v - (\sum_{j=i}^{m} q_{ij}\,\eta_j + \tau)\}}{1 + \exp\{\beta_v - (\sum_{j=1}^{m} q_{ij}\,\eta_j + \tau)\}}$$

It has all the properties of the simple logistic model and some additional specific properties (Fischer 1977a):

(1) The number of correct responses is a sufficient statistic for β_v, given the item parameters δ_i.
(2) The ability parameters β_v are independent of the sample of items.
(3) The item parameters δ_i^* are independent of the sample of subjects.
(4) The operation parameters η_j are independent of the sample of items.
(5) The validity of the SLM and the LLTM can be tested by means of graphical methods as well as the likelihood ratio test.
(6) The item parameters δ_i lie on a difference scale.
(7) The elementary parameters η_j lie on an absolute scale if $m < k$ and the rank of q_{ij}-matrix is m.

Assumption of the rank of Q-matrix can be easily ascertained by computing the eigenvalues of matrix Q´Q. If all of them are positive the rank of Q-matrix is equal to m.

Fischer (1977a, p. 212) gives the following guidance for applied research:

- In a formal sense, it would be sufficient for the rank of matrix Q to be m. It is good if columns of Q are as ortogonal as possible (this does not mean that basic parameters must be uncorrelated).
- The number of items must be great enough compared to the number of elementary parameters.

- It would be good to get matrix Q stable at the very initial stages of research, because elements of Q and basic parameters cannot be estimated simultaneously and no column can be added afterwards without re-working of the data.

3.6. Some other latent trait models

During the last twenty years several models have been developed from the simple logistic model. Dynamic models (Kempf 1977a) include transfer parameter which means that the probability of getting an item correct is dependent on success in previous items. Different versions of the linear logistic models have been developed for measuring change (Fischer 1977b). Poisson models have also been developed for analysing frequency data in studying effects of treatments (Rasch 1960; Fischer 1977b). For attitude data Andersen (1977) and Andrich (1975, 1977a, 1977b, 1978, 1979) have developed rating scale models. The same computer programme used for the rating scale model can also be used in the case of the SLM (Andrich & Sheridan 1980). Methods for estimating parameters can be roughly classified into conditional and unconditional (Gustafsson & Lindblad 1978; Gustafsson 1979; Wright & Douglas 1977). Wright (1980) presents a classification of estimating methods and applications of Rasch measurement. Masters (1980) gives a good summary of rating models and their parameters. Application of different models have been published e.g. in journals: Journal of Educational Measurement (summer 77 among others), Studies in Educational Evaluation, Applied Psychological Measurement, Psychometrica, British Journal of Mathematical and Statistical Psychology and in the book edited by Spada and Kempf (1977).

The number of Rasch models has increased very quickly. Douglas (1980) gives a generalization and synthesis of the principles which underlie the structure of Rasch models.

His generic model incorporates all properties that we usually demand from any latent trait model. The generic model represents a probabilistic definition of the Rasch model in the sense that all models which appear in the literature under this rubric are derivable from the generic form. Models which do not fit into the mould cannot genuinely be called Rasch models. The generic model takes the form

$$(3.24) \qquad P(X_{i_1 i_2 \ldots i_t} = x_{i_1 i_2 \ldots i_t}) = \frac{\exp\left(\sum_{h=0}^{m} \theta_{i_1 i_2 \ldots i_t h} x_{i_1 \ldots i_t h}\right)}{\sum_{h=0}^{m} \exp\left(\theta_{i_1 i_2 \ldots i_t h}\right)}$$

where

m = number of categories

$h = 0, 1, \ldots, m$

$\theta_{i_1 i_2 \ldots i_t h}$ = general function of parameters, which is factorable into additive components

$x_{i_1 i_2 \ldots i_t h}$ = indicator variable (0 or 1)

For the special case of the SLM (3.24) can be written:

$$(3.25) \qquad P(X_{vi} = x_{vi}) = \frac{\exp\left(\sum_{h=0}^{1} (\beta_v - \delta_i) x_{vih}\right)}{\sum_{h=0}^{1} \exp\left(h(\beta_v - \delta_i)\right)} = \frac{\exp(\beta_v - \delta_i) x_{vi1}}{1 + \exp(\beta_v - \delta_i)}$$

after taking into account that

$i_1 = v$ and $i_2 = i$

$h = 0, 1$

$\theta_{i_1 i_2 h} = h\beta_v - h\delta_i$

$x_{vih} = \begin{cases} 1 \text{ if item correct} \\ 0 \text{ otherwise} \end{cases}$

Giving guidelines for future research Douglas (1980) emphasizes the use of conditional algorithms in numerical analysis problems and paying attention to developing powerful test of fit.

As we can easily see from the previous paragraphs, development in the area of Rasch models has happened very quickly in the last few years. Several new models have been developed, and methods for estimating parameters have become more accurate. A lot of knowledge is available from Rasch models, even if their mathematical and statistical form seems to be a hindrance for applying them in education and other areas of empirical research. In this study an attempt has been made to solve statistical problems in an analysis of Mathematics teaching by means of Rasch models. The problems are concerned with structural learning and could not be solved at all without the use of the linear logistic test model.

4. SIMPLE LOGISTIC MODEL AND TRADITIONAL METHODS IN THE MEASUREMENT OF ACHIEVEMENTS IN MATHEMATICS

The main aim of this chapter is
(1) to make a trial of item analysis using the simple logistic model side by side with the traditional methods, and linking together the items in primary level of comprehensive school;
(2) to do research work in comparing methods for item and person fit and comparing ways of getting reliable information from items and tests.

4.1. Item and person statistics

The data for this study consists of 14 tests, making a total of 360 items. The number of subjects answering the tests is 2467. Methods for analysing goodness of items arise partly from latent trait theory and partly from traditional test theory. Both viewpoints complement each other. On the contrary "person fit" is a concept of latent trait theory and it does not exist in traditional test theory.

Examination of items is interesting not only as a theoretical comparison of the two methods but also in a didactical sense for finding guidelines for test construction. The end of this chapter will highlight the topic of test

construction and it will be discussed again in chapter 5 from the viewpoint of the linear logistic test model.

4.1.1. Results of the SLM

The Rasch Simple Logistic Model includes only one item parameter δ_i:

$$(4.1) \qquad P\{X_{vi}=x_{vi} \mid \beta_v, \delta_i\} = \frac{\exp\{(\beta_v - \delta_i)x_{vi}\}}{1 + \exp(\beta_v - \delta_i)}$$

where

$x_{vi} = \begin{cases} 1 \text{ in the case of a correct answer} \\ 0 \text{ otherwise} \end{cases}$

β_v = person ability parameter

δ_i = item difficulty parameter

This means that the only property of an item is its relative difficulty. If an easy item is "good", it is easy for each person in the sample, and similarly for difficult and average items - they will be difficult or average for each person in the sample. In other words each item operates in the same way at different positions of the person continuum. An objective comparison of item difficulty is based on this fact. That is why we can consider the total score of a person to be a sufficient statistic for his β_v and the total score of an item to be sufficient for δ_i. This is usually the case in schools. The teacher gives his final marks in an exam by adding the marks for each item. The use of the SLM is based exactly on the same property of the test. In goodness of fit tests of the model, it is a question of checking if this kind of "adding" is allowable, that is to say if the sum of marks is sufficient for the person parameter.

4.1.1.1. Different t-statistics for item fit

Criteria for goodness of items are based on how well the items fit the SLM. Test statistics for item fit are based on t-distribution of residuals $\sum z_{vi}$ or χ^2-statistic comparing observed and expected item characteristic curves. In the NEWRATE-program, which has been used in this study there are two statistics of the former type and one of the latter (Andrich & Sheridan 1980).

Statistic T_1 is based on the sum of squares of the differences between x_{vi} and its expected value,

$$z_{vi} = \frac{x_{vi} - E(x_{vi})}{\sqrt{V(x_{vi})}} . \qquad (4.2)$$

The residual (4.2) has an expectation of 0 and a variance of 1 and its distribution is approximately t-distribution with degrees of freedom

$$f = \frac{(n-1)(k-1)}{nk} \qquad (4.3)$$

where

n = number of persons,

k = number of items.

For large samples the distribution can be considered to be normal. The test statistic is

$$T_1 = \frac{z_i^2 - E(z_i^2)}{\sqrt{V(z_i^2)}} = \frac{z_i^2 - f_i}{\sqrt{V(z_i^2)}} \qquad (4.4)$$

where

$$z_i^2 = \sum_v z_{vi}^2 .$$

However, the distribution of T_1-statistic proves to be non-symmetrical and the corrected version

$$(4.5) \qquad T_2 = \frac{\ln\left(\frac{z_i^2}{f_i}\right)}{\sqrt{V\left(\frac{z_i^2}{f_i}\right)}}$$

has been used for getting the shape nearer to the normal distribution. Even in the distribution of T_2-statistic there tend to be too many extreme negative values which skew the distribution. This does not overly distrub the test of fit. The statistic works very well in giving misfitting items $|T_2|>2$. Other variance stabilising transforms of T_1 could also be used and the same advantage for distribution would be gained. For example

$$(4.6) \qquad T_3 = \sqrt{\frac{z_i^2}{f_i}} \cdot \frac{z_i^2 - f_i}{\sqrt{V(z_i^2)}}$$

is a suitable test of fit statistic because of its approximately normal distribution, mean $\cong 0$ and variance $\cong 1$. The main drawback is that the shape of the negative tail of the distribution cannot be avoided, whether one used statistic T_2 or T_3. Variances of T_1, T_2 and T_3 statistics are approximately the same. They can be calculated from the first approximation of the Taylor series of g(t)

$$(4.7) \qquad V\left(g(t)\right) \cong \left(g'(t_o)\right)^{-2} V(t)$$

where t_o is the expected value of t (Ord 1972).

4.1.1.2. Comparison of statistics T_1 and T_2

A closer look at T_1 and T_2 statistic shows that

$$(4.8) \qquad T_2 = \frac{\ln\left(\frac{\Sigma z^2}{f}\right)}{\frac{\Sigma z^2}{f} - 1} \cdot T_1$$

where the transformation coefficient (symbolized as ω) is positive and, having called $\sum z^2/f = x$, we have

$$T_2 = \frac{\ln x}{x - 1} \cdot T_1 \tag{4.9}$$

The transformation coefficient

$$\omega = \frac{\ln \left(\frac{\sum z^2}{f}\right)}{\frac{\sum z^2}{f} - 1} \tag{4.10}$$

is illustrated in Figure 4. It is easy to see that $\lim_{x \to 1} \frac{\log x}{x-1} = 1$, which means that T_1 and T_2 give exactly the same results if $\sum z^2 = E\left(\sum z^2\right) = f$. Otherwise T_2-transformation always gives smaller values than T_1 does. The reason for this argument is that

$$\begin{cases} |T_2| < |T_1| & \text{if } \frac{\sum z^2}{f} > 1 \\ |T_2| > |T_1| & \text{if } \frac{\sum z^2}{f} < 1 \end{cases} \tag{4.11}$$

or we can say that

$$\begin{cases} |T_2| < |T_1| & \text{if } T_1 > 0 \text{ (that is } \sum z^2 - f > 0) \\ |T_2| > |T_1| & \text{if } T_1 < 0 \end{cases} \tag{4.12}$$

$$T_2 < T_1 \quad \forall \frac{\sum z^2}{f} \tag{4.13}$$

because ω is always positive which implies that both T_1 and T_2 are both positive, both negative or both equal to zero.

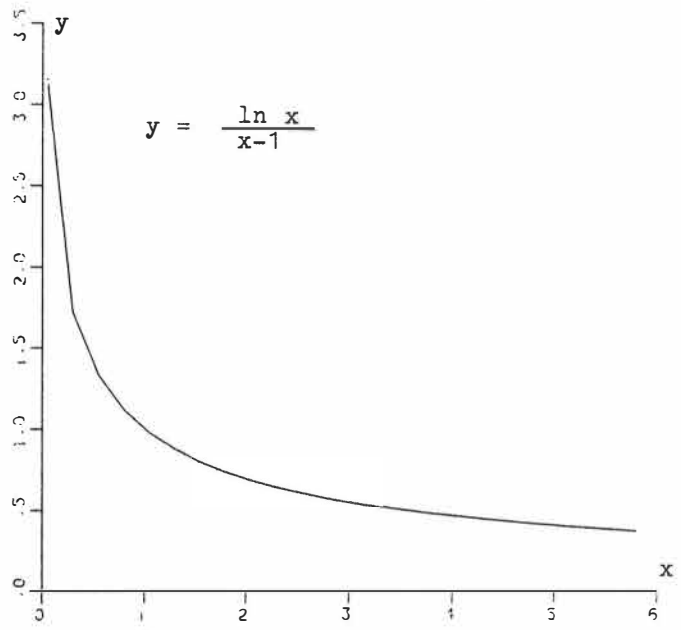


Figure 4. Transformation coefficient ω

Both T_1 and T_2 are functions of variance $V(\sum z^2)$. If we want to compare the values of those two statistics this can be done for each value of variance separately. As an example let us take the test for the highest level group of 9th grades (Figure 5). The bigger the variance of $\sum z^2$ is the effectively the values of T_2 are after transformation concentrated to zero.

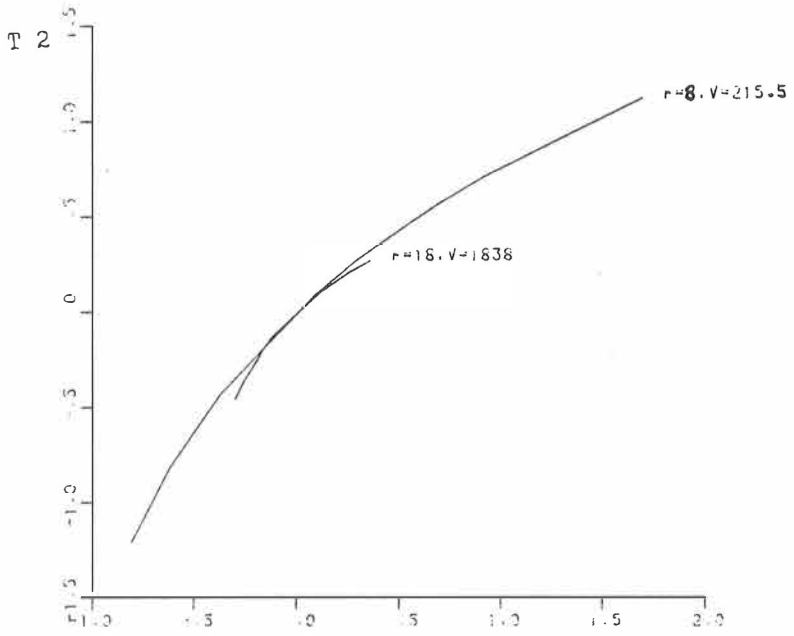


Figure 5. T_2 as a function of T_1 when $V(\sum z^2)$ is constant

If the test fits the model $|T_2|$ is relatively small. That is why it is difficult to draw illustrating curves for constant variances in case of tests used in this study. However, if we take test number 9 with several coding errors (the test has been rejected from the final analysis) we can find many misfitting persons and curves for constant variances are very clear (Figure 6). If variance is constant it implies that T_2 is a function of $\sum z^2$ and degrees of freedom only.

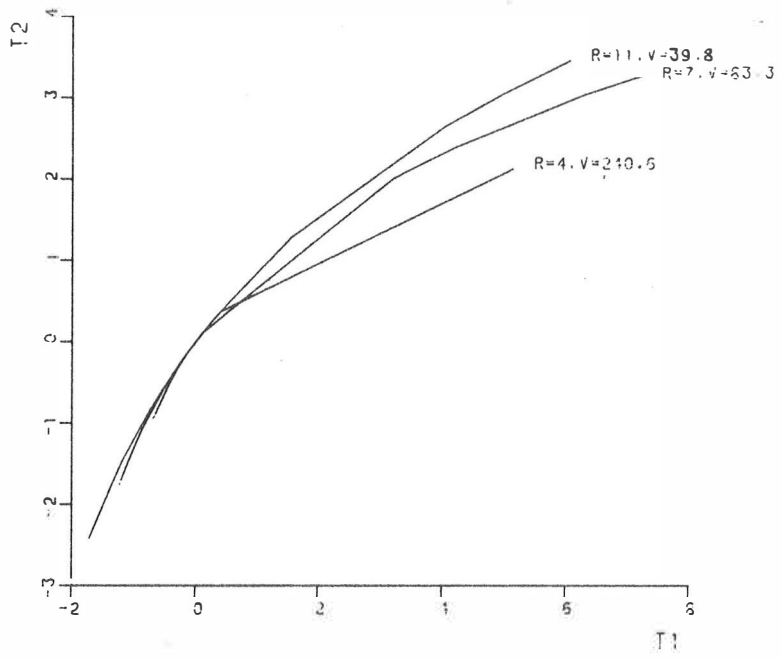


Figure 6. T_2 as a function of T_1. Tests 7, 8 and 9 together.

A Closer look at skewness and kurtosis of T_1 and T_2 distributions is worth doing in order to give reasons for using T_2-statistic in the test of fit. Skewness is

$$Sk = \frac{\sum_{v=1}^{n} (\text{variable} - \text{expected value})^3}{n \cdot (sd)^3} \qquad (4.14)$$

In the equation of kurtosis there is 4 instead of exponent 3 in (4.14). Variances of T_1 and T_2-distributions are approximately the same which implies that the difference in skewness of the distributions in based on the differences in numerators. The numerator in skewness of distribution of T_2-statistic is symmetrical in exponential scale. In the case of small values of $\sum z^2/f$ both statistics are equally skewed, the difference becomes evident if $\sum z^2 << f$ or $\sum z^2 >> f$ (Figure 7).

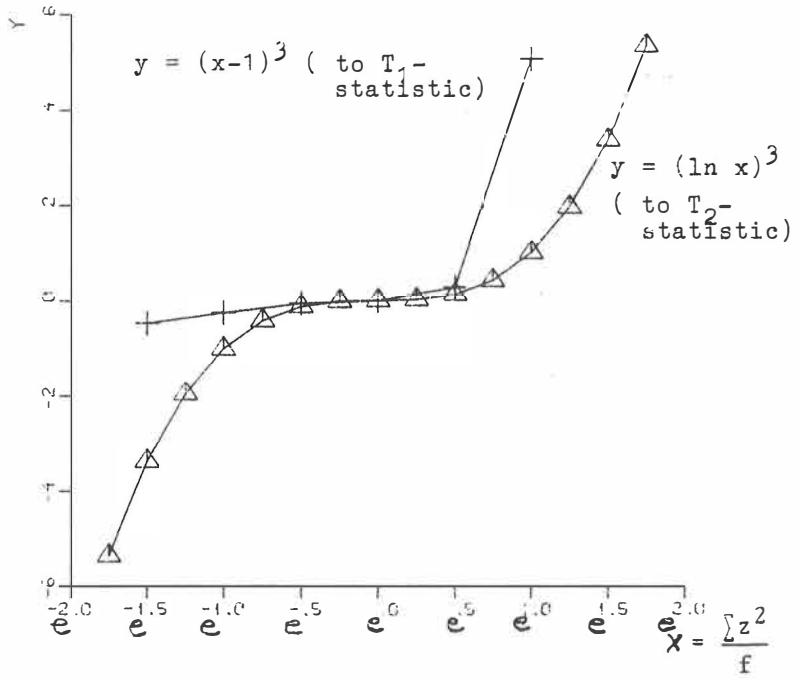


Figure 7. One term in expression of numerator of skewness

In analysis of kurtosis the corresponding curves are presented in Figure 8.

Also in this case the curves differ clearly if $\sum z^2 >> f$ or $\sum z^2 << f$. The symmetry of the curves in the case of T_2-statistics is the main reason for considering T_2-statistic to be better than T_1.

4.1.1.3. Variance of z_{vi}^2 as a function of $\beta_v - \delta_i$

In the denominator of skewness and kurtosis there exists the variance $V(\sum z_{vi})$ which is a function of $\beta_v - \delta_i$. Andrich and Sheridan (1980, p. 16) give the function

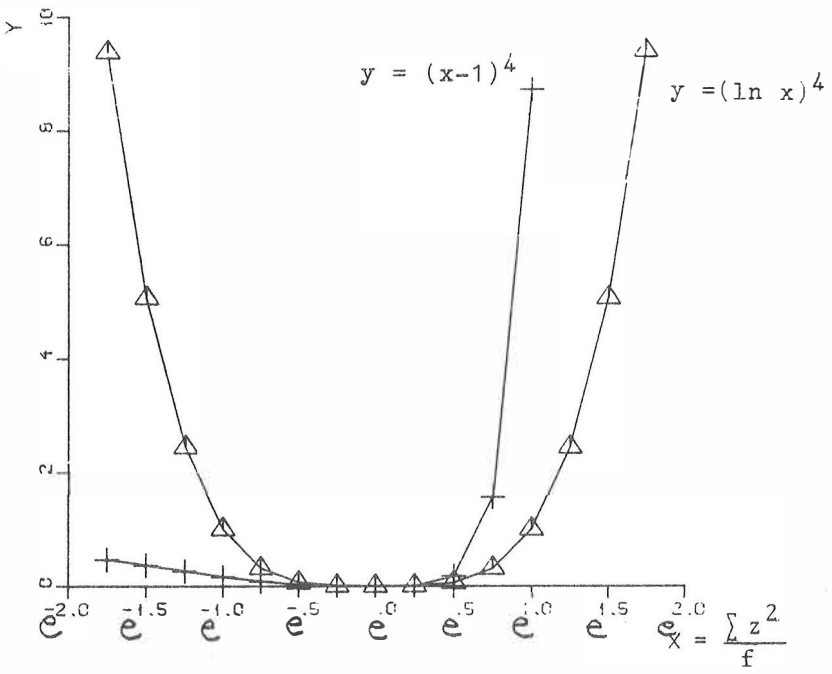


Figure 8. One term in expression of numerator of kurtosis

(4.15) $$V\left(z_{vi}^2\right) = 2 \tanh\left(\frac{\beta_v - \delta_i}{2}\right) \cdot \sinh\left(\beta_v - \delta_i\right) .$$

This expression can be simplified $\left(\frac{\beta_v - \delta_i}{2} = x\right)$

$$V\left(z_{vi}^2\right) = 2 \cdot \frac{\sinh x}{\cosh x} \cdot \sinh(2x)$$

$$= \frac{e^x - e^{-x}}{e^x + e^{-x}} \cdot \left(e^{2x} - e^{-2x}\right)$$

$$= \left(e^x - e^{-x}\right)^2$$

(4.16) $$V\left(z_{vi}^2\right) = \left(e^{\frac{\beta_v - \delta_i}{2}} - e^{\frac{\delta_i - \beta_v}{2}}\right)^2$$

$$= e^{\beta_v - \delta_i} + e^{\delta_i - \beta_v} - 2$$

Figure 9 gives the function and shows that the values are big if β_v is far from δ_i. On the logarithmic scale the corresponding curve is near to the straight line $y = x^* - 2$, where $x^* = \ln |\beta_v - \delta_i|$ (Figure 10).

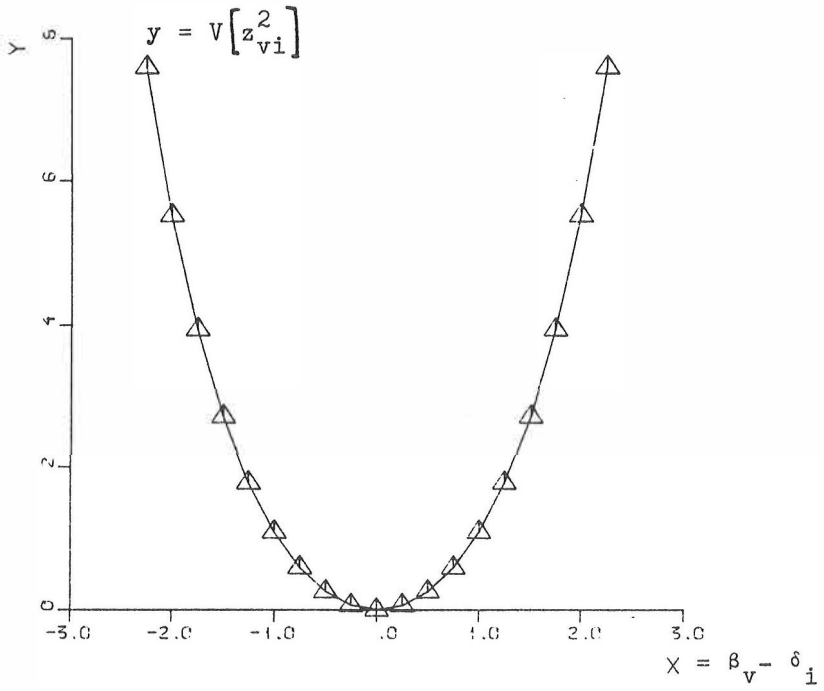


Figure 9. $V\ (z^2_{vi})$ as a function of the difference $\beta_v - \delta_i$

$V\ (z^2_{vi})$ can be expressed also as a function of probability of getting the item right:

$$V\ (z^2_{vi}) = \exp\ (\ln \frac{P}{1-P}) + \exp\ (\ln \frac{1-P}{P}) - 2$$

$$= \frac{P}{1-P} + \frac{1-P}{P} - 2$$

(4.17)
$$V\ (z^2_{vi}) = \frac{(2P - 1)^2}{P\ (1 - P)}$$

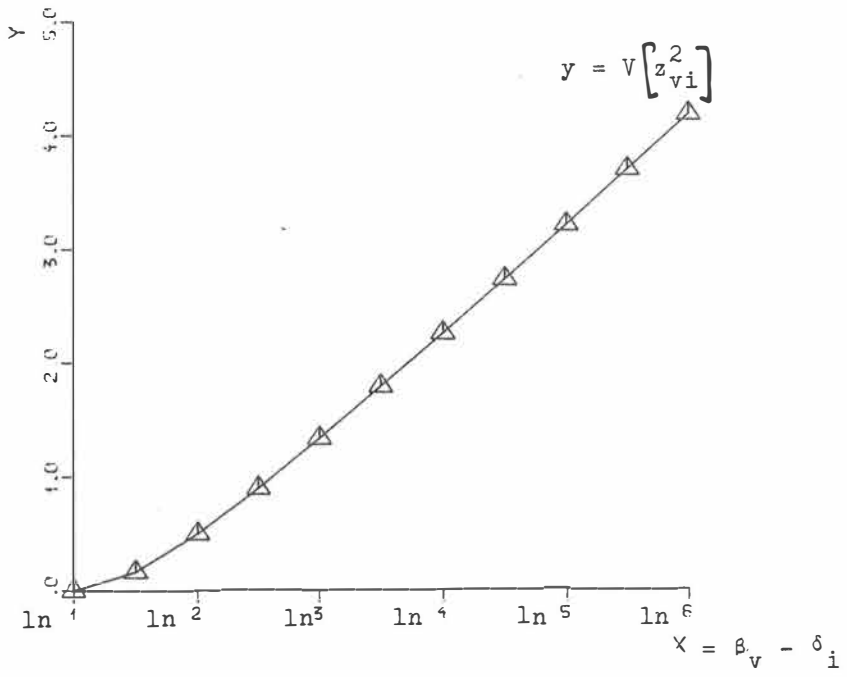


Figure 10. Figure 9 with logaritmic scale

The variance of z_{vi}^2 gives the result that the power of the test of fit is low in the case of small $|\beta_v - \delta_i|$.

4.1.1.4. The χ^2-statistic for item fit

The other test of fit for items involves cheking that the observed item characteristic curve follows that expected from the model. Persons are divided into class intervals according their total scores and then in each class interval the total scores on an item are compared with their expected values. In each group the total score of item i is

$$s_{gi} = \sum_{v \in g} x_{vi} \qquad (g = 1,\ldots,G). \tag{4.18}$$

For each item the statistic

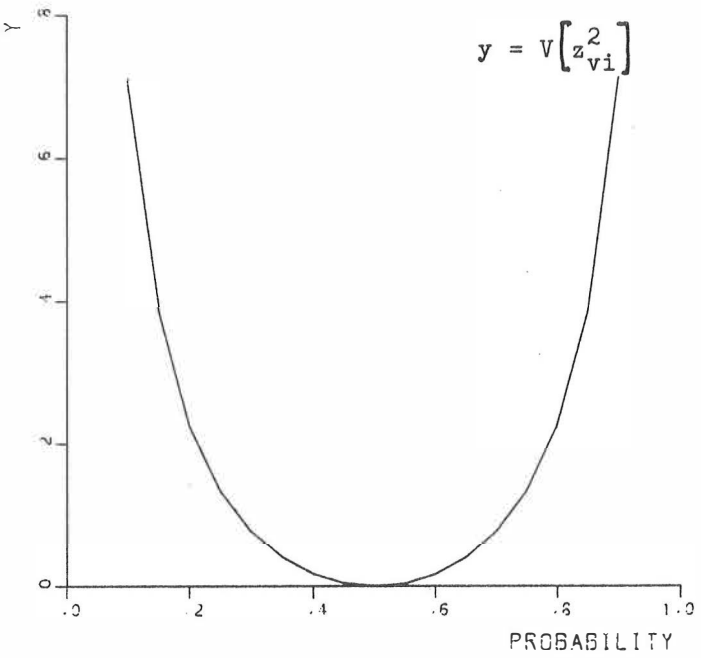


Figure 11. $V(z_{vi}^2)$ as a function of probability of getting the item right

$$\chi_i^2 = \sum_{g=1}^{G} \frac{(s_{gi} - E(s_{gi}))^2}{V(s_{gi})} \tag{4.19}$$

may be taken to approximate χ^2-distribution with

$$f = \frac{(k-1)(G-1)}{k} \tag{4.20}$$

if the data conforms to the model. For each group standardized residuals are computed in NEWRATE,

$$d_{gi} = \frac{n_{gxi} - n_g \hat{p}_{xgi}}{\sqrt{n_g \hat{p}_{xgi}}} \tag{4.21}$$

where

$\hat{p}_{xgi}$ = estimated probability that a person from class interval g will respond in category x (x=1 or 0),

n_g = number of persons in g,

n_{gxi} = observed number in g who respond in category x.

Methods based on T-statistics and χ^2-statistics do not necessarily give the same results. For example, in the data of the first graders the correlation between order of the item fit is .27.

4.1.1.5. Transformation of fitting items to the SLM metric

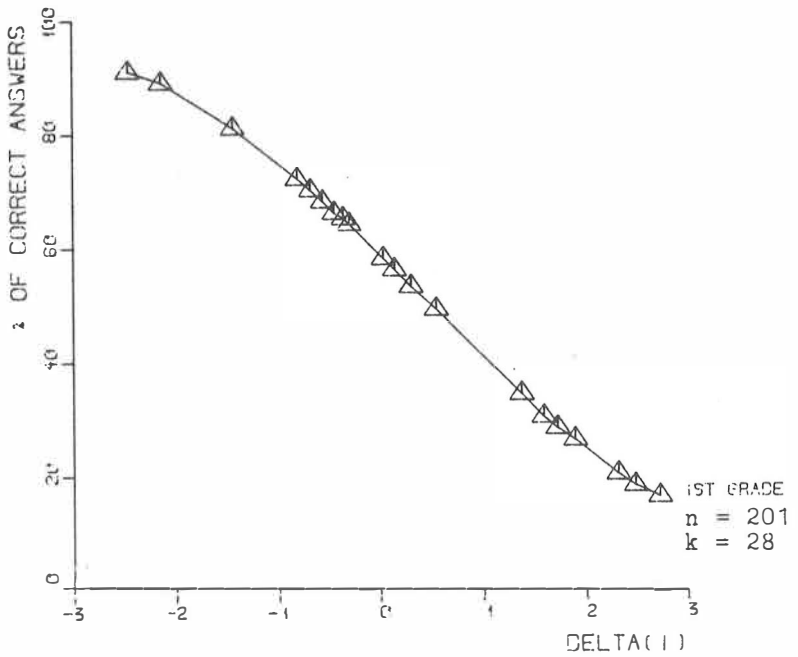


Figure 12. The transformation from the percentage of correct responses to item difficulty

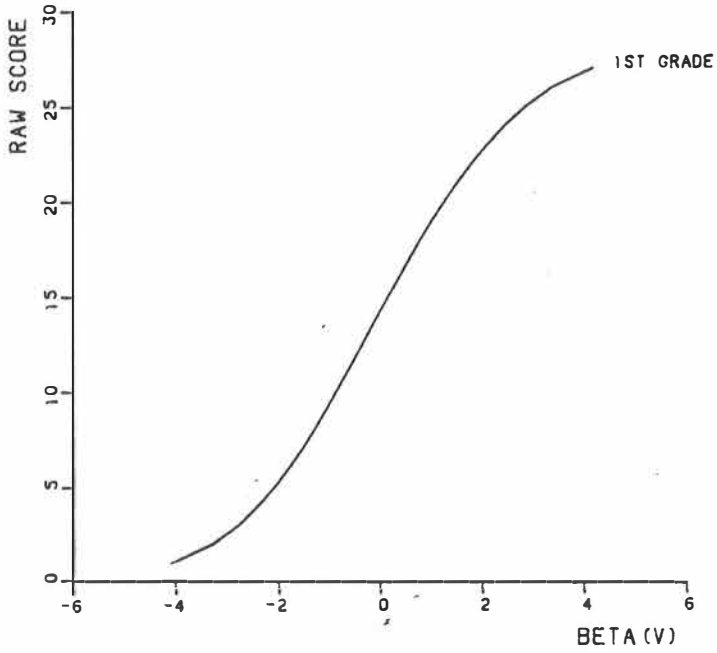


Figure 13. Transformation of raw scores to abilities, 1st grade

The transformation from the percentage of correct responses to estimates of item parameters is illustrated in Figure 12 in the case of test 1.

The corresponding transformation from total scores to abilities in presented in Figure 13. After this transformation frequence distribution of transformed scores is nearer normal distribution than it was before transformation (Figure 14)

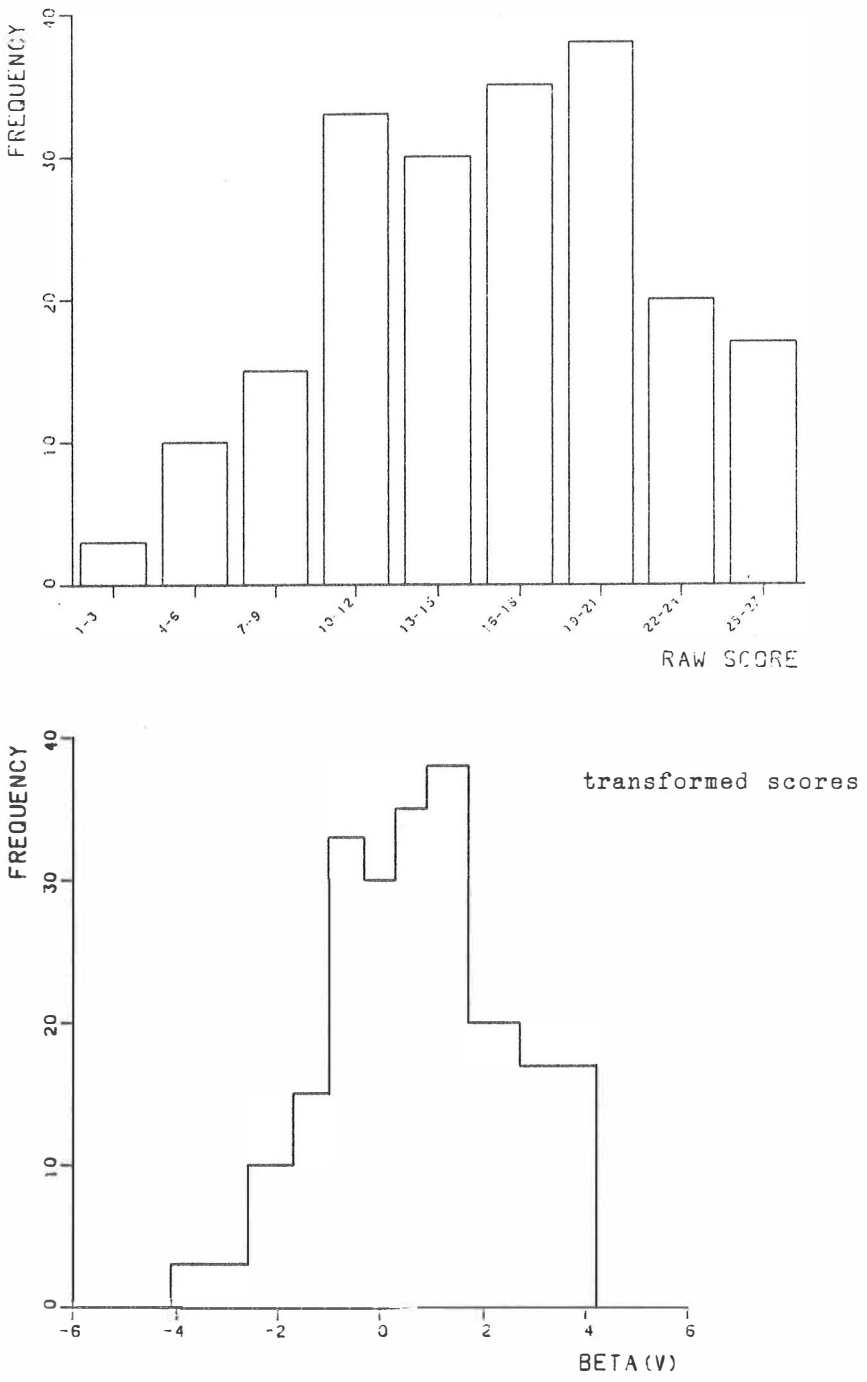


Figure 14. Raw score and transformed score frequency distributions, 1st grade

4.1.1.6. Person fit and separation

In the tests of person fit the NEWRATE-program uses T_1 and T_2-statistics mentioned before. The only difference in computing item fit is that now sums are computed over items, and degrees of freedom are related to the number of persons instead of items. The program gives response vectors of misfitting persons and compares them to vectors of expected responses. In the case of misfit ($T < -2$ or $T > 2$) the response vector is inconsistent: a person has solved some difficult items correctly but not the easy ones. The model does not usually give the most capable or the least capable persons in the list of misfits. This is obvious because of their responses and also because of a big variance $V(\sum z_{vi}^2)$ in extreme cases which is in the denominator of the T_1-statistic.

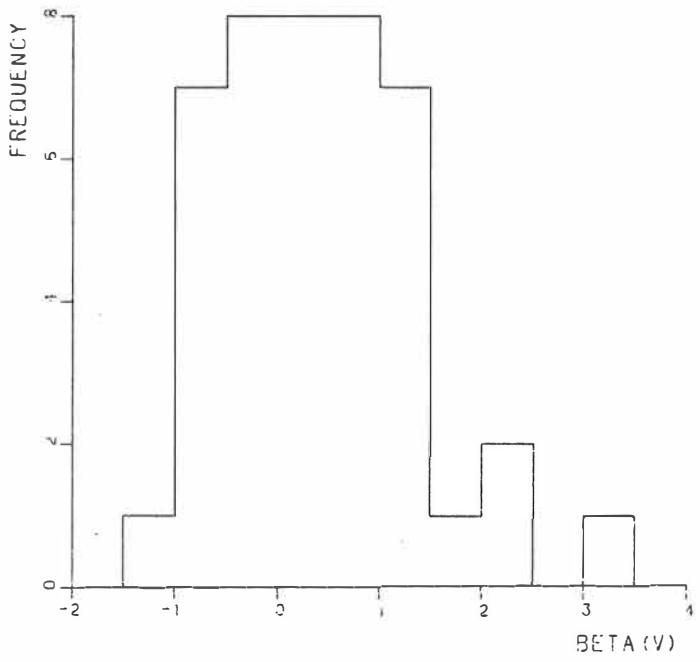


Figure 15. Frequency distribution of abilities of misfitting persons

The test of person fit has high power if the inter-subject separation index r_β is high (Andrich & Sheridan 1980). This index

(4.22) $$r = \frac{V(\hat{\beta}) - \bar{\delta}^2_{\hat{\beta}V}}{V(\hat{\beta})}$$

is relatively high in each grade (Table 15). The mean of all grades is .86. The corresponding index in traditional test theory is Cronbach's alpha. For the lowest level groups in the 7th, 8th and 9th grades the separation is not as high because of the small variance in ability distribution.

Table 15. Separation indices (only fitting items included)

Grade	r_β	
1	.8692 ≅ .87	
2	.8585 ≅ .86	
3	.8736 ≅ .87	
4	.8905 ≅ .89	
5	.8561 ≅ .86	$\bar{r}_\beta$ = .86
6	.8837 ≅ .88	sd = .03
7	.7987 ≅ .80	
8	.8200 ≅ .82	
9	.8645 ≅ .86	
7 lower level group	.5611 ≅ .56	
7 upper	.8010 ≅ .80	
8 lowest	.4953 ≅ .50	$\bar{r}_\beta$ = .70
8 middle	.6382 ≅ .64	sd = .14
8 highest	.7878 ≅ .79	
9 middle	.8241 ≅ .82	
9 highest	.8164 ≅ .82	

The only misfitting person who has got $\hat{\beta}_v > 3$ is a person in the third grade. Because of the high ability estimate all his resposes are expected to be correct. However, he has errors in items 1 (the easiest item with $\hat{\delta}_1 = -4.14$) and 3 ($\hat{\delta}_3 = .12$). These two items are enough for making a person misfitting because of the high sum of squares (Wright & Stone 1979, p. 76). In this case there is a considerable deviation in the values of T_1- and T_2-statistics even if both show that person to be misfitting (Table 16). In the case of the next person in the same test there are 5 items which he has answered correctly even though their difficulty is greater than his ability ($\hat{\beta}_v = -.6407$) and 3 items which he has got wrong although they are expected to be correct ($\hat{\delta}_i < \hat{\beta}_v$).

The number of misfitting persons in these tests is small (Table 17). If there was more misfit in person statistic it would be reasonable to have a closer look at the reasons for misfit. It might be that some topics in the equations have not been taught to some pupils which would be inconsistent with unidimensionality of persons. If the same items for each misfitting person caused misfit there should be a check to see if something is wrong in the items (language used, etc.) or if it is a question of lack in unidimensionality either at the item or the person level.

Table 16. Response vector of a misfitting person

total score	ability	residual sum of squares	variance	T_1	T_2
28	3.2606	1905.30	3165.93	33.35	2.15
observed	0 1 0	1111111111111111111111111			
expected	1 1 1	1111111111111111111111111			
difference	-1 0-1	0000000000000000000000000			

Table 17. Summary of number of misfitting items and persons

Grade	Number of items	Number of misfitting items	Number of misfitting persons $T_2 > 2$	Number of misfitting persons $T_2 < -2$	Percentage of misfitting persons	Percentage of misfitting items
1	30	2	6	1	3.5	6.7
2	30	3	3	3	2.8	10.0
3	30	2	2	0	1.0	6.7
4	30	0	6	1	4.8	0.0
5	30	0	4	0	1.7	0.0
6	30	0	2	0	0.8	0.0
7 lower	22	3	0	0	0.0	13.6
7 upper	22	1	2	0	1.0	4.5
7 all	22	1	1	0	0.3	4.5
8 lowest	22	2	0	0	0.0	9.1
8 middle	22	2	0	0	0.0	9.1
8 highest	24	2	0	0	0.0	8.3
8 all	22	2	3	0	0.7	9.1
9 middle	22	3	1	0	0.5	13.6
9 highest	24	3	2	0	0.9	12.5
9 all	22	3	6	0	1.3	13.6

In conclusion we can say that less than 7.5 % of items and less than 1.5 % persons are misfitting in terms of the T_2-statistic in the NEWRATE program.

4.1.1.7. Reasons for misfit

In the previous paragraph we gave examples of misfitting persons for which $T_2 > 2$. Those persons had solved some hard items correctly and some easy items incorrectly. In the case of $T_2 < -2$ it is a question of a person whose response model is "too-good-to-be-true" and cannot be accepted in the probabilistic model. For example person number 49 in the test I has

got the 17 easiest items right and the 11 hardest items wrong from 28 fitting items. He is a misfitting person because of too complete a response pattern (T_2 = -2.35).

Misfitting items in the sense of T_2-statistic in the first grade are items 1 (5 + 3 = $\frac{8}{16}$) and 24 ($-\frac{48}{16}$). Item 1 is the easiest item in the whole test (96 % correct answers): only 9 pupils have got the item wrong. Total scores of these 9 are: 4, 5, 11, 12, 16, 17, 17, 18 and 20. It can be seen immediately that most of total scores are relatively high. It means that residuals in the T_2-statistic become high, which makes T_2 > 2. Also in the case on χ^2-statistic, misfit can be seen in deviations of observed and expected item characteristic curves, (Figure 16). Both observed curves are surprisingly similar. In both cases the item does not discriminate between lower and higher ability groups well enough. In the case of item 1.1. 4 % pupils is enough for the item to misfit. Item 1.24 exists also in the test for second graders. It is misfitting also in test 2 (item 2.6) for the same reason as in test 1, even if misfit is not as clear (T_2 = 2.04) as it was in test 1 (T_2 = 2.79).

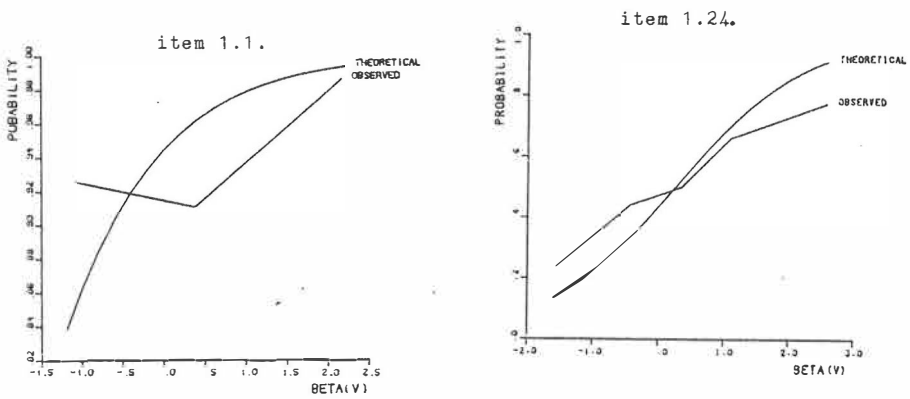


Figure 16. The misfitting items in test 1, item characteristic curves (ICC).

Having drawn all item characteristic curves for misfitting items we can see that in most cases the reason for misfit is too low a discrimination. Only in items 2.22, 2.26, 3.17 and 8.21 are there deviations from monotonity and in item 2.22 there is too high a separation in some class intervals and too low a separation in some others. All ICC´s for misfitting items are presented in Figures 17 and 18.

The remaining items are good in the sense of Rasch latent trait theory. That is to say:

(1) they are suitable for objective measurement,
(2) the sum of scores in the items gives all the information of the whole test.

The main idea on the Rasch models is the assumption that the probability of a correct response can be presented in additive form:

$$(4.23) \qquad M(p_{vi}) = \beta_v + \delta_i$$

for all persons (v) and items (i) in the case of some monotonic transformation M. In the case of the SLM the transformation is

$$(4.24) \qquad M(p_{vi}) = \ln \left(\frac{p_{vi}}{1 - p_{vi}}\right) .$$

Obviously, the p_{vi} are unobservable, as are the ability and item parameters. Estimates can be calculated by several methods (e.g. Perline & Wright & Wainer 1979, Wright 1980).

It has been shown that in the case of our 14 tests, when misfitting items are rejected, the probability of correct answer can be presented by means of ability and difficulty parameters, because the model holds.

4.1.2. Results of traditional test theory

Traditional analysis of tests is based on two indices (Wilmut 1975, p. 25)

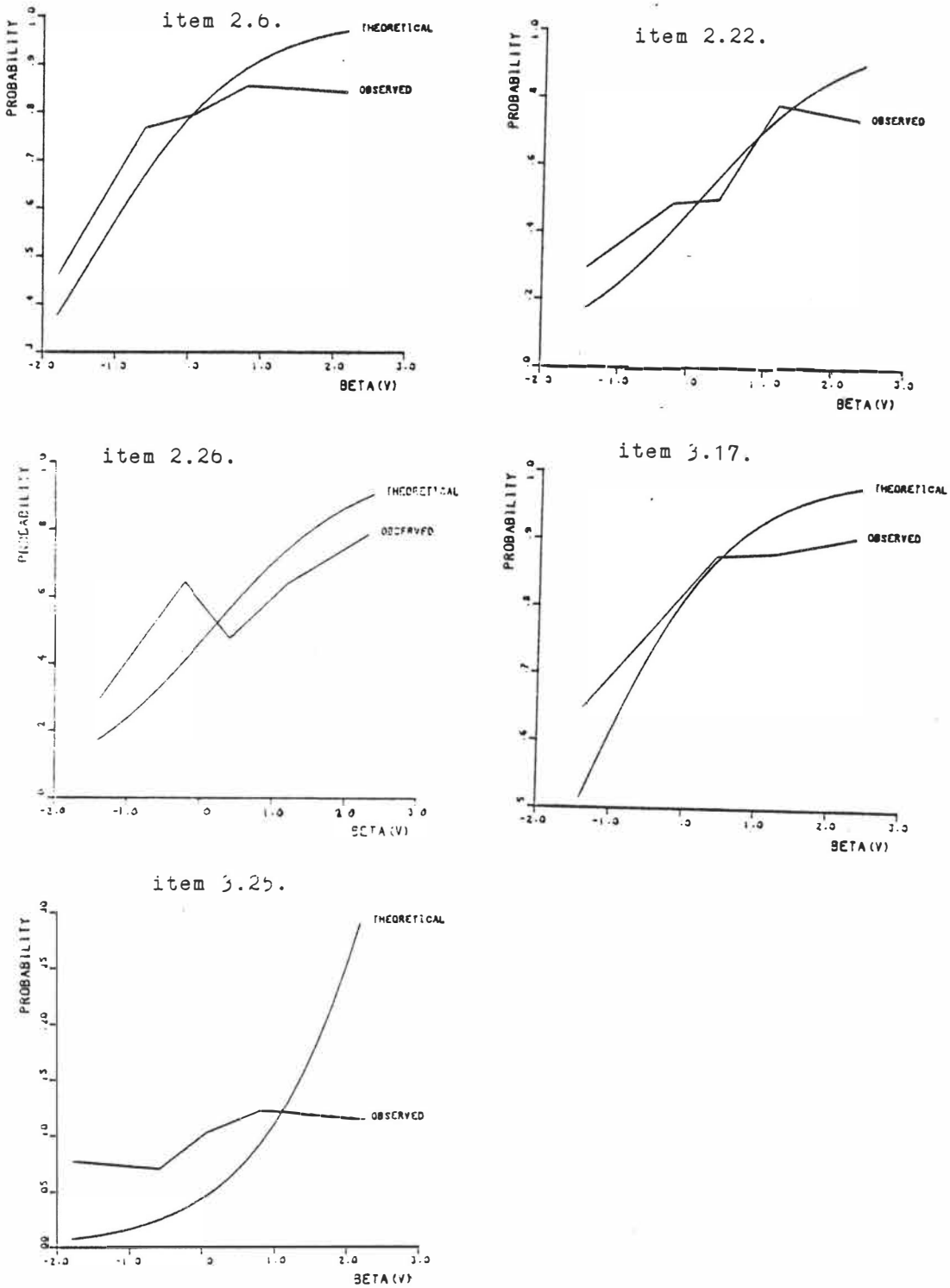


Figure 17. The misfitting items in 2nd and 3rd graders, ICC-curves

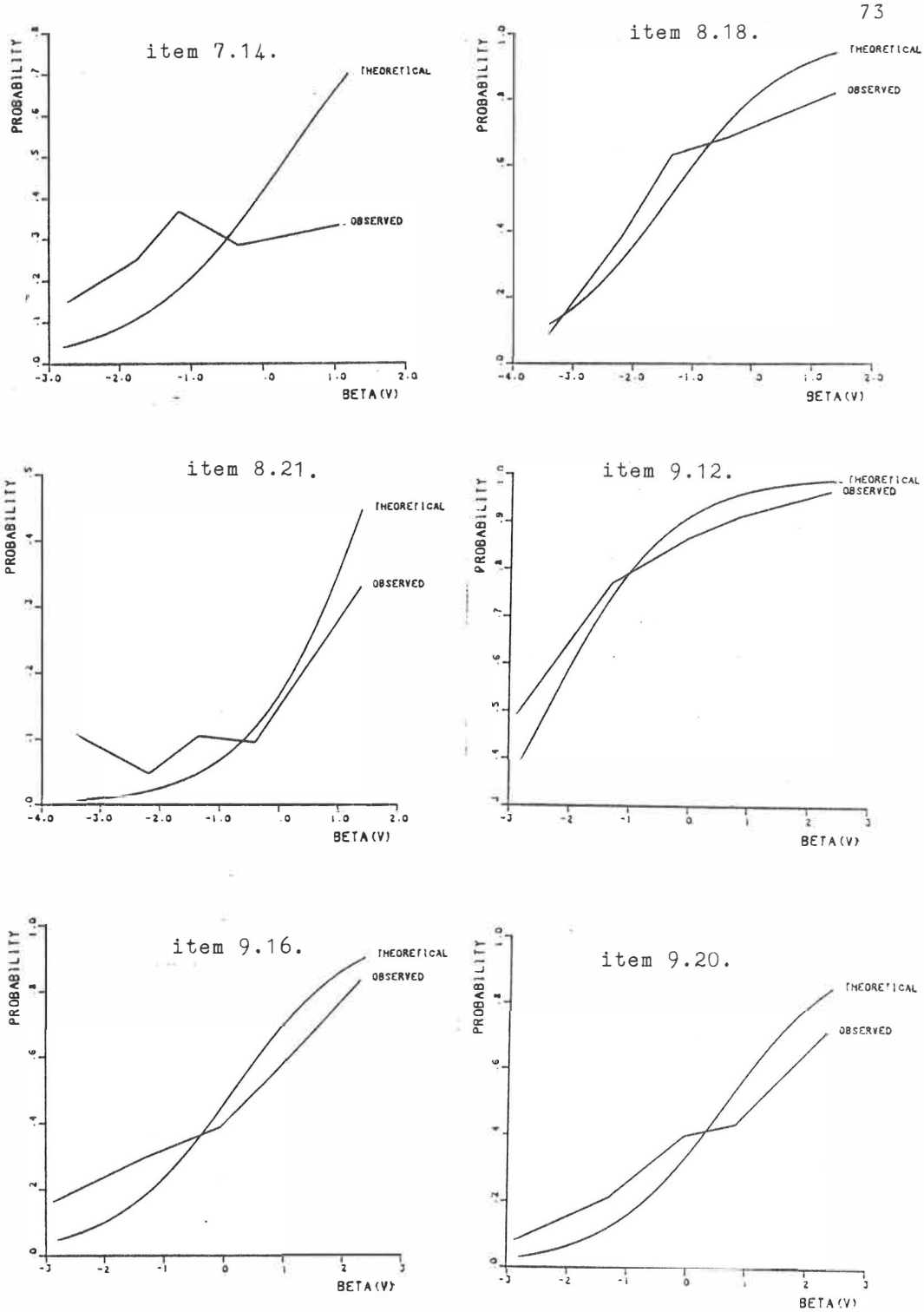


Figure 18. Misfitting items in 7th, 8th and 9th grades

(1) the Facility Index (the proportion of the candidature attempting the item, who get it right) and
(2) The Discrimination Index (a correlation between item scores on a 0/1 basis and candidate test scores).

It has been usual to accept an item whose discrimination index exceeds some arbitrary value (often taken as .15 or .20) and whose facility index is not excessively large or small.

4.1.2.1. Percentage of correct answers

As an example of the facility index we may take a closer look at the common items of grades 7, 8 and 9. (The item which is common to each grade from the first to the ninth has already been analysed). The distribution of 11 items has been presented in Figures 19a, b and c. The x-axis is the grade and level group, the y-axis the percentage of correct responses. The items in Figure 19a represent the type of equations in which learning has happened in each level group and in each grade. These are "basic" equations, and learning has occured just as desired. The percentage of correct responses is a good way of illustrating learning.

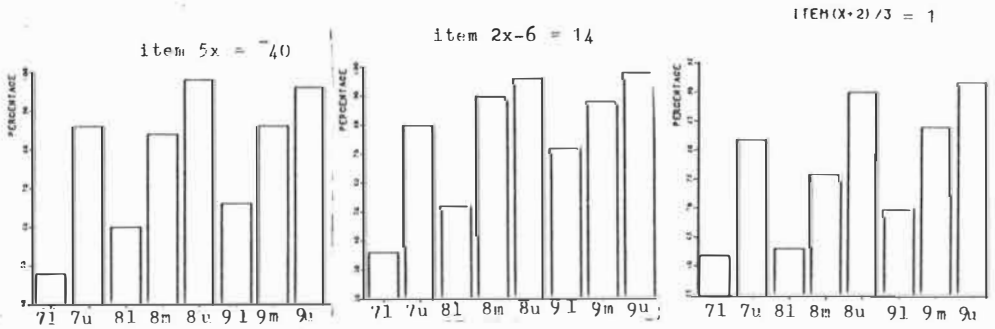


Figure 19a. Percentage of correct responses of three common items for junior secondary level

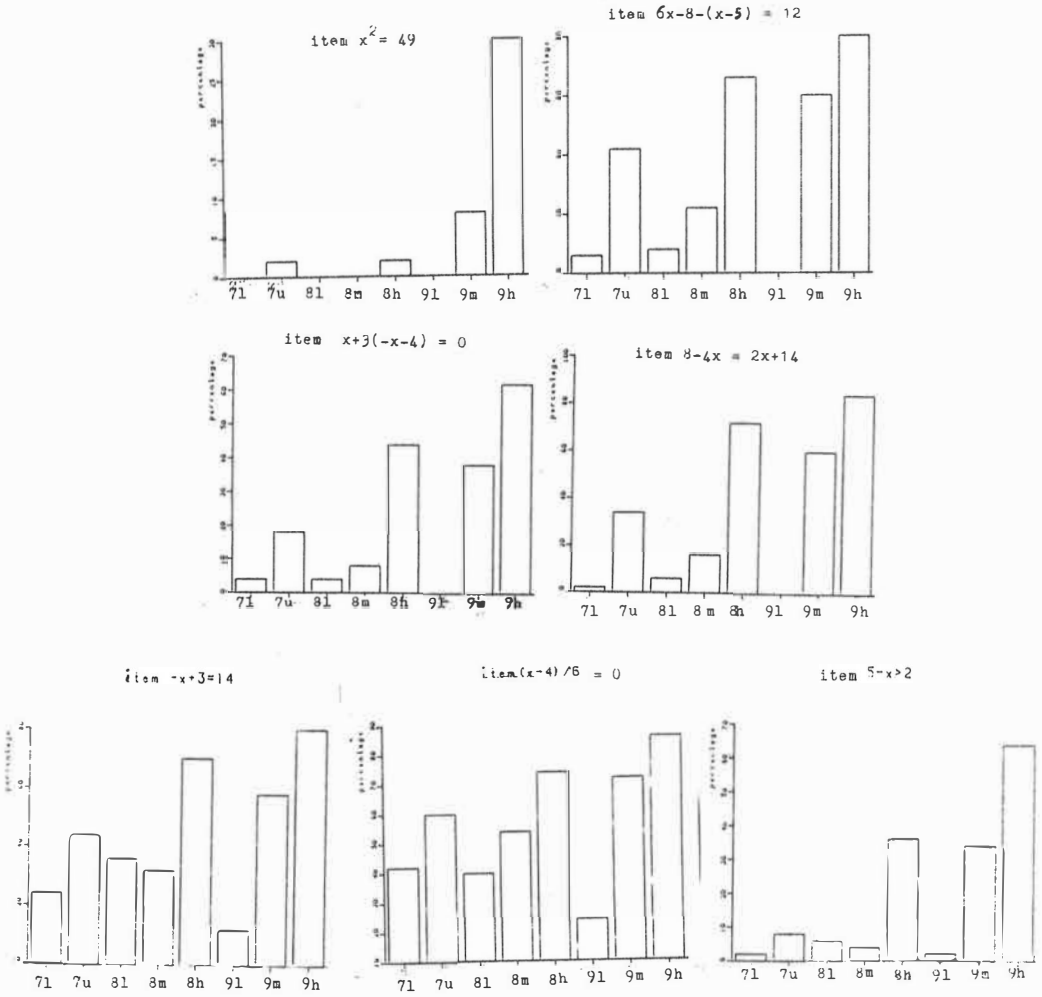


Figure 19b. Percentage of correct responses of seven common items for junior secondary level

Figure 19b shows the items for which the 9th lowest level group is the worst group of all in solving them. At the same time the upper level groups have learned these kinds of equations quite well. This figure gives a hint for the Mathematics syllabus: too many things (in this case, too complicated equations) have been taught to the lowest level groups, learning does not take place in the desired amount.

In Figure 19c there is one item which does not belong to the previous groups but is somewhere between the two groups. Also in the case of other the inequation (5 - x > 2) eighth graders in the middle level group have got fewer correct answers than expected.

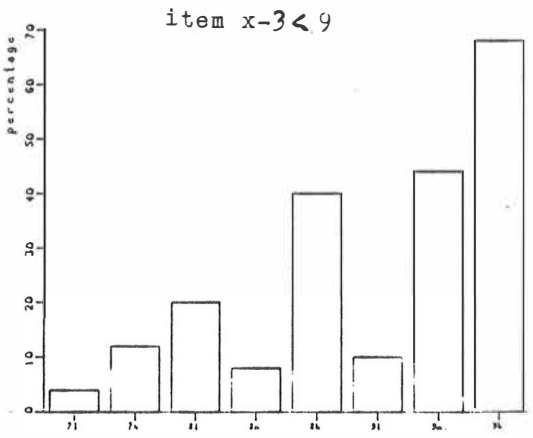


Figure 19c. Percentage of correct responses of a common item for junior secondary level

4.1.2.2. Item-test correlation

In traditional test theory the goodness of items is based on item test correlations. Table 18 is a summary of correlations in the OSANA-program (Konttinen & Kortelainen 1979) at primary level and Figure 20 illustrates the frequence distribution of correlations.

Table 18. Biserial item-test correlations of 180 primary school items

Grade	r_{bis} < .35	.35-.39	.40-.44	.45-.49	.50-.54	.55-.59	.60-.64	.65-.69	.70-.74	.75-.79	.80-.84	$\bar{r}_{bis}$	s_r	$\bar{p}$	s_p
1	1	1	1	2	1	3	5	7	6	2	1	.62	.13	59	23
2	0	3	3	0	2	2	3	8	5	3	1	.62	.13	60	19
3	1	1	1	2	0	5	8	2	3	4	3	.63	.15	54	23
4	0	0	0	0	1	2	5	8	8	4	2	.69	.08	52	20
5	2	0	1	3	3	3	9	5	3	1	0	.59	.13	53	21
6	0	0	0	2	0	3	8	6	6	3	2	.67	.09	53	22

In the previous table $\bar{p}$ is the mean of the percentage of correct answers in each grade and s_p its standard devitation.

For "poor" items we can choose different limiting values of r_{bis}, for example:

Table 19. Possible limits of r_{bis} for poor items

maximum r_{bis}	number of poor items in primary grades
.15	1
.20	3
.25	3
.30	3
.35	4
.40	9
.45	15
.50	24

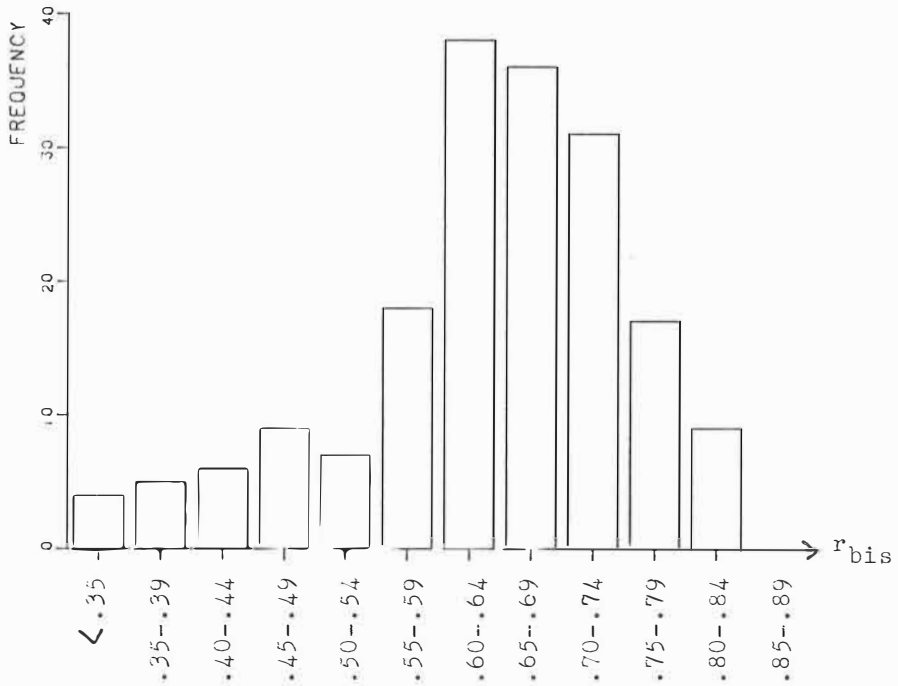
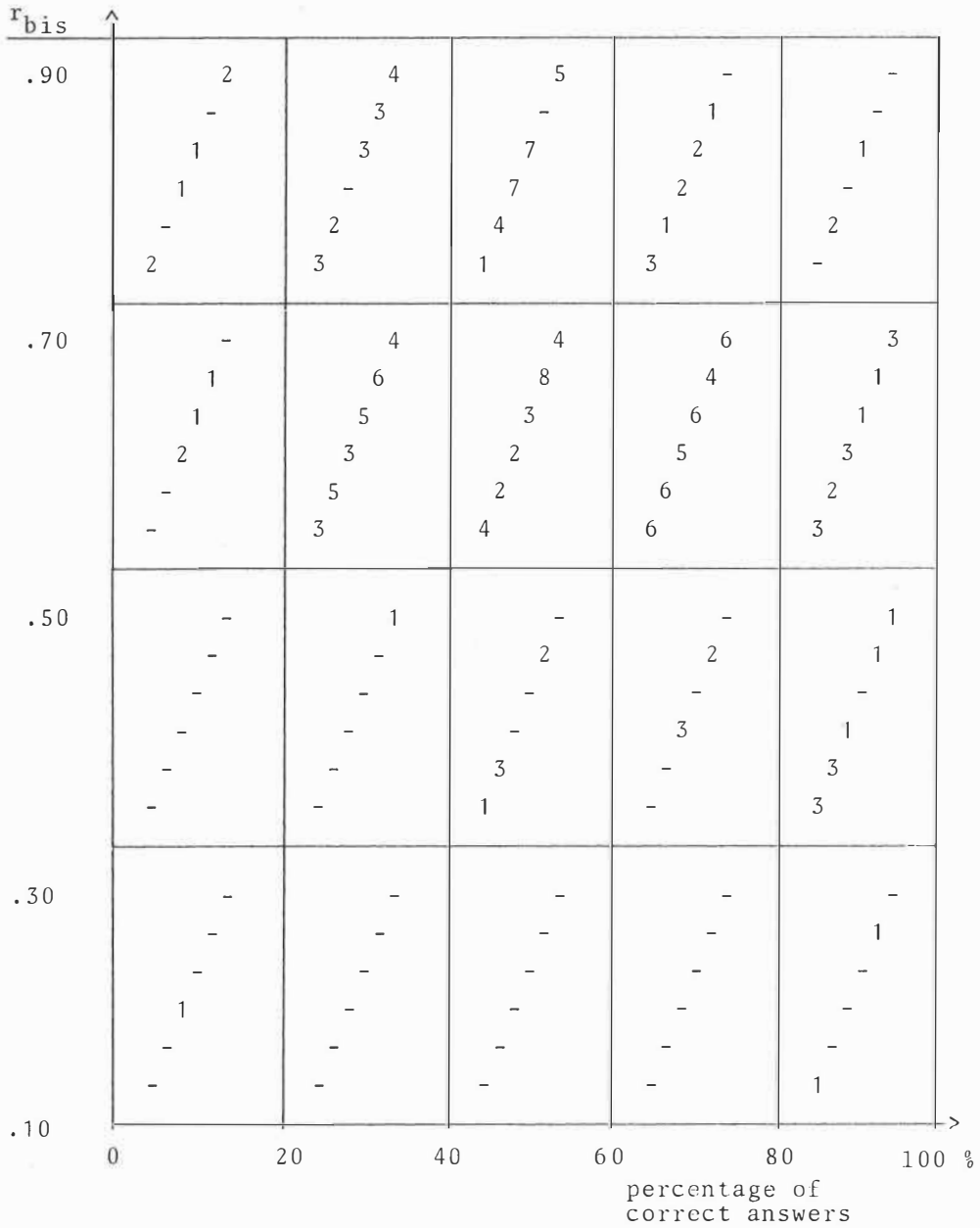


Figure 20. Frequency distribution of biserial item-test correlations in the primary grades

It seems reasonable to take r_{bis} = .40 as the limit. Poor items are 1.1, 1.7, 2.1, 2.6, 2.26, 3.5, 3.25, 5.1 and 5.10. In the sense of the SLM the misfitting items are 1.1, 1.24, 2.6, 2.22, 2.26, 3.17 and 3.25. The number of common rejected items in each method is half of the total number of rejected items.

Item-test correlation is the relation of the percentage of correct responses. It is interesting to compare the primary grades and the lowest level groups of the 7th, 8th and 9th grades. The relation at the primary level is very

Table 20a. Item-test biserial correlations in primary grades as a relation of the percentage of correct answers. Table of frequencies.



different from that of the lowest level groups. Correlation Table 20 is constructed so that the frequencies of each grade can be seen separately:

. . .
3rd grade
2nd grade
1st grade

The corresponding contingency table (Table 21a) shows that 72 % of cases are inside the square 20 % < percentage < 80 %, .50 < correlation < .90. Item-test correlation is > .50 in 87 % of all items.

Table 21a. Contingency table (summary of previous table)

r					
.90	6	15	24	9	3
.70	4	26	23	33	13
.50	1	1	6	5	11
	20	40	60	80	> %

The corresponding tables 20b and 21b for the lowest level groups show that in most cases the percentage of correct answers is low and so also is the item-test correlation. One reason for this must be that these tests include 10 multichoice items of the 22 items. Only 21 % of cases are situated in the square with limits 20 % ... 80 % and .50 ... .90. Spearman-Brown reliabilities of the tests are as low as .588, .555 and .344 respectively.

Table 20b. Item-test biserial correlations in lowest level groups of the 7th, 8th and 9th grades as a relation of the percentage of correct answers

r_{bis}	20	40	60	80	100
.90					
.70	1 4 -	1 - -	- - -	- - -	- - -
.50	1 2 5	2 2 1	3 3 -	1 - 1	- 1 -
.30	2 3 4	5 2 1	- - -	- - 1	- - -
.10	1 2 1	- 1 -	1 - -	- - 1	1 - -
.00	3 2 5	- - 1	- - -	- - -	- - 1

Table 21b. Contingency table (summary of previous table)

r	20	40	60	80
.90	5	1	0	0	0
.70	8	5	6	2	1
.50	23	10	1	2	2

4.1.2.3. Reliability and validity

Table 22 gives a summary of reliability coefficients in each test. With the exception of the tests of the lowest level groups for 7th, 8th and 9th graders reliability is high. The reason for high reliability coefficients is that items are mostly open-ended and the area tested is relatively compact: equations and verbal applications of them.

Table 22. Reliability coefficients

Grade	(OSANA) Spearman-Brown (based on r_{bis})	(SPSS) Spearman-Brown (based on $r_{Pearson}$)	(SPSS) Guttman split-half	(SPSS) α_1 and α_2	(SPSS) Cronbach's alpha
1	.876	.831	.826	.770 .808	.87
2	.876	.807	.799	.773 .815	.87
3	.882	.852	.852	.786 .807	.88
4	.909	.885	.883	.829 .849	.91
5	.863	.770	.763	.744 .813	.86
6	.897	.881	.879	.804 .830	.90
7 lower	.588	.634	.573	.282 .470	not computed
7 upper	not computed	.707	.693	.656 .700	.78
8 lowest	.555	.418	.415	.584 .272	.55
8 middle	not computed	.615	.613	.481 .518	.64
8 highest	not computed	.749	.748	.663 .664	.78
9 lowest	.344	.340	.330	.314 .180	.33
9 middle	not computed	.757	.743	.772 .700	.83
9 highest	.819	.789	.773	.675 .730	.81

From the second last column of the table we can see that α ($x = 1,2$) tends to be bigger in the part of the test where verbal items are situated. This is true in all primary tests and all open-ended tests of 7th and 8th grades. Similarly α_x tends to be smaller in the case of multi-choice items. This occurs in the case of the lowest groups of the 8th and 9th grades.

In validity estimation the last marks in mathematics have been used. For the two lowest grades they are missing because of verbal evaluation; also some other pupils had not given their marks. There are 2092 pupils in the calculations of the Pearson correlation coefficient which is $.5946 \cong .59$ (the correlation between the total score and last recorded grades in mathematics).

4.1.3. Comparison of logistic and traditional views

In many respects the two approaches are near to each other. For example the separation index r_β is nearly the same as Cronbach´s alpha (Tables 15 and 22). An interesting comparison of item fit has been presented in Figure 21 and Table 23 where r_{bis} and T_2 are compared. The best fitting items in the sense of T_2-statistic are those which have the highest item-test correlation (data from the primary level). The phenomenon is very clear. The analogous result in the case of high T_2 values is not as obvious. Items which are close to misfitting tend to have low item-test correlation even if some items with relatively high r_{bis}-values are included.

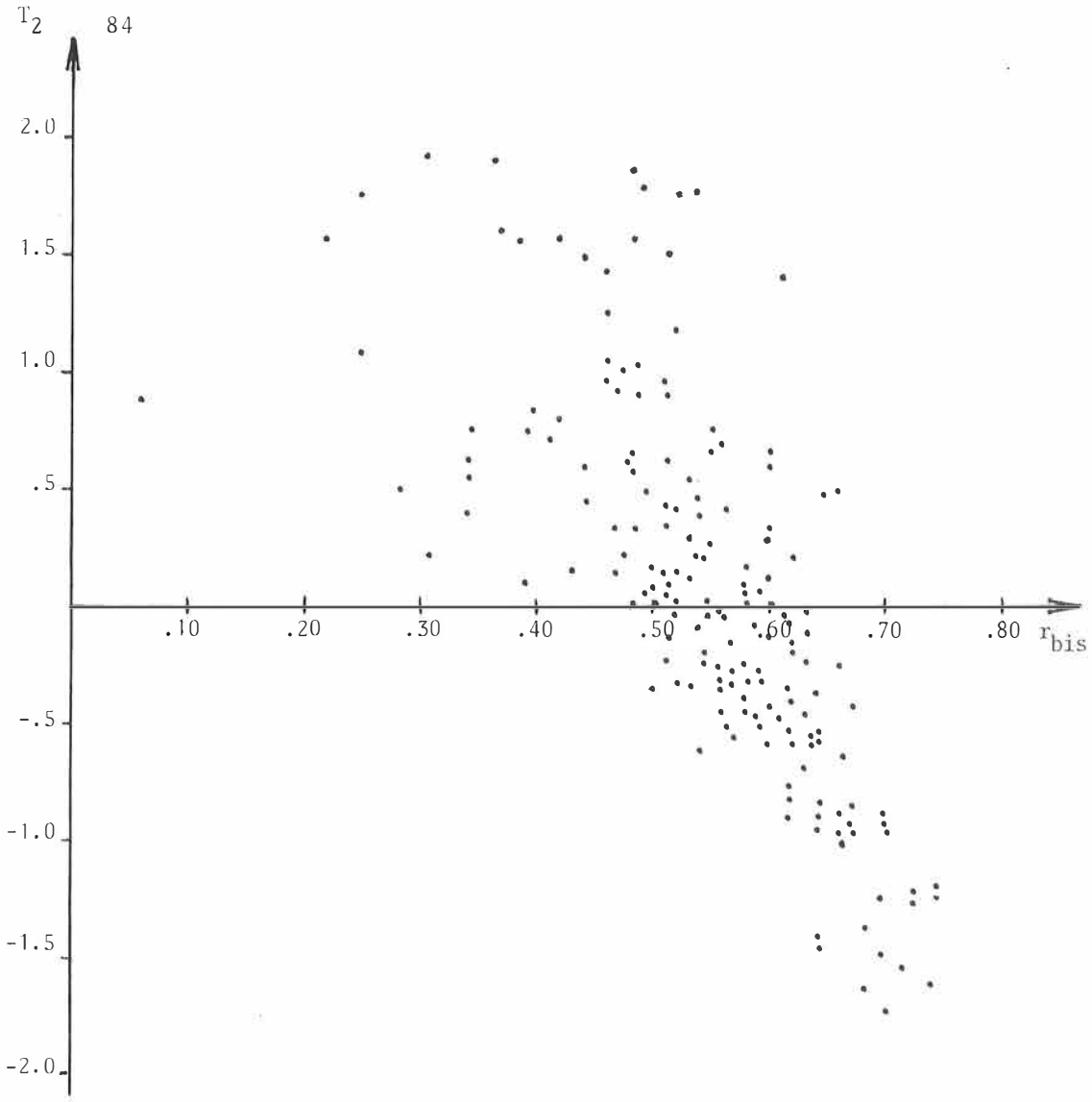


Figure 21. Relation between biserial item-test correlation and test of fit statistic T_2 in primary grades (fitting items)

Table 23. Contingency table based on Figure 21

T_2	.10	.20	.30	.40	.50	.60	.70	.80	.90
2									
1	-	-	3	4	10	4	1	-	
0	1	-	1	8	18	36	3	-	
-1	-	-	-	-	2	32	36	-	
-2	-	-	-	-	-	-	8	6	

r_{bis}

Table 23 shows a high negative correlation between r_{bis} and T_2. For finding good items, one criterion could be a high item-test correlation. For finding poor items, it is not necessarily very useful to reject items with a low r_{bis}-coefficient.

4.2. Item and test information

A theoretically interesting viewpoint for the comparison of the SLM and the LLTM is to examine more closely the information which those models can give about an item.

4.2.1. Definitions of information

Birnbaum´s (1968a) definition for information is in the case of the SLM for an item i

$$(4.25) \qquad I_i = \frac{(P'_\theta)^2}{P_\theta Q_\theta} \quad \theta = \beta_v - \delta_i \quad = \quad \frac{\exp(\beta_v - \delta_i)}{(1 + \exp(\beta_v - \delta_i))^2}$$

Figure 22 illustrates the information of item i as a function of the difference $\beta_v - \delta_i$. Maximum information has been achieved when a person´s ability is equal to the difficulty of the item.

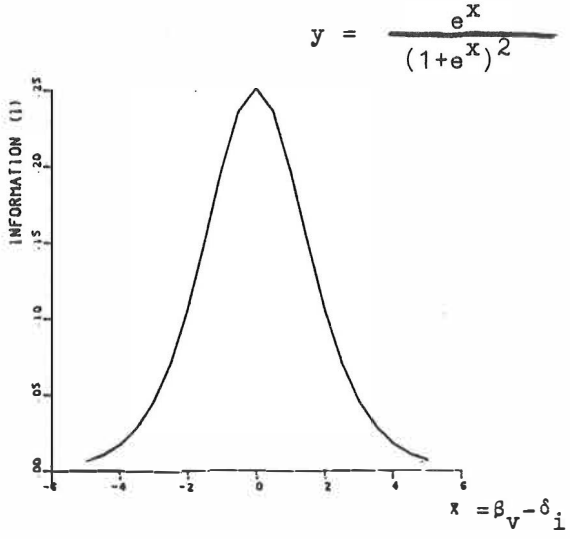


Figure 22. Birnbaum's information function for one item

In information theory information is defined by equation (4.26) (Young 1971):

$$I_i = \sum_x - p_x^{\,2} \log p_x = \sum_x p_x^{\,2} \log \frac{1}{p_x} \quad \text{bits or}$$

(4.26)

$$I_i = \sum_x p_x \ln \frac{1}{p_x} \quad \text{natural units}$$

in which p_x's are the probabilities of different answers. In the SLM (when $0 < p < 1$) the information of an item is derived from the information from correct and incorrect answers:

(4.27) $$I_i = p_1 \ln (1/p_1) + p_o \ln (1/p_o)$$

where

$$p_1 = \frac{1}{1 + \exp(\delta_i - \beta_v)}$$

$$p_o = \frac{1}{1 + \exp(\beta_v - \delta_i)}$$

For each item

$$I_i = \frac{\ln(1 + e^{-\theta})}{1 + e^{-\theta}} + \frac{\ln(1 + e^{\theta})}{1 + e^{\theta}} \qquad (4.28)$$

where $\theta = \beta_v - \delta_i$

Maximum information in this case (when $\beta_v = \delta_i$) is $\ln 2 = .69$.

In the case of the LLTM, $\delta_i = \sum_{j=1}^{m} q_{ij} \eta_j$ and the information is the same as in the SLM. The divergence from the SLM becomes evident if we think that each of m operations would be solved separately. In this case

$$p_1 \ln(1/p_1) = \left(\frac{e^{\beta-\eta_1}}{1+e^{\beta-\eta_1}}\right)^{q_1} \cdots \left(\frac{e^{\beta-\eta_m}}{1+e^{\beta-\eta_m}}\right)^{q_m} \cdot \ln \left(\left(\frac{1+e^{\beta-\eta_1}}{e^{\beta-\eta_1}}\right)^{q_1} \cdots \left(\frac{1+e^{\beta-\eta_m}}{e^{\beta-\eta_m}}\right)^{q_m}\right)$$

$$= \left(\prod_j (1+e^{\eta_j-\beta})^{-q_j}\right) \left(\ln \prod_j (1+e^{\eta_j-\beta})^{q_j}\right)$$

$$= \frac{1}{x} \ln x, \text{ where } x = \prod_j (1 + e^{\eta_j-\beta})^{q_j}$$

Similarly

$$p_o \ln(1/p_o) = (1-\tfrac{1}{x}) \cdot \ln \left(\frac{1}{1-\frac{1}{x}}\right) = \frac{x-1}{x} \ln \frac{x}{x-1} .$$

For each item

$$(4.29) \qquad I_i = \frac{1}{x} \ln x + \frac{x-1}{x} \ln \frac{x}{x-1}$$

$$\text{where } x = \prod_j \left(1+e^{\eta_j - \beta}\right)^{q_j} > 1$$

For both models (4.28) and (4.29) the maximum information is the same. In Figure 23 both information functions have been considered to be sums of the two components.

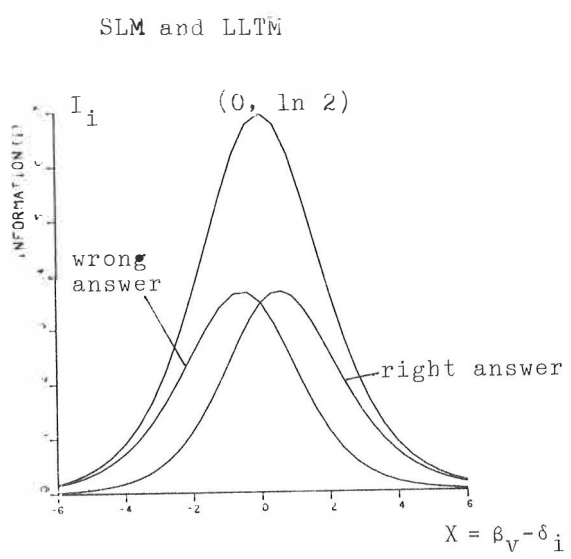


Figure 23. Information function based on information theoretical definitions

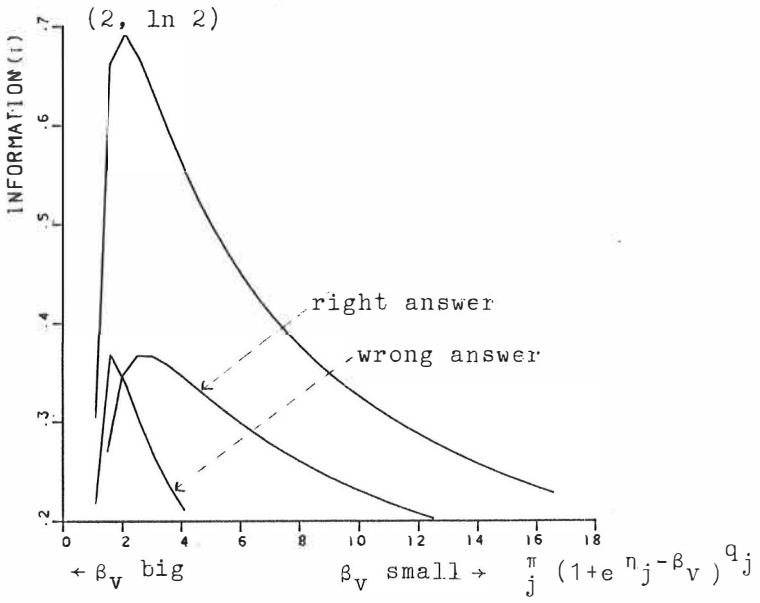


Figure 24. Information, if each operation in an item is a separate task (Suppes model)

If measuring could be performed at the level of operations, the amount of information would not decrease as fast as it does in the case of the whole item. High values of

$$\pi_j (1 + e^{\eta_j-\beta_v})^{q_j}$$

mean that the item is a combination of quite difficult operations (compared to the ability of a person solving it). The information curve shows that, even in this case, a lot of

information can be received from this item, particularly in the case of the right answer.

A third definition of information is named Fisher information. It is defined by means of log-likelihood function $\mathcal{L}$ (Fisher 1974, p. 293)

$$I_i = E\left(\left(\frac{\partial \mathcal{L}}{\partial \beta}\right)^2\right) = -E\left(\frac{\partial^2 \mathcal{L}}{\partial \beta^2}\right) . \tag{4.30}$$

In the case of the SLM:

For a correct answer we have $\mathcal{L} = \ln P_i(\beta_v) = \ln P_{vi}$
and for a wrong answer $\mathcal{L} = \ln (1-P_{vi})$ which implies

$$I_i = \left(\frac{\partial \ln P}{\partial \beta}\right)^2 \cdot P + \left(\frac{\partial \ln (1-P)}{\partial \beta}\right)^2 (1-P) = P(1-P) \tag{4.31}$$

This is the same as Birnbaum's information mentioned in equation (4.25).

Similarly

$$I_i = -\left(\frac{\partial^2 \ln P}{\partial \beta^2}\right) \cdot P - \left(\frac{\partial^2 \ln (1-P)}{\partial \beta^2}\right)(1-P) = P(1-P) \tag{4.32}$$

The information of the whole test is the sum of the information of items. Information can be defined not only for items but also analogously for persons.

$$I_{test}(\beta_v) = \sum_{i=1}^{k} I_i = \sum_{i=1}^{k} P_i(\beta_v)\left(1 - P_i(\beta_v)\right) \tag{4.33}$$

$$I_{persons}(\delta_i) = \sum_{v=1}^{n} I_v = \sum_{v=1}^{n} P_i(\beta_v)\left(1 - P_i(\beta_v)\right) .$$

Also in the Fisher information function, the information of an item consists of two parts: information given by right and wrong answers.

$$(4.34) \qquad I_i = I_{correct} + I_{wrong} = \frac{e^{\beta-\delta}}{1+e^{\beta-\delta}} + \frac{e^{\beta-\delta}}{1+e^{\beta-\delta}}$$

Both types of answers give 50 % of the amount of information.

4.2.2. Information functions form the data

In the literature, information has been defined without exception according to Fisher´s (= Birnbaum´s) function (e.g. Hambleton & Traub 1971; Samejima 1977; Lord 1977; Wright & Stone 1979). It is not paractical to use the expression

$$I_{test} = \sum \frac{e^{\beta v-\delta i}}{(1+e^{\beta v-\delta i})^2}$$

for calculating information of the whole test. It is easier to use the equation of standard error (3.11) from which it is obvious that

$$(4.35) \qquad I_{test} = \sum_{i=1}^{k} P_{vi}\ (1-P_{vi}) = \left(\frac{1}{SE(\hat{\beta}_v)}\right)^2$$

Figure 25 gives the information functions of 9th graders. For both of them the maximum value of the information has been dervied when $\beta_v \cong .5$. It is not the case in the figure which presents information functions for each level group of 8th graders.

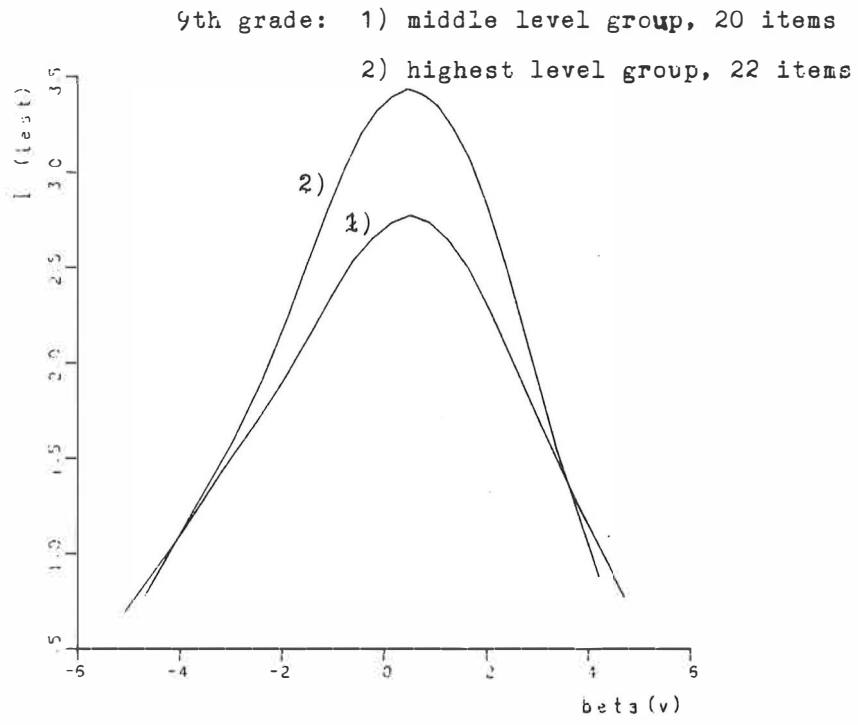


Figure 25. Information of the test for 9th graders

To take the primary grades as an example, we present the information curve of the first graders (Figure 26). Maximum information is bigger than in Figure 25. One reason for this is that there are 28 items in the test of first graders; also some standard errors for $\hat{\beta}_v$ are smaller in the test of first than ninth graders.

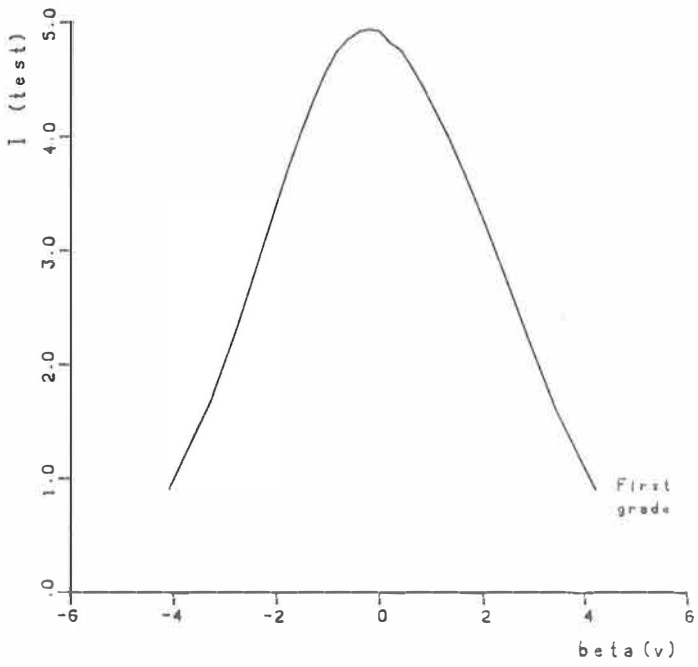


Figure 26. Information of the test for 1st graders

Also other information functions for primary grades are very similar to those of first graders. The best reliability coefficient exists in the case of test 4. The information function for this test has been drawn in Figure 27, and it has been compared to the test for 6th graders. For 4th graders I_{max} is higher than in any other primary tests. This means that the test gives more information from the average ability group ($\beta_V \cong 0$) than any other test. The mean information from each item, when $\beta_V = 0$ is about .19 (the theoretical maximum is .25). Even if we had the best possible items, with $I_i = .25$ for each of them, we would need 23 items for getting the same amount of information from the test as has been obtained from test 4.

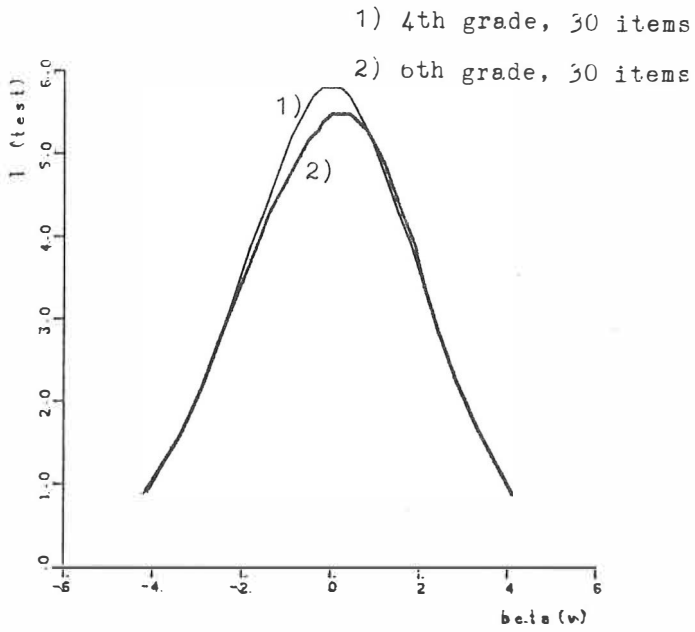


Figure 27. Information functions for test 4 and test 6.

For 8th grade information curves can be compared easily in the case of the lowest and middle level groups because of a similar number of items. Items are partly the same: 12 common items at the beginning of each test, the remaining 9 items being multiple-choice type for the lowest and open-ended for the middle level group. With the exception of 9 items both tests are the same. The test for the middle level group gives more information of extreme persons ($\beta_v < -2.5$ or $\beta_v > 2.0$).

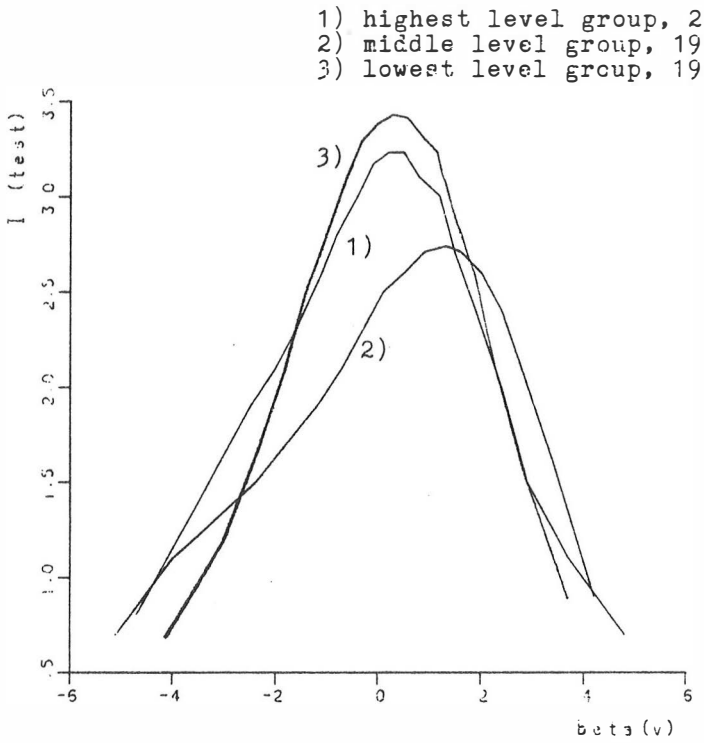


Figure 28. Information functions for each level group in 8th grade

4.2.3. Reliability and information

In traditional test theory the reliability of a test has been calculated using different coefficients (paragraph 4.1.2.3). These coefficients are, however, not independent of the group of persons. Samejima (1977) and Lumsden (1976) make this point in their criticism of traditional test theory.

> "Thus it is clear that the reliability coefficient in classical test theory is at the mercy of the heterogeneity of the group of examinees, which has nothing to do with the test itself. We can easily make an erroneous test look good by using a heterogeneous group of subjects and obtaining a large value for the "reliability coefficient". Similarly we can make a good test look bad by using a homegeneous group of subjects.
> ... A fatal deficiency of classical test theory is that it cannot specify the standard error of estimation independently from the reliability coefficient, and thus independently from a specific group of examinees." (Samejima 1977, p. 196)

The paragraph is based on the fact that in traditional test theory reliability is by definition the correlation between two sets of parallel measurements, which implies that

$$r_{xx'} = 1 - \frac{S_e^2}{S_x^2} \tag{4.36}$$

(reliability = 1 - error variance / variance of observed scores), from which we get the standard error of measurement

$$S_e = S_x \sqrt{1 - r_{xx'}} \,. \tag{4.37}$$

The corresponding standard error in latent trait theory is

$$SE(\hat{\delta}_i) = \frac{1}{\sqrt{\sum_{v=1}^{n} P_{vi}\,(1-P_{vi})}} \tag{4.38}$$

$$SE(\hat{\beta}_v) = \frac{1}{\sqrt{\sum_{i=1}^{k} P_{vi}\,(1-P_{vi})}}$$

In equation (4.37) both $r_{xx'}$ and S_x are derived from a specific group of examinees. They are not population-free and that is why reliability is dependent on the population on which the test has been administered. Standard errors of $\hat{\delta}_i$ and $\hat{\beta}_v$ are dependent on the number of persons and items

respectively: otherwise they are population-free. They are derived from the 2nd derivative of the log-likelihood function using maximum likelihood estimation:

$$P_{vi} = \frac{e^{x_{vi}(\beta_v - \delta_i)}}{1 + e^{\beta_v - \delta_i}}$$

$$L = \prod_{i=1}^{k} \prod_{i=1}^{n} P_{vi} = \frac{e^{\sum_i \sum_v x_{vi}(\beta_v - \delta_i)}}{\prod_i \prod_v (1 + e^{\beta_v - \delta_i})}$$

$$\ln L = \mathcal{L} = \sum_i \sum_v x_{vi} (\beta_v - \delta_i) - \ln \prod_i \prod_v (1 + e^{\beta_v - \delta_i})$$

$$= \sum_v \beta_v (\sum_i x_{vi}) - \sum_i \delta_i \sum_v x_{vi} - \sum_i \sum_v \ln (1 + e^{\beta_v - \delta_i})$$

$$(4.39) \quad \mathcal{L} = \sum_v \beta_v r_v - \sum_i \delta_i S_i - \sum_i \sum_v \ln (1+e^{\beta_v - \delta_i})$$

Where r_v is the sufficient statistic for β_v and S_i for δ_i. From equation (4.39) we can easily derive equations for derivates:

$$(4.40) \quad \frac{\partial \mathcal{L}}{\partial \delta_i} = - S_i + \sum_v \frac{e^{\beta_v - \delta_i}}{1 + e^{\beta_v - \delta_i}} = - S_i + \sum_v P_{vi} \quad (i=1,\ldots,k)$$

$$(4.41) \quad \frac{\partial^2 \mathcal{L}}{\partial \delta_i^2} = - \sum_v P_{vi} (1-P_{vi}) = - \frac{1}{(SE (\hat{\delta}_i))^2}$$

Information which one item can give of persons is

$$(4.42) \quad I_i = - E \left(\frac{\partial^2 \mathcal{L}}{\partial \beta^2}\right) = P_{vi} (1-P_{vi}) = E \left(\left(\frac{\partial \mathcal{L}}{\partial \beta}\right)^2\right) .$$

This relation shows that the amount of information, such as the likelihood function, is additive. Among consistent estimates, that which conserves the greatest amount of information is the estimate of maximum likelihood (Fisher 1959).

Lumsden´s criticism is based on assumptions of traditional model (model T): O = T + E (observed score = true score + error) (Lumsden 1976, 1977, 1978). He relies on information function:

> "A different and promising approach to the problems of reliability (and some others as well) is given by the information measure developed by Birnbaum. This measure I (θ, x) is considered by Birnbaum to be a kind of index of precision which for a given test and scoring formula reflects the information provided by the test in the vicinity of a given value of the attribute θ. It should be noted that this measure is not based on any of the model T assumptions. This is why it is described as promising." (Lumsden 1976, p. 262)

In latent trait definition of weakly parallel tests Samejima (1977) gives an interesting way of comparing tests in our data in the sense of information.

> "By weakly parallel tests we mean any pair of tests measuring the same ability or latent trait whose test information functions are indetical. Note that in this definition of weakly parallel tests nothing is required for the number of test items, the number of item score categories of each item or the operating characteristics of the score categories." (Samejima 1977, p. 194)

One important advantage of this definition compared to weakly parallel tests in a traditional sense (Konttinen 1979a) is that in constructing parallel test we can add or discard an item in either of the two tests to make the two test information functions practically equal. All information functions of primary tests (with fitting items) are presented in Figure 29. Units on the x-axis are selected so that each curve is separate. Curves are based on Table 24 and equation (4.35). Tests 1, 2 and 3 can be considered to be weakly

parallel. They would become even more parallel if one item were added to test 1 (and perhaps also to test 2). Tests 4 and 5 are weakly parallel, if one good item were added to test 6, this test would also become parallel to tests 4 and 5.

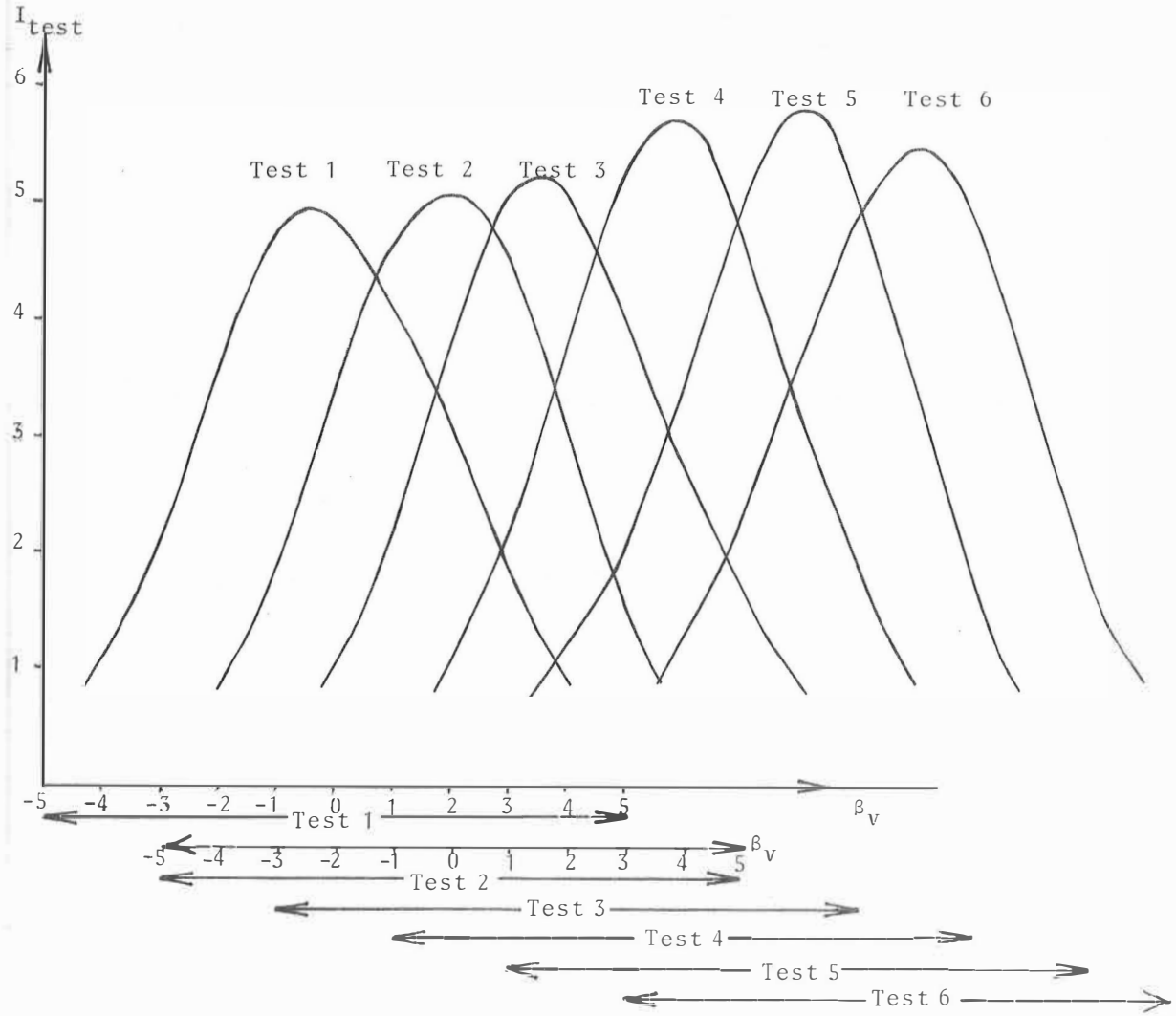


Figure 29. Information functions of primary tests

Table 24. Estimates of person abilities and their standard errors in primary grades

Total score	Test 1 (k=28) $\hat{\beta}_v$	$SE(\hat{\beta}_v)$	Test 2 (k=27) $\hat{\beta}_v$	$SE(\hat{\beta}_v)$	Test 3 (k=28) $\hat{\beta}_v$	$SE(\hat{\beta}_v)$	Test 4 (k=30) $\hat{\beta}_v$	$SE(\hat{\beta}_v)$	Test 5 (k=30) $\hat{\beta}_v$	$SE(\hat{\beta}_v)$	Test 6 (k=30) $\hat{\beta}_v$	$SE(\hat{\beta}_v)$
1	-4.09	1.049	-3.86	1.041	-3.93	1.036	-4.05	1.052	-4.40	1.133	-4.21	1.054
2	-3.30	.774	-3.08	.766	-3.16	.758	-3.25	.776	-3.46	.846	-3.41	.782
3	-2.79	.658	-2.58	.651	-2.68	.641	-2.75	.657	-2.87	.709	-2.89	.667
4	-2.41	.592	-2.20	.585	-2.31	.574	-2.36	.587	-2.42	.626	-2.49	.601
5	-2.08	.549	-1.89	.543	-2.01	.531	-2.05	.541	-2.07	.571	-2.16	.557
6	-1.80	.518	-1.61	.514	-1.74	.501	-1.77	.508	-1.76	.531	-1.87	.526
7	-1.54	.496	-1.36	.493	-1.50	.479	-1.53	.483	-1.50	.501	-1.60	.502
8	-1.30	.480	-1.12	.477	-1.28	.464	-1.30	.464	-1.26	.479	-1.36	.483
9	-1.08	.468	-.90	.466	-1.07	.452	-1.10	.450	-1.04	.461	-1.14	.468
10	-.87	.459	-.69	.457	-.87	.445	-.90	.439	-.83	.447	-.92	.456
11	-.66	.454	-.48	.450	-.68	.439	-.71	.430	-.64	.437	-.72	.447
12	-.45	.451	-.28	.446	-.48	.437	-.53	.424	-.45	.428	-.52	.439
13	-.25	.450	-.08	.444	-.29	.436	-.35	.420	-.27	.422	-.33	.434
14	-.05	.451	.11	.443	-.10	.437	-.17	.417	-.09	.418	-.15	.430
15	.16	.454	.31	.444	.09	.441	.00	.417	.08	.416	.04	.427
16	.37	.479	.51	.447	.29	.446	.17	.417	.25	.415	.22	.426
17	.58	.466	.71	.452	.49	.454	.35	.420	.43	.416	.40	.428
18	.80	.475	.92	.460	.70	.465	.53	.424	.60	.419	.59	.430
19	1.03	.487	1.14	.471	.92	.479	.71	.430	.78	.424	.77	.435
20	1.28	.500	1.36	.486	1.16	.497	.90	.439	.96	.431	.97	.443
21	1.53	.517	1.61	.506	1.42	.520	1.09	.450	1.15	.441	1.17	.453
22	1.81	.539	1.88	.535	1.70	.550	1.30	.464	1.35	.455	1.38	.466
23	2.12	.568	2.19	.577	2.03	.588	1.53	.483	1.56	.472	1.60	.485
24	2.46	.609	2.56	.643	2.40	.640	1.77	.508	1.80	.496	1.85	.509
25	2.87	.673	3.04	.760	2.86	.713	2.05	.541	2.06	.529	2.12	.541
26	3.40	.785	3.81	1.037	3.45	.831	2.36	.588	2.36	.576	2.44	.587
27	4.20	1.055			4.34	1.095	2.75	.658	2.73	.646	2.82	.656
28							3.25	.777	3.23	.767	3.33	.775
29							4.06	1.053	4.01	1.046	4.12	1.051

Table 25 and Figure 30 give information functions for 7th graders. For the test of the 7th lower level group 12 is the highest observed score (the corresponding $\hat{\beta}_V$ = .66.). In spite of missing total scores 13...18 in the data, the information curve can be drawn in the usual way by means of SE ($\hat{\beta}_V$).

1) lower level group
2) upper level group

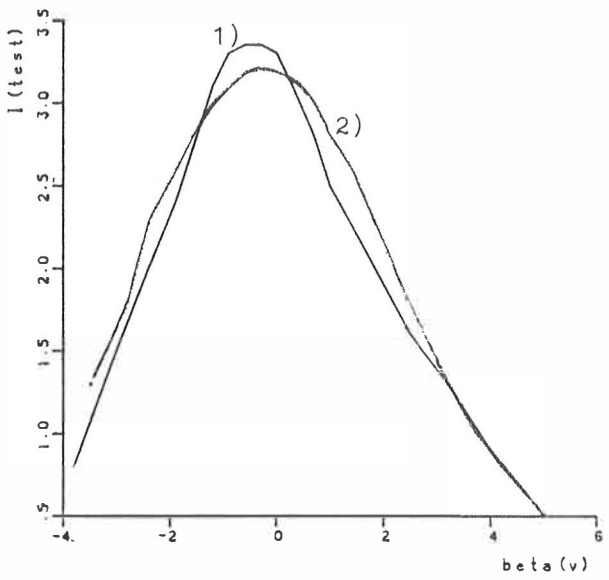


Figure 30. Information curves for tests of 7th grades

Table 25. Ability estimates and their standard errors in the tests of the 7th grade

total score r	lower level group (k=19)		upper level group (k=21)	
	$\hat{\beta}_v$	$SE(\hat{\beta}_v)$	$\hat{\beta}$	$SE(\hat{\beta}_v)$
1	-3.84	1.101	-4.44	1.141
2	-2.95	.828	-3.47	.862
3	-2.37	.708	-2.84	.736
4	-1.92	.640	-2.36	.666
5	-1.54	.596	-1.94	.623
6	-1.20	.569	-1.57	.596
7	-.89	.553	-1.23	.578
8	-.59	.546	-.90	.567
9	-.29	.546	-.59	.560
10	.01	.554	-.27	.558
11	.33	.569	.04	.560
12	.66	.594	.36	.566
13	1.04	.629	.68	.577
14	1.46	.673	1.02	.595
15	1.95	.726	1.39	.621
16	2.52	.791	1.80	.661
17	3.22	.889	2.28	.720
18	4.20	1.130	2.86	.817
			3.67	.996
			5.01	1.352

4.2.4. Summary of the data in the NEWRATE program

The data in this research is comparatively large including 13 tests from 9 different grades. One way of summarizing the results is to look at the means and standard deviations of the distributions of item fit and person fit statistics and residuals (Table 26 and 27).

Table 26. Item fit statistics in all grades

Grade/level group	sd($\hat{\delta}$)	mean of T_1	sd(T_1)	mean of T_2	sd(T_2)	df
1	1.470	.197	1.0034	.079	.8727	192.8
2	1.202	.127	.9651	.023	.8847	199.3
3	1.412	.008	.7810	-.077	.7788	197.6
4	1.269	.216	.8379	.116	.7596	140.1
5	1.371	.295	1.0708	.145	.8863	227.1
6	1.397	.156	.8680	.067	.8140	146.5
7/lower	1.742	.387	.8878	.255	.7518	102.3
7/upper	2.175	.242	.8242	.138	.7366	186.6
7/all	2.108	.415	.9074	.304	.8878	290.4
8/lowest	1.619	.170	.7322	.011	.7263	91.8
8/middle	2.354	.236	1.2465	.064	1.0089	162.9
8/highest	2.206	.095	.7144	.022	.6746	159.0
8/all	1.809	.670	1.5169	.348	1.3727	412.1
9/middle	2.418	.396	1.1182	.139	.7542	178.6
9/highest	1.959	.187	.9308	.047	.7527	208.0
9/all	2.015	.553	1.4750	.230	1.1422	432.2

Mean of distribution of T_2-statistic is nearer to 0.0 than that of T_1 and its standard deviation is nearer to 1.0 than that of T_1.

Table 27 gives the corresponding information for persons. All means of T_2-statistic are slightly negative and its standard deviation a bit smaller than one in each test. The biggest difference in standard deviations is in the test of the upper level group of 7th graders: for them sd(T_1) $\cong$ 2.13 and sd(T_2) $\cong$ .47. The reason for high sd(T_1) is one misfitting person with $\hat{T}_1 \cong 28.3$, after ln-transformation $\hat{T}_2 \cong 2.3$. For the next misfitting person $\hat{T}_1 \cong 7.2$ and the corresponding $\hat{T}_2 \cong 2.0$. This is an example of transforming high T_1-values too much in transformation.

Table 27. Person fit statistics in all grades

Grade/level group	mean of $\hat{\beta}_v$	sd($\hat{\beta}_v$)	mean of T_1	sd(T_1)	mean of T_2	sd(T_2)	df
1	.503	1.5041	.0978	1.2689	-.1644	.9244	26.9
2	.535	1.3991	.1167	1.0334	-.0809	.9428	25.9
3	.262	1.4896	.1197	.9999	-.0791	.8039	26.9
4	.164	1.5658	.1591	1.1876	-.0502	.8607	28.8
5	.257	1.2873	.1948	2.4322	-.0939	.7235	28.9
6	.144	1.5080	.0473	.9024	-.1145	.7651	28.9
7/lower	-1.305	.9573	.0777	.8090	-.0864	.4923	17.8
7/upper	-.315	1.4078	.1923	2.1344	-.0918	.4688	19.9
7/all	-.816	1.3946	.1045	1.7440	-.0929	.4027	19.9
8/lowest	-1.789	1.0790	-.0230	.6771	-.2396	.6365	17.8
8/middle	-1.696	1.3540	-.0106	.9487	-.3492	.6734	17.9
8/highest	.359	1.3894	.1237	.9204	-.1089	.5706	19.9
8/all	-.698	1.6836	.1844	1.5093	-.2355	.8540	18.0
9/middle	-1.043	1.7661	.1661	1.4450	-.1891	.5395	18.9
9/highest	.782	1.4350	.1679	1.1793	-.1227	.6775	20.9
9/all	-.051	1.8832	.3133	2.0838	-.1575	.7369	19.0

4.3 Linking of tests

4.3.1. Aims and methods in linking

The main idea in the linking of tests in (Rasch 1960, p. 29): "When the degrees of difficulty of all tests in a chain have been determined, the ability of a pupil may be estimated from any test of set of tests which is suitable for him." Our aim in calibrating items is to build up an item bank in which items measure the same latent trait. The difficulties of items are relative to the particular set of items used in the test. That is why items cannot be merely

thrown into the same bank together. Linking of the items is required. Linking is based on common items used in the tests a and b in consecutive grades. A summary of the main elements of this method proposed by Wright (1977) and Wright and Stone (1979) follows (Morgan 1980):

1. Begin by separately calibrating the items in test a and test b, which give two independent sets of estimated item difficulties for the linked items. Let $\hat{\delta}_{ai}$ and $\hat{\delta}_{bi}$ represent the estimated item difficulties of the i th item in the link, in test a and test b respectively.
2. Calculate the translation constant which effectively translates all item difficulty estimates from the calibration of test b to the calibration scale of test a, using the formula

$$\hat{t}_{ab} = \frac{\sum_{i=1}^{K} (\hat{\delta}_{ai} - \hat{\delta}_{bi})}{K} \qquad (4.43)$$

   where K is the number of items in the link. This translation constant is the difference in average estimated item difficulties of the common items in the two calibrations. The standard error of the estimated translation constant $SE(\hat{t}_{ab})$ is approximately 3.5/NK, where N is the calibration sample size of the linked items.
3. The validity of the link between test a and test b may be tested using the statistic

$$\frac{N}{12} \cdot \frac{K}{K-1} \cdot \sum_{i=1}^{K} (\hat{\delta}_{ai} - \hat{\delta}_{bi} - \hat{t}_{ab})^2, \qquad (4.44)$$

   which is distributed approximately as a chi-square with K degrees of freedom.
4. The validity of an item in the link may be tested using the statistic

$$\frac{N}{12} \cdot \frac{K}{K-1} \cdot (\hat{\delta}_{ai} - \hat{\delta}_{bi} - \hat{t}_{ab})^2 \qquad (4.45)$$

which is distributed approximately as a chi-square with one degree of freedom.

5. The validity of the link and the items in the link may also be ascertained visually by plotting the estimated difficulty estimates of the common items from the two calibrations, and observing the amount of scatter in the points.

4.3.2. Linking of consecutive tests in primary level

In two consecutive tests there are about 5 common items. At first the validity of each link has been tested. The common items misfitting in the link (criteria 3 and 4 in previous section, α = 5 %) are rejected. It appears that two items in the first/second test link and two items in the link of the 5th and 6th graders had to be rejected. Table 28 gives a summary of the linking process in primary grades.

An alternative way of studying the validity of linking is to use the standardized residuals (Morgan 1980)

$$z_{ab} = \frac{\hat{\delta}_{ai} - \hat{\delta}_{bi} - \hat{t}_{ab}}{\sqrt{SE(\hat{\delta}_{ai})^2 + SE(\hat{\delta}_{bi})^2}} \qquad (4.46)$$

to see if they estimate the expected mean to be equal to zero and the expected standard deviation equal to one.

Because of the small number of items in each link some deviations from expected mean and standard deviation can be detected. Some supplementary common items would make the estimates closer to their expected values.

If we know approximately the ability level of a pupil, we can choose for him a particular set of items for testing his ability. Only a few items are enough if items conform to the simple logistic model. These kinds of item pools could be recommended for use in testing achievements in basic

Table 28. Items used in linking in primary grades

Tests a & b	Linking items test a	test b	Smaller calibration sample size	$\hat{t}_{ab}$	SE($\hat{t}_{ab}$)
1 & 2	15	8	201	1.32	.006
	27	23			
	30	28			
2 & 3	11	1	206	1.41	.003
	13	2			
	18	3			
	21	4			
	19	6			
	30	29			
3 & 4	6	2	146	.58	.005
	8	3			
	14	7			
	29	25			
	28	26			
4 & 5	2	2	146	.34	.005
	5	9			
	22	21			
	25	25			
	28	26			
5 & 6	9	11	236	.74	.005
	8	12			
	21	13			

Figure 31 illustrates the validity of linking graphically (according to criterion 5).

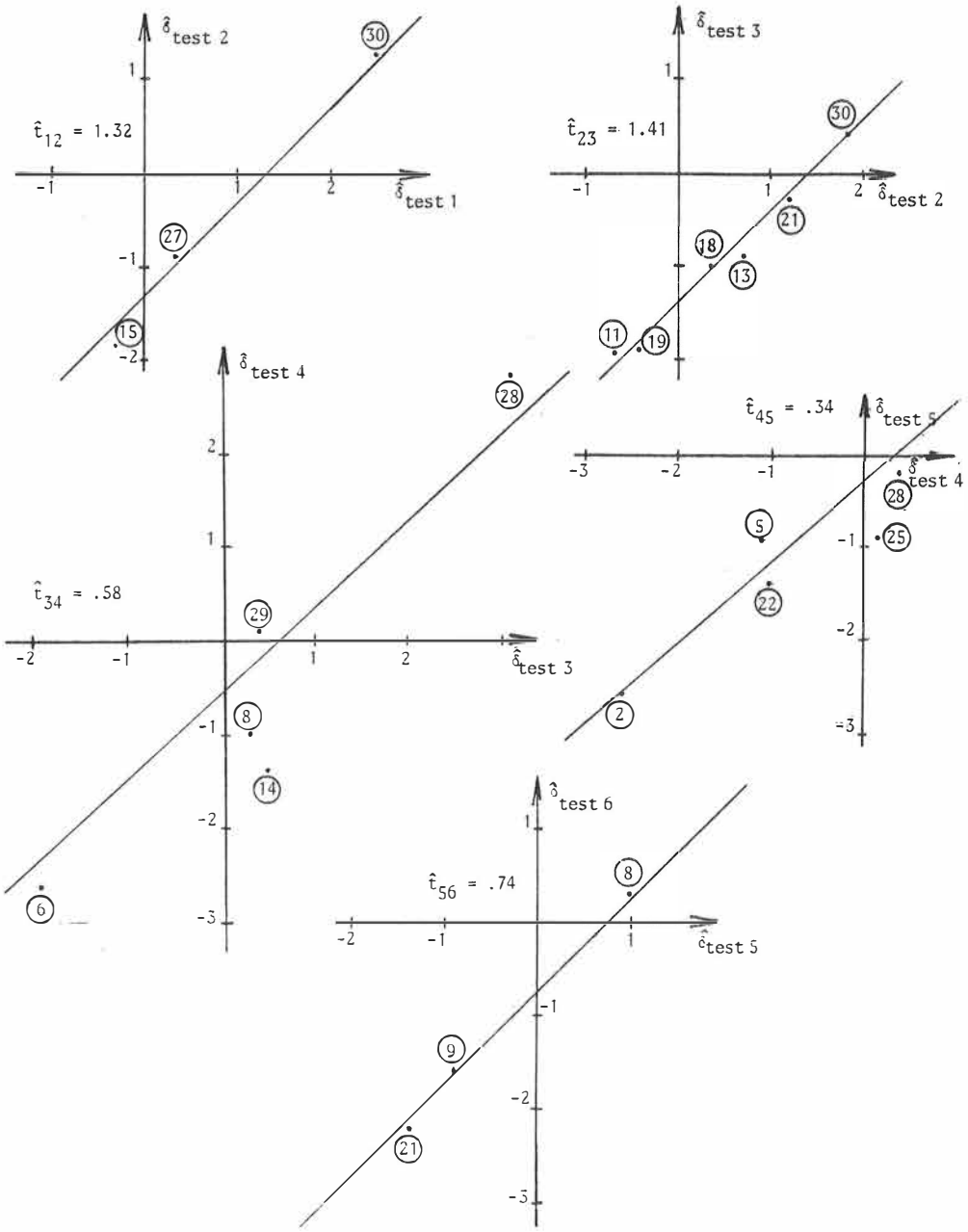


Figure 31. Validity of linking (criterium 5)

Table 29. Validity of linking based on standardized residuals

Tests	Number of items in linking	$\bar{z}_{ab}$	$SD(z_{ab})$
1 & 2	3	-.03	.89
2 & 3	6	.01	.64
3 & 4	5	.02	.82
4 & 5	5	.12	1.84
5 & 6	3	-.03	.46
	Mean	.02	.93

skills in Mathematics. At the end of second, fourth and sixth grade some common tests in the main areas of school Mathematics could be constructed so that two tests could be linked together as presented before. In order to avoid tests that are too long the items could be arranged in order of increasing difficulty and pupils could determine their own starting point independently.

5. LINEAR LOGISTIC TEST MODEL IN THE MEASUREMENT OF ACHIEVEMENTS IN PRIMARY SCHOOL MATHEMATICS

The purpose of this chapter is to continue the analysis of learning difficulties. At the beginning, the viewpoint of structural learning is examined to find the basis for operational structure. The methodological tool in this chapter will be the linear logistic test model (LLTM).

The use of the LLTM will be introduced studying the pupils attainments in the data of equations and inequations in the primary school, and the learning difficulties will be found using the LLTM and its basic parameters. The linking methods will be developed for achievement tests using the LLTM. The relation of the difficulty of the whole task and the difficulties of the basic operations will be studied theoretically.

5.1. Logistic models in Mathematics education

The first phase in the history of mathematical models in psychology and education started about 150 years ago with attempts to find analogies between psychological theories and Newton's mechanics (Kempf 1977a). The second phase of development began about 35 years ago. It is connected with the names Fischer, Guttman, Lazarsfeld and Rasch among

others. This stage can be separated from the first because more attention has been paid to the philosophy of science. Concepts such as specific objectivity, unidimensionality, goodness of fit and many other concepts which are related to assumptions of mathematical models have become essential.

At the same time as probabilistic models and group theoretical structure have turned out to be useful in the interpretation of the properties of materials in physics and chemistry, psychologists and educators interested in building fascinating models have seen probabilistic models as being of considerable value in their studies of human behaviour. Logistic models have proved to be an excellent tool for finding some structure behind a person´s answers in a text or questionnaire or for finding deviations from ratings performed by many different people. Logistic models are basically probabilistic: some of them, particularly the linear logistic model has a deterministic component and for this reason it can be seen as a bridge between probabilistic and deterministic models and, in addition, it gives some extra information from the structural point of view from the data.

The main principles in selecting or constructing mathematical models are listed briefly on page 42. This chapter concentrates on the properties of the linear logistic test model. Each of those five principles has been taken into detailed examination in the context of this model. Not only the theoretical characteristics of the model but also the usability of it in the development of the Mathematics syllabus has been highlighted.

The data consists of six Mathematics tests for the lower level of the Finnish comprehensive schools. Each test contains 30 items. Each item has been decomposed to basic operations (or rules) which it is essential to know to get the correct answer. The difficulty estimates of basic operations give a great deal of information about sources of learning difficulties in each grade. They can also be used when comparing learning effects in consecutive grades. This

kind of information is important for developing the syllabus in Mathematics and in particular the part of the syllabus which is supposed to from the field of basic skills and basic objectives. If logistic test models can be used for improving the syllabus, a big step has been made in finding empirical uses for statistical theories. A lot of work has been done explaining human learning in terms of the structure behind the learning tasks or instruction (e.g. Kempf 1977a, 1977b, Dienes 1979, Scandura 1979). The structural point of view has been used not only in Mathematics but also for example in linguistics and the physical sciences as well. Latent trait theory is a tool which can be used when the interpretation of achievements is based on the structure of task used in testing achievements.

5.2. Structural learning

Dienes (1979) defines learning as being a personal adaption to the environment at the end of which the learner is able to cope with something more effectively from the point of view of his own goals than he was before. Essential parts of learning are:

(i) interaction with the learning environment,
(ii) discovery and manipulation of the rules apparent in the environment,
(iii) comparison of one rule structure with another,
(iv) representation of rule structures,
(v) development of a symbolic language for describing the rule structures and
(vi) a formalization stage which means finding some order in the symbolized way of describing rules.

Structural learning is concerned with the following questions (Scandura 1979):

- Firstly, what does it mean to know something, and how can one respect competence so that it has behavioural relevance?
- Secondly, how does one find out what people know about some content domain? How does one measure knowledge?
- Thirdly, what are the mechanisms by which knowledge is put to use (and this includes the whole gamut, ranging from perception and memory to problem solving)?
- Fourthly, and finally, what are the relationships among these conceps and their implications for instruction and communication.

In one sense then, structural learning is an area of interdisciplinary pursuits. It fills gaps between such fields as artificial intellingence, instructional design, cognitive psychology, educational measurement, philosophy, mathematics, education, etc. In other sense, structural learning is multi-disciplinary, since many different disciplines can contribute to the solution of its problems. The field is also cross-disciplinary or trans-disciplinary, since it is conserned with ideas common to variety of disciplines. The use of rules and rule-based thinking, for example, pervades the various disciplines listed above.

In structural learning models many theoretical issues exist concerning the structure of memory, the hierarchy of different rules and individual use of the rules. Spada (1977) points out differences between basic types of models. Scandura´s deterministic model assumes rules to be applied in a deterministic fashion: a student who has learned a certain rule will apply it correctly to all tasks that require this rule. Scandura´s memory-free theory of structural learning is deterministic in the sense that it is possible to predict an individual´s actual behaviour with respect to a given class of tasks, from the set of rules he or she knows (Häussler 1978). Suppes´ probabilistic model assumes that all subjects use the same operations with equal error rates when solving a given problem. In the logistic variant of Suppes´ probabilistic model operation,

probabilities are allowed to differ from student to student and the probability of getting the whole item correct is considered to be the product of the probabilities of getting each operation correct separately.

In the stochastic theory of structural learning, in contrast to Scandura´s theory, three basically psychological assumptions have been made (Kempf 1975):

(i) It will be assumed that an individual is more or less capable of performing a rule.

(ii) It will be assumed that an individual does learn something from the performance of a task, that is, learning is dependent on the structure of the task and that is generalized to other tasks as well.

(iii) It will not be assumed that an individual will solve any item which is composed of rules that are known by the individual, regardless of the number and difficulty of the rules to be applied.

Two of these assumptions (i and iii) will be studied more closely in this chapter.

One way to deal with the problem of the task with several operations is to use the linear logistic test model (LLTM). This is a Rasch model in which the difficulty of each item δ_i has been decomposed to "elementary" or "basic" parameters η_j (Spada & Kempf 1977). The probability of getting an item correct is thus dependent on the difficulty estimates of each of the basic parameters (operations) but not the product of the probabilities of getting each of them correct:

$$(5.1.) \qquad P\left(+ \mid \beta_v, \delta_i\right) = \frac{\exp\left(\beta_v - \left(\sum_{j=1}^{m} q_{ij}\,\eta_j + c\right)\right)}{1 + \exp\left(\beta_v - \left(\sum_{j=1}^{m} q_{ij}\eta_j + c\right)\right)}$$

where

η_j = basic parameter (operation)

q_{ij} = frequency of occurence of the operation j in item i

Properties of the LLTM are those of the Rasch simple logistic model (SLM) plus some others (Fischer 1977a).

If some structural model has been used and the results show that the model does not hold, the reason for the misfit may be that assumptions of the model are not valid in the data. Deviations in the case of the LLTM have often been explained quite similarly (e.g. Fischer 1973 ; Häussler 1978; Spada 1980). Among other things they tend to mention that algorithms used by some students are sufficiently different from the algorithms used by others or that the complexity of the solution process is not adequately represented by the relatively simple set of operations used. Another explanation is that the linear approach is altogether inadequate (Scheiblechner 1975).

5.3. The idea of using operational structure

Traditional approaches to testing ignore individual differences between subjects and structural relationships among test items (Hilke, Kempf & Scandura 1977). Completely deterministic structural theory can hardly be used in analysing the curriculum of learning results. Some of its assumptions must be toned down. For example, instead of the statement: "Subject S knows and has a rule available <=> S uses the rule successfully when needed to solve problems" it is more realistic to use the corresponding probabilistic statement: "Subject S_1 knows rule n_j better than S_2 <=> S_1 has a higher probability of using n_j successfully when needed than S_2". If this is the case, it means that ability estimate of S_1 is bigger than that of S_2 if items consist only of the difficulty n_j. That is the case in the SLM in which unidimensionality is assumed and only one item parameter used. In many educational applications, however, information about the structure of a task and success in solving different parts of the task are at least as interesting research problems as information about carrying out the whole task.

If $|\hat{\eta}_j|$ is very small the operation η_j has no meaning in the structure of a task. Big absolute values of $\hat{\eta}$´s are essential with regard to the difficulty estimate of the whole item and to person parameter estimates. To keep the order of person parameters the same, the correct solutions to difficult operations should be given only by persons with the highest β-estimates. That is analogous to the SLM. Each item in the SLM must work similarly, whatever the sample of persons is, if the model holds. In the LLTM each operation η is assumed to work similarly: operations have the same difficulty order in each item (Spada 1980). The property of "sample-free items" could be extended in the case of the LLTM to cope not only with the whole item (which is still supposed to be independent in the person sample) but also with operations which are assumed to be in a sense independent of the item sample. They are thus independent of both person and item samples, and can be measured, in addition, on the absolute scale, if the number of them is smaller than the number of items (Fischer 1977a, p. 205).

These properties of η s make them a useful tool in curriculum research. It is a question about developing the curriculum it always includes the question of what was wrong in the previous version of the curriculum. In many cases this means: what parts of the syllabus were too difficult to learn in some grade of the school? For finding solutions to questions like this some structural analysis of the syllabus is needed at first. In this research it means that before constructing tests for each grade, analysis of basic objectives in the primary grades of the Finnish comprehensive school will have been done. The tests are constructed so that items concentrate on measuring achievements in "basic skills". The test includes some easy items with only a few operations (and easy ones) and also some very difficult items with relative complicated structure and many operations in the same task. Operational structure of a test should be found so that it would be as stable as possible. One reason for criticism of the LLTM is the dubious stability of operations from test

to test. In this research the operations used as the basis of the Q-matrix are received from analysing errors for each test separately. Each wrong answer has been classified and the collection of the most common wrong answers has been labelled as can be seen from Figure 3 (page 22) in which the final error structure for primary levels has been presented by means of a flow diagram. It shows that reasons for errors are basically of two main types: Firstly, choosing an incorrect calculation and secondly leaving the task unfinished. Both main types of errors have been analysed more closely for finding a suitable category for each single error.

5.4. Research problems

The linear logistic test model has been used basically when it is a question of analysing some complicated learning task. Typical test are constructed so that each item is a combination of relatively few operations (rules) which must be known for solving items correctly. Even if dichotomous scaling had been used for items, the estimates for the difficulty of operations ($\hat{\eta}_j$, j=1,...,m) are on an absolute scale if the number of operations is smaller than the number of items and the frequency matrix $Q = ((q_{ij}))$ is of full rank (Fischer 1977a).

Also this chapter the main interest will be to find a relevant cognitive structure for 6 Mathematics tests administered to primary school pupils in Finland. This research problem can be divided into the following smaller problems.
Problem 1: The whole item has been divided into elementary operations. If the subject can solve each operation separately, this does not necessarily mean that he can solve the whole item, or conversely: if he can solve the item it does not mean that he can solve each elementary operation.
Problem 2: Estimates of difficulty parameters in the SLM and the LLTM will be compared to each other using the

graphical method. The main interest in this problem will be in misfitting items and finding reasons for misfit.
Problem 3: Linking of consecutive grades and all the 6 grades using estimates of elementary operations will give some information about the learning process in the primary grades in the area of equations.
Problem 4: Properties of a good mathematical model (page 42) will be taken into consideration after empirical research problems in the case of LLTM.
Problem 5: The last problem will deal with the benefit of the LLTM - research for developing Mathematics syllabus.

5.5. The item and basic operation in the LLTM

In order to find data which fits the LLTM, basic parameters must be selected so that every subject can be assumed to be solving items using the same stages (operations, rules) for finding the correct solution. In each item basic operations are also assumed to work similarly. In other words, the operations are assumed to have a constant rank order of difficulty (Spada 1980).

The latter property is analogous to the SLM which is "sample-free": that is to say, all items work similarly in each subgroup of persons if the model holds. They have the same order of difficulty whatever the sample of subjects is. It could be said that elementary parameters are "item-free" in the LLTM: they have the same difficulty order, whatever the item is, if the test is constructed so that the LLTM holds.

Some deviations can always be expected, because different people do not use exactly the same operations in their solutions. On the other hand, somebody may know all the rules which are necessary for getting the correct answer; he is, however, not able to apply rules to an item. It is possible for him to solve each operation separately but it may be too difficult to solve the whole item.

Spada (1980) considers that it is a serious drawback of the LLTM that the task solution probabilities cannot be understood as the products of the corresponding operation probabilities, because in general

$$(5.2) \qquad \frac{\exp(\beta_v - \sum_{j=1}^{m} q_{ij}\eta_j)}{1 + \exp(\beta_v - \sum_{j=1}^{m} q_{ij}\eta_j)} \neq \prod_{j=1}^{m} \{\frac{\exp(\beta_v - \eta_j)}{1 + \exp(\beta_v - \eta_j)}\}^{q_{ij}}$$

Equation (5.2) means that the probability of the correct answer to the whole item cannot be considered to be the product of the probabilities of getting each elementary operation correct. That can easily be seen using an artificial example. Let´s suppose that η_1 = to drive a car and $\hat{\eta}_1 = 3.0$, η_2 = to read a newspaper, $\hat{\eta}_2 = 2.0$. It is less likely that one will succeed in doing bouth things simultaneously than in doing both things separately. If β_v is assumed to be 2.0, the probability of succeeding in the combination of the two operations is

$$(5.3) \qquad P_+ = \frac{\exp(\beta_v - \delta_i)}{1 + \exp(\beta_v - \delta_i)} = \frac{e^{2-5}}{1 + e^{2-5}} = .047$$

and the product of the two separate probabilities is

$$(5.4) \qquad P_{\eta_1 +} \cdot P_{\eta_2 +} = \frac{\exp(\beta_v - \eta_1)}{1 + \exp(\beta_v - \eta_1)} \cdot \frac{\exp(\beta_v - \eta_2)}{1 + \exp(\beta_v - \eta_2)} = .134.$$

The latter probability is higher which means that the whole task is "more" than the combination of the two operations. If $\beta_v = 0$, it is the other way round.

An interesting theoretical question is to compare the difficulties of the whole item and those of the operations which form the item. The probability of getting the whole item i correct is (index i omitted for simplicity)

$$(5.5) \qquad P_+ = \frac{\exp\left(\beta_v - \sum_{j=1}^{m} q_j \eta_j\right)}{1 + \exp\left(\beta_v - \sum_{j=1}^{m} q_j \eta_j\right)}$$

For different operations the corresponding probability is the product

$$P_{\eta_1+} \ldots P_{\eta_m+} = \left(\frac{\exp(\beta_v - \eta_1)}{1 + \exp(\beta_v - \eta_1)}\right)^{q_1} \ldots \left(\frac{\exp(\beta_v - \eta_m)}{1 + \exp(\beta_v - \eta_m)}\right)^{q_m}$$

$$(5.6) \qquad P_{\eta_1+} \ldots P_{\eta_m+} = \frac{\exp\left(\beta_v \sum_{j=1}^{m} q_j - \sum_{j=1}^{m} q_j \eta_j\right)}{\prod_{j=1}^{m} \left(1 + \exp(\beta_v - \eta_j)\right)^{q_j}}$$

Probabilities can be compared in different ways (Cox 1970). In this case the most useful is to calculate log odds for each probability and to find the difference between them.

Log odds in the case of the whole item are

$$(5.7) \qquad \Lambda_v = \ln \frac{P_+}{1 - P_+} = \beta_v - \sum_{j=1}^{m} q_j \eta_j$$

and in the case of operations

$$(5.8) \qquad \lambda_v = \beta_v \sum_{j=1}^{m} q_j - \sum_{j=1}^{m} q_j \eta_j - \ln \left\{ \prod_{j=1}^{m} \left(1+\exp(\beta_n - \eta_j)\right)^{q_j} - \exp\left(\beta_v \sum_{j=1}^{m} q_j - \sum_{j=1}^{m} q_j \eta_j\right) \right\}$$

For simplicity's sake, also indices v will be left out, we are concerned with the same person v (and the same item i) all the time. The difference of log odds is

$$(5.9) \qquad \lambda - \Lambda = \left(\sum q_j - 1\right) \beta - \ln \left\{ \prod_j \left(1 + e^{\beta - \eta_j}\right)^{q_j} - e^{\beta \sum q_j - \delta} \right\}$$

This is a general form for the difference of log odds in the case of the linear logistic model. In the simple logistic model the equation (5.9) can be simplified because $\sum q_j = 1$. In this case $\lambda - \Lambda = 0$, which means that the whole item is at the same time the only basic operation. It can be considered that for the SLM the Q-matrix is the identity matrix I.

The sum of row frequencies in the Q-matrix can be considered to indicate the complexity of the item:

$$\sum_{j=1}^{m} q_j = \gamma. \tag{5.10}$$

After this definition equation (5.9) can be written

$$\lambda - \Lambda = (\gamma - 1)\beta - \ln\left(\prod_j (1 + e^{\beta - \eta_j})^{q_j} - e^{\gamma\beta - \delta}\right) \tag{5.11}$$

The case below examines the situation when $\lambda > \Lambda$ in general case of the LLTM (SLM is excluded, that is to say $\gamma > 1$).

$$\begin{aligned}
& \lambda - \Lambda > 0 \\
\Leftrightarrow\ & (\gamma - 1)\beta > \ln(\ ) \\
\Leftrightarrow\ & \ln e^{(\gamma-1)\beta} > \ln(\ ) \\
\Leftrightarrow\ & e^{(\gamma-1)\beta} > \prod_j (1+e^{\beta-\eta_j})^{q_j} - e^{\beta\gamma-\delta} \\
\Leftrightarrow\ & e^{(\gamma-1)\beta} + e^{\beta\gamma-\delta} > \prod_j (1+e^{\beta-\eta_j})^{q_j} \\
\Leftrightarrow\ & e^{\beta\gamma}(e^{-\beta}+e^{-\delta}) > \prod_j (1+e^{\beta-\eta_j})^{q_j} \\
\Leftrightarrow\ & e^{-\beta}+e^{-\delta} > \frac{\prod_j (1+e^{\beta-\eta_j})^{q_j}}{e^{\beta\gamma}} \\
\Leftrightarrow\ & e^{-\delta} > \frac{\prod_j (1+e^{\beta-\eta_j})^{q_j}}{e^{\beta\gamma}} - e^{-\beta}
\end{aligned} \tag{5.12}$$

so it is easier to solve each operation separately if inequation (5.12) holds. In Figures 32a and b both cases can be seen.

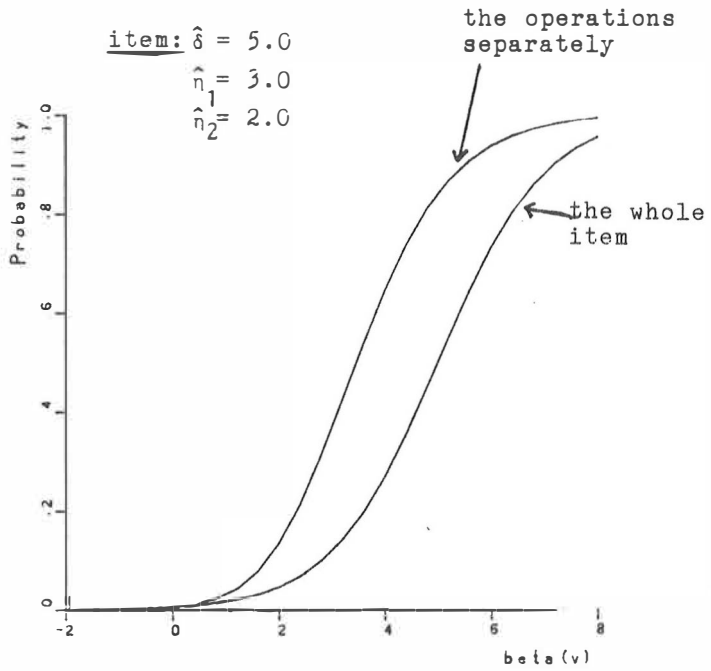


Figure 32a. Probabilities of solving the operations separately is higher than that for the whole item.

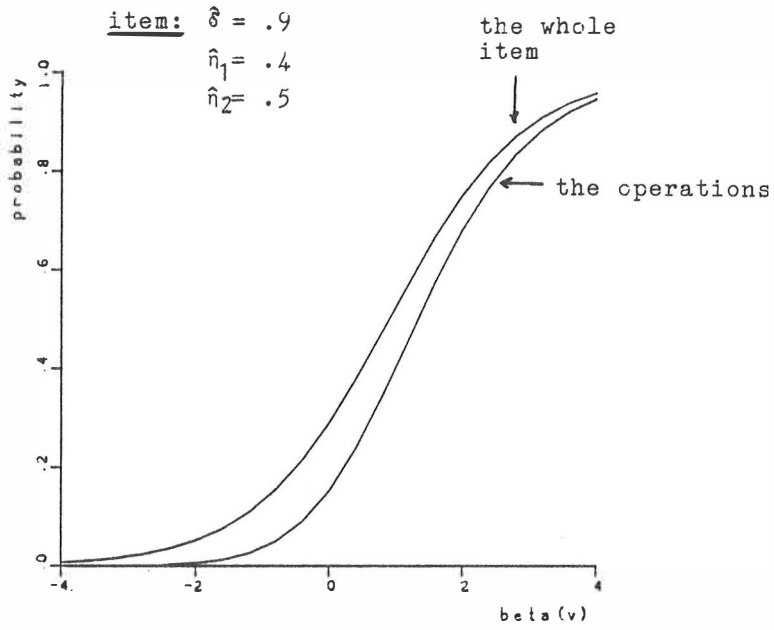


Figure 32b. Probabilities of solving the whole item is higher than that of solving operations separately.

Probabilities of correct answers and log odds both measure the same thing. In extreme cases (very low or very high probability of solving the item correctly) the log odds is a more sensitive measure of change. Difference of log odds is at least 4 times as big as the difference of the corresponding probabilities. The derivate function of log odds, $y = \frac{1}{x(x-1)}$, illustrates clearly what happens in extreme cases. The inverse function of log odds is $y = \frac{e^x}{1 + e^x}$,

which is the logistic function in Rasch models. The absolute value of derivative function increases very fast when the probability p -> 1 or p -> 0. The same effect, but in opposite form, can be seen in logistic function: the slope of that curve is very small in extreme cases ($\beta \to \infty$ or $\beta \to -\infty$).

5.6. Basic operations and Q-matrix for the LLTM

In this study the selection of basic operations is based on error analysis. If an error occurs very often in different items it means that the same difficulty η_x is included in different tasks and that special difficulty makes these items difficult. When the Q-matrix is based on errors it is likely that operations η_1, η_2,..., η_m have a constant order of difficulties in each item, as far as each error exists consistently if an item gives the oportunity to make that error. If only a few basic parameters are used in the LLTM (for example, only addition, subtraction, multiplication and division) they are more likely to follow the assumption of constant rank order from item to item. Even though this is desirable, it cannot be used very easily because it tends to make the columns of two basic parameters linearly dependent. In each item of test 1 (except the last one) either addition or subtraction is included (in the last one both addition and substraction). The matrix of frequencies must be constructed so that columns are independent and if possible the number of zeroes (and ones) is not close to the number of items. If it is, the confidence interval of η will become big and estimates will not be accurate enough for any decisions. Very extensive elementary parameters are hardly useful: usually they must be divided into smaller ones.

Table 30. Labels for basic parameters

Name	Grade 1	2	3	4	5	6
ADDITION, variable is the sum total	η_1					
ADDITION, variable is an addendum (for ... graders)	η_2	η_2				
SUBTRACTION, variable is the difference (for ... graders)	η_3	η_3				
SUBTRACTION, variable is first or second term (for ... graders)	η_4	η_4				
Is there more than one number which is $\geq$ 10? How many more?	η_5					
Are there any figures carried over or is there borrowing in the item?	η_6	η_6	η_4	η_4	η_4	η_4
Is it a verbal problem or an inequation	η_7	η_7	η_5	η_5	η_5	η_5
Does the item give possibility to use "short cut" and get a wrong answer?	η_8	η_8	η_6	η_6	η_6	η_6
ADDITION or MULTIPLICATION		η_1				
Is there more than one number which is $\geq$ 20? How many more?		η_5				
ADDITION for ... graders			η_1	η_1	η_1	η_1
SUBTRACTION for ... graders			η_2	η_2	η_2	η_2
MULTIPLICATION for ... graders			η_7	η_7	η_7	η_7
DIVISION for ... graders			η_8	η_8	η_8	η_8
Is there more than one number which is $\geq$ 100? How many more?			η_3			
Is there more than one number which is $\geq$ 100 or are there decimal numbers in item? How many?				η_3		
Is there more than one number which is $\geq$ 100 or are there decimal numbers or fractions in item? How many?					η_3	η_3

Table 31. Q-matrix for 1st grade (N = 202)

Item n:o	Elementary parameters								Percentage of correct answers
	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	
1	1	0	0	0	0	0	0	0	94
2	0	1	0	0	0	0	0	1	89
3	0	1	0	0	0	0	0	1	81
4	1	0	0	0	0	1	0	0	81
5	0	1	0	0	0	1	0	1	73
6	0	1	0	0	0	1	0	0	72
7	0	0	1	0	0	0	0	0	91
8	0	0	0	1	0	0	0	0	91
9	0	0	0	1	0	0	0	1	57
10	0	0	0	1	1	1	0	1	58
11	0	0	0	1	1	0	0	0	66
12	0	0	0	1	1	1	0	1	33
13	0	0	0	1	2	0	0	0	34
14	0	1	0	0	2	0	0	1	67
15	0	0	0	1	2	0	0	0	64
16	0	0	1	0	0	0	1	0	49
17	1	0	0	0	0	0	1	0	56
18	0	0	0	1	0	0	1	1	81
19	1	1	0	0	0	0	0	1	70
20	0	0	1	1	0	1	0	0	65
21	1	1	0	0	3	2	0	1	30
22	0	0	1	1	2	1	0	0	20
23	1	0	0	0	2	1	0	0	28
24	0	0	2	0	2	0	0	0	54
25	0	0	1	0	0	1	1	1	68
26	0	0	1	0	1	0	1	0	72
27	2	0	0	0	1	2	1	0	53
28	0	0	1	0	2	2	1	0	26
29	0	1	0	0	2	1	1	1	16
30	1	0	1	0	3	2	1	0	18

At the same time it should be kept in mind that new parameters must have an item-free rank order of difficulty.

Different η's are not necessarily completely independent. They may be correlated. For example, in the first grade η_3 = subtraction, the variable is the difference and η_4 = subtraction, the variable is the first or the second term are highly correlated. If a person cannot solve the item 8 - 3 = $\square$ he usually cannot solve the item 6 - $\square$ = 2 either. However, each η-parameter must be defined so clearly that the structure of the Q-matrix is unique (Table 31). The rows of Q-matrix express, in a sense, the complexity of an item. The sum of frequencies on a row can be considered to measure complexity $\gamma = \sum_{j=1}^{m} q_j$. Complex items also tend to be difficult. Difficult items need not be complex. It may be that all verbal problems, for example, are difficult, although only one calculation is enough for solving them. The difficulty in this case is not the complexity but the interpretation of the problem. Having selected η-parameters, the construction of the Q-matrix is usually a routine task. However, some frequencies seem to become too big very easily: for example, one subtraction, one multiplication and one division. Each of these basic operations is very easy to calculate. The structure of this type of item becomes too complicated merely because of consistency in the construction of the Q-matrix. The goodness of fit test reveals these kind of deviations.

In the previous examples the item is easier to solve than it is expected to be. Some other item may have exactly the same vector of basic operations but it is more difficult because numbers are complicated fractions or decimal numbers. The domain of numbers cannot be without influence on four basic calculations although it has been taken into account as a separate basic parameter. Another evidence of this is that in the higher grades of the primary school subtraction in the new domain of numbers is a new task which must be learned and practised again. It cannot be taken as granted

that the idea of subtraction is the same, whatever the numbers are. Every new domain is a new difficulty for pupils. In this research the names of some basic operations have been specified afterwards for that reaseon, e.g. "subtraction" has become "subtraction in the 4th grade". Basic parameters are listed in Table 30.

5.7. Standardizing of item difficulties

Table 30 gives the names of elementary parameters. It includes three basic operations which are common for each grade. However, one of them - "figures carried over or borrowing" - is dependent on the grade. Using two common parameters linking has been done later.

In this research it seems to be reasonable to standardize difficulty estimates so that both the SLM and the LLTM estimates will have the same standard deviation 1.0. The reason for standardizing is the small range of the LLTM-estimates compared to those of the SLM. If no standardizing has been used it means that misfitting items will mostly be the easiest or the hardest ones. A small range of the LLTM-estimates is a sign of the fact that at the primary level the operations are not as different in difficulty levels as would be expected when the SLM is in question. By standardizing the percent of misfitting items will be changed from 25 % to 13 % of all items.

5.8. Empirical results

5.8.1. Operations and their confidence limits

The LLTM programme (Niehusen, Mach, Hansen, Rost & Kempf 1978) gives conditional maximum likelihood estimates and confidence limits for basic parameters. In this programme they are expressed in the form of "easiness". For convenience

they have been changed in this study to the scale of difficulty by taking an opposite sign for each $\hat{\eta}$ and simultaneously changing lower and upper limits of the confidence intervals. The same thing has also been done to item difficulties because of test of fit. It is only a question of two slightly different ways of using the same Rasch model:

$$(5.13) \qquad P_+ = \frac{\theta_v \varepsilon_i}{1 + \theta_v \varepsilon_i} = \frac{\exp(\beta_v - \delta_i)}{1 + \exp(\beta_v - \delta_i)}$$

Both interpretations are alternatives. In this study the latter form has been used consistently.

Figures 33a, b and c give values on $\hat{\eta}$'s and their 95 % confidence limits on a difficulty scale. The most interesting thing is that in the main the order of $\hat{\eta}$'s is the same (grades 1 & 2, 3 & 4, 5 & 6) with some important exceptions. For example the most difficult operation for first graders is η_5 = big numbers. For second graders it turns out to be the easiest one. This result can be interpreted as meaning that, in the second grade, pupils have learned the idea of the ten-based system of numbers and they have had a sufficient amount of practice at overcoming problems which they had in the first year with big numbers. For first graders the operation η_1 = addition, sum is variable is very easy but for second graders it is much harder. In fact it is not a question of the same operation because of a different domain of numbers.

1st grade		2nd grade	
estimates	95 % confidence intervals	estimates	95 % confidence intervals
$\hat{\eta}_1 = .018$	(-.20, .23)	$\hat{\eta}_1 = .45$	(.33, .57)
$\hat{\eta}_2 = .087$	(-.20, .37)	$\hat{\eta}_2 = -.046$	(-.23, .14)
$\hat{\eta}_3 = .20$	(-.027, .44)	$\hat{\eta}_3 = .31$	(.11, .50)
$\hat{\eta}_4 = .56$	(.33, .80)	$\hat{\eta}_4 = .35$	(.13, .57)
$\hat{\eta}_5 = .82$	(.74, .90)	$\hat{\eta}_5 = -.077$	(-.17, 01)
$\hat{\eta}_6 = .48$	(.35, .61)	$\hat{\eta}_6 = .19$	(.088, .28)
$\hat{\eta}_7 = .66$	(.49, .83)	$\hat{\eta}_7 = .57$	(.46, .67)
$\hat{\eta}_8 = .17$	(-.046, .38)	$\hat{\eta}_8 = .16$	(.0018, .31)

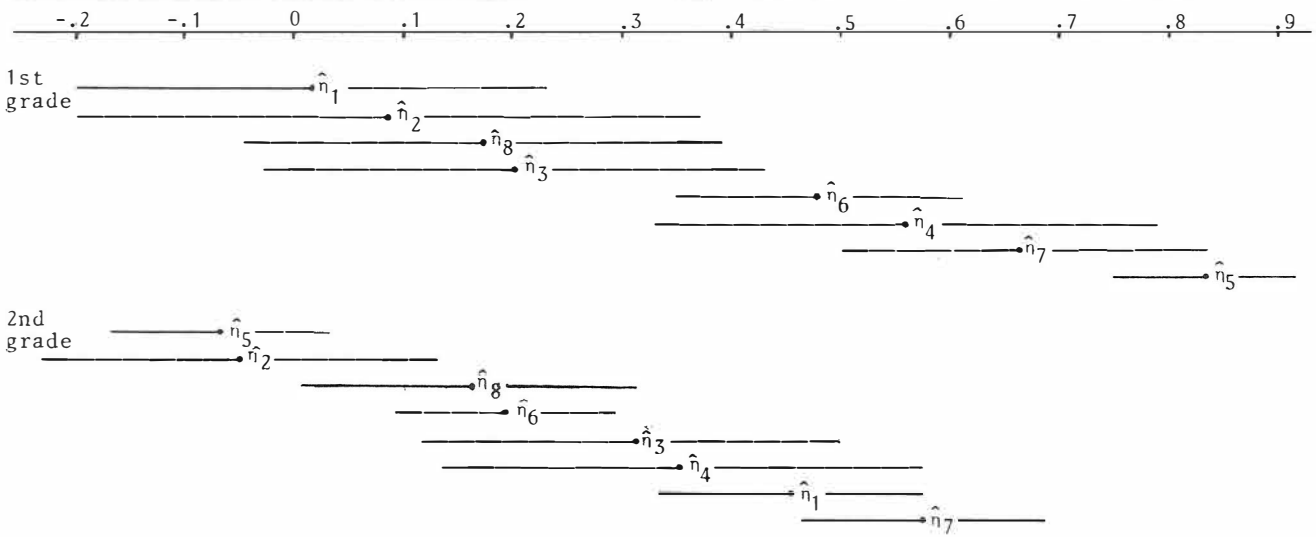


Figure 33a. Estimates of basic parameters, 1st and 2nd grade.

3rd grade		4th grade	
estimates	95 % confidence intervals	estimates	95 % confidence intervals
$\hat{\eta}_1 = .43$	(.28, .57)	$\hat{\eta}_1 = .022$	(-.11, .15)
$\hat{\eta}_2 = -.12$	(-.31, .063)	$\hat{\eta}_2 = .31$	(.11, .50)
$\hat{\eta}_3 = .55$	(.47, .64)	$\hat{\eta}_3 = .35$	(.22, .48)
$\hat{\eta}_4 = .22$	(.087, .36)	$\hat{\eta}_4 = -.058$	(-.17, .053)
$\hat{\eta}_5 = .87$	(.73, 1.02)	$\hat{\eta}_5 = .90$	(.72, 1.07)
$\hat{\eta}_6 = -.06$	(-.21, .091)	$\hat{\eta}_6 = -.079$	(-.26, .11)
$\hat{\eta}_7 = .31$	(.21, .41)	$\hat{\eta}_7 = .75$	(.57, .92)
$\hat{\eta}_8 = 1.42$	(1.24, 1.60)	$\hat{\eta}_8 = 1.35$	(1.16, 1.55)

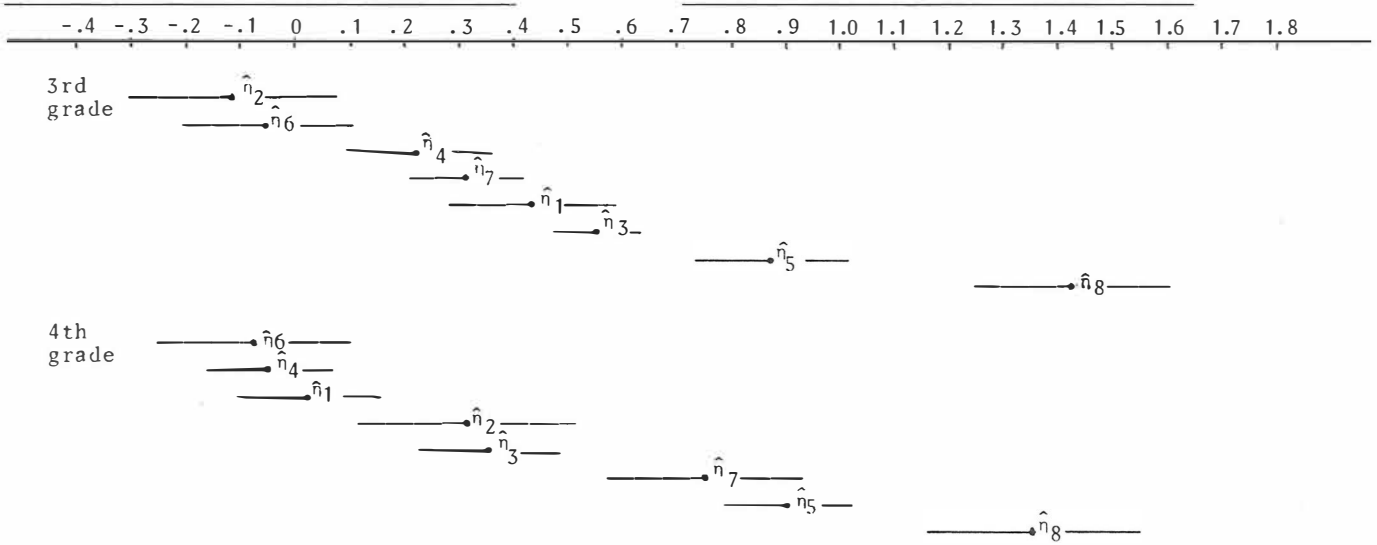


Figure 33b. Estimates of basic parameters, 3rd and 4th grade.

5th grade		6th grade	
estimates	95 % confidence intervals	estimates	95 % confidence intervals
$\hat{\eta}_1 = .93$	(.72, 1.13)	$\hat{\eta}_1 = .13$	(-.046, .31)
$\hat{\eta}_2 = .35$	(.17, .53)	$\hat{\eta}_2 = .67$	(.48, .85)
$\hat{\eta}_3 = .14$	(.074, .21)	$\hat{\eta}_3 = .53$	(.48, .59)
$\hat{\eta}_4 = .21$	(.11, .30)	$\hat{\eta}_4 = -.12$	(-.22, -.012)
$\hat{\eta}_5 = .71$	(.57, .84)	$\hat{\eta}_5 = .63$	(.48, .78)
$\hat{\eta}_6 = -.42$	(-.60, -.25)	$\hat{\eta}_6 = -.15$	(-.28, -.023)
$\hat{\eta}_7 = .62$	(.48, .75)	$\hat{\eta}_7 = .94$	(.85, 1.03)
$\hat{\eta}_8 = 2.07$	(1.89, 2.25)	$\hat{\eta}_8 = 1.20$	(1.03, 1.38)
$\hat{\eta}_9 = .80$	(.72, .87)	$\hat{\eta}_9 = .67$	(.58, .75)

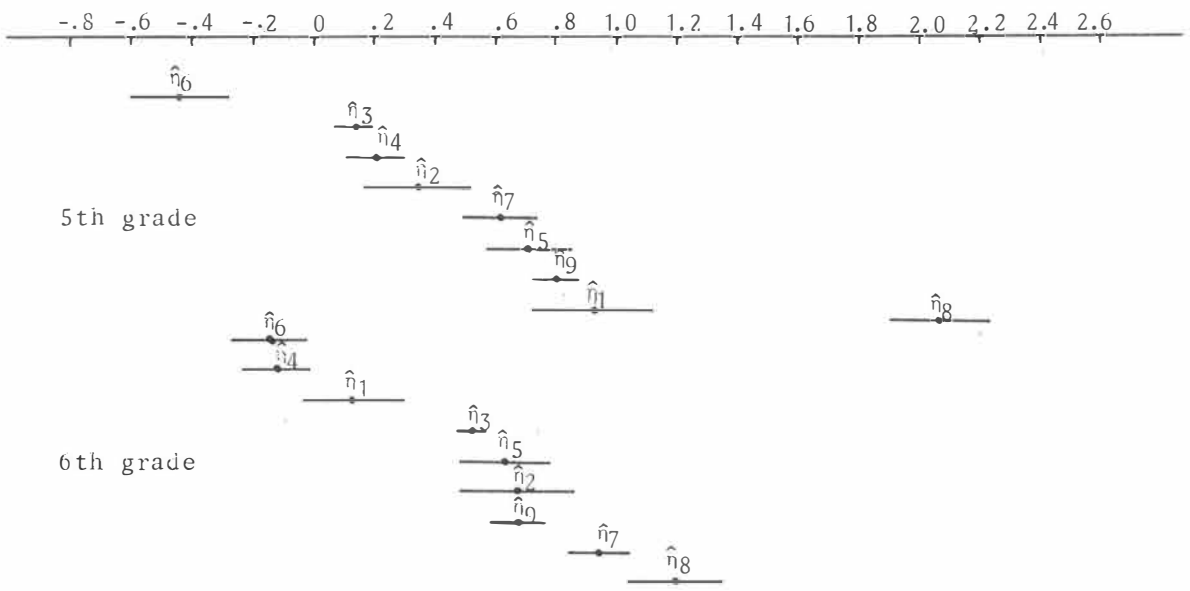


Figure 33c. Estimates of basic parameters, 5th and 6th grade

At first sight it may seem confusing that addition has become more difficult and simultaneously dealing with big numbers has become easier for second graders. A closer look at the items reveals that addition, where the sum is variable, has been used not only in direct addition items like 23 + 12 = ☐ but also in items like $\begin{smallmatrix}14\\+18\end{smallmatrix}$ or 18 + 14 = 14 + ☐. The latter of these does not necessarily need any addition. However, pupils´ errors gave a hint to the effect that at least those who had not found the correct answer had tried to add 18 and 14 at first and then subtract 14 from the sum. The same holds for item 25 + 17 + 5 ☐ 17 + ☐. In the operation vector of this item two additions and one subtraction exist. An equally correct way would be to argue that there are no subtractions and only one addition. Also multiplication for second graders has been interpreted as going together with addition. For item 3 · 124 = ☐ and some others the operation vector includes addition because basically only multiplication tables from the 1st to the 5th are known by second graders, in all other tasks they must use addition. For reasons mentioned before the "addition" is later called "addition or multiplication in the 2nd grade". In most cases in each consecutive grade (1 & 2, 3 & 4, 5 & 6) operations become easier. In exceptional cases it is question of some of four basic calculations and a different domain of numbers being used for each grade that makes operations different.

5.8.2. Test of fit

The simple logistic model gives item difficulties δ_i which are based on the number of correct answers in each item. The corresponding item difficulties δ_i^* in the LLTM are compared with them in test of fit. Both estimates are scaled by computer so that

$$\sum_{i=1}^{k} \delta_i = \sum_{i=1}^{k} \delta_i^* = 0. \qquad (5.14)$$

Particularly in the data of primary pupils, operations in early grades are not so different that the variance of δ_i^* would become as big as the variance of δ_i is. That is why in this study both statistics for item difficulties have been standardized so that not only the sum of estimates is zero but also the variance of each distribution is equal to one. The goodness of fit test is based on the graphical method. If no standardizing has used, the easiest and hardest items (25 % of all 180 items) are misfitting. Standardizing decreases the percentage to 13. For each misfitting item some credible reason for misfit can be found. Standard deviations of item difficulties initially are:

Table 32. Standard deviations of item difficulty estimates before standardizing

Grade	$sd(\delta_i)$	$sd(\delta_i^*)$
1	1.50	1.13
2	1.14	.57
3	1.47	.84
4	1.27	.81
5	1.37	.99
6	1.40	.87

Standardizing has been done using the formula

$$(5.15) \qquad \hat{\delta}_{is} = \frac{\hat{\delta}_i}{sd(\delta_i)} \quad \text{and} \quad \hat{\delta}^*_{is} = \frac{\hat{\delta}^*_{is}}{sd(\delta^*_i)}$$

Figures 34 and 35 give two examples of the graphical test of fit.

5.8.3. Statistical analysis of misfit

This statistical analysis is based on a comparison of the order of difficulty of items (in SLM) and their simplified basic structure. Difficulty of items should increase in both Rasch models (SLM and LLTM) similarly. Deviations from this structure have been interpretated to reveal misfit in item structure.

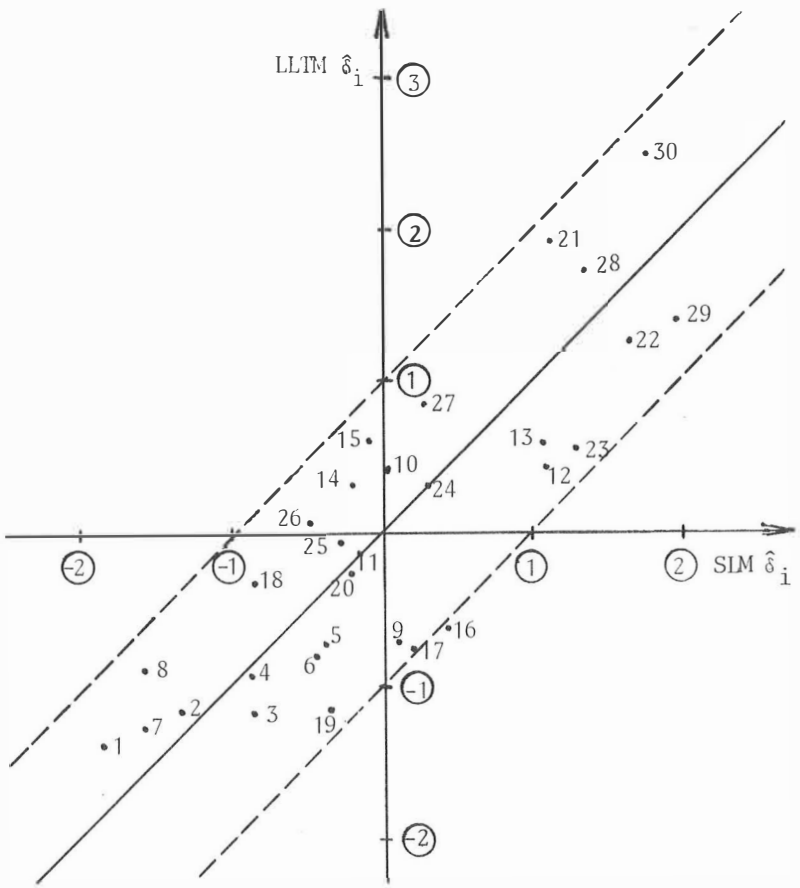


Figure 34. Standardized item difficulties, graphical test of fit (1st grade)

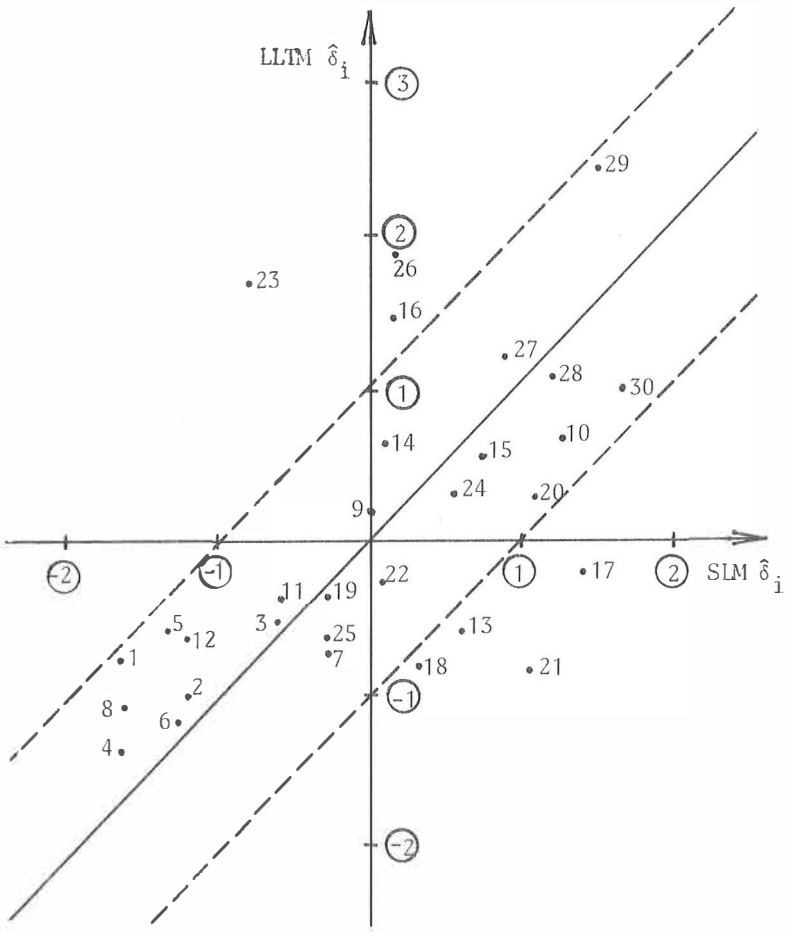


Figure 35. Standardized item difficulties, graphical test of fit (2nd grade)

For each test the same steps have been used for illustrating the significance of some items not fitting the model as well as some others do. Steps are:

(1) Items have been arranged in order of difficulty according to SLM estimates. In each test all 30 items have been used in spite of the fact that some of them (only a few) may be misfitting.

(2) Basic parameters (8 parameters for 1st, 2nd, 3rd and 4th graders and 9 parameters for 5th and 6th graders) have also been arranged in order of difficulty.

(3) If estimates of some basic parameters are very near to 0 (0 belongs to their confidence interval) they have been left out of the next analysis because these operations are not difficult.

(4) If estimates of some basic parameters are very close to each other, they are considered to be the same.

By these four steps the number of basic parameters has been reduced from eight or nine to five in each grade. The idea behind this is to take only different η's and those which have some influence on item difficulty into account.

(5) The first positive η-estimate (or the mean of the first two or three if they are very close to each other) has been replaced by 1. The estimate of each basic parameter has been compared to the first η-estimate and its value has been replaced by the nearest integer. For example for first graders the difficulty order of $\hat{\eta}$'s is:

$\hat{\eta}_1$=.018, $\hat{\eta}_2$=.087, $\hat{\eta}_8$=.17, $\hat{\eta}_3$=.20, $\hat{\eta}_6$=.48, $\hat{\eta}_4$=.56, $\hat{\eta}_7$=.66, $\hat{\eta}_5$=.82.

Estimates on η_1, η_2 and η_8 can be considered to be equal to zero, the integers 1, 2, 3, 3, 4 can be substituted for the next five and they are still on a ratio scale. Each operation has been compared to η_3. The integers reveal how many times as difficult the other basic parameters are compared to η_3.

(6) For each item, the weighted sum of five operations has been calculated. For example for the most difficult item 29 this is:

$1 \cdot \hat{\eta}_6^* + 1 \cdot \hat{\eta}_7^* + 2 \cdot \hat{\eta}_5^* = 13$

(where $\hat{\eta}_i^*$ is the integer-estimate of η_i)
The sum is a rough estimate of "structural difficulty" (LLTM-difficulty) of the item. It turns out to be accurate enough for revealing misfit (Table 34).

(7) Cumulative frequencies have been calculated using the sums mentioned before (Table 34).

(8) In the frequency polygon of cumulative frequencies deviations can be seen clearly in the height of steps (Figure 36).

The integers mentioned in step 5 can be calculated for each grade:

Table 33. Integer estimates for operations

Grade	Operations
1	$\hat{\eta}_3^* = 1$ $\hat{\eta}_6^* = 2$ $\hat{\eta}_4^* = 3$ $\hat{\eta}_7^* = 3$ $\hat{\eta}_5^* = 4$
2	$\hat{\eta}_8^* = \hat{\eta}_6^* = 1$ $\hat{\eta}_3^* = 2$ $\hat{\eta}_4^* = 2$ $\hat{\eta}_1^* = 3$ $\hat{\eta}_7^* = 3$
3	$\hat{\eta}_4^* = \hat{\eta}_7^* = 1$ $\hat{\eta}_1^* = 2$ $\hat{\eta}_3^* = 2$ $\hat{\eta}_5^* = 3$ $\hat{\eta}_8^* = 5$
4	$\hat{\eta}_2^* = 1$ $\hat{\eta}_3^* = 1$ $\hat{\eta}_7^* = 2$ $\hat{\eta}_5^* = 3$ $\hat{\eta}_8^* = 4$
5	$\hat{\eta}_6^* = -2$ $\hat{\eta}_3^* = \hat{\eta}_4^* = \hat{\eta}_2^* = 1$ $\hat{\eta}_7^* = \hat{\eta}_5^* = 3$ $\hat{\eta}_9^* = \hat{\eta}_1^* = 4$ $\hat{\eta}_8^* = 9$
6	$\hat{\eta}_6^* = \hat{\eta}_4^* = -1$ $\hat{\eta}_1^* = 1$ $\hat{\eta}_3^* = \hat{\eta}_5^* = \hat{\eta}_2^* = \hat{\eta}_9^* = 5$ $\hat{\eta}_7^* = 7$ $\hat{\eta}_8^* = 9$

Table 34. Finding misfitting items in the case on LLTM 2nd grade

Item	0		1		2		3		Weighted sum	Cumulative sum
	η_5	η_2	η_8	η_6	η_3	η_4	η_1	η_7		
1	1						1		3	3
4	1	1	1						1	4
8	2					1			2	6
5	2			1			1		4	10
6	2				1				2	12
2	1	1	1	1					2	14
12							1		3	17
23	1			2			2	1	[11]	28
3	1			1			1		4	32
11			1				1		4	36
7	1			1	1				3	39
19	1		1	1		1			4	43
25	1	1						1	3	46
9		1	1	2			1		6	52
14			1				1	1	7	59
16			1				2	1	[10]	69
22	2			2			1		5	74
26					1		1	2	[11]	85
18	2			1	1				(3)	88
24							1	1	6	94
13							1		(3)	97
15		1	1				1	1	7	104
27	1						1	2	9	113
20				1		1	1		6	119
21	1						1		(3)	122
28	3			2	3		1	1	10	132
10	2	1	1	2			2		9	141
17			1					1	(4)	145
29	2			3	1		2	1	14	159
30	1						2	1	9	168

SLM difficulty order

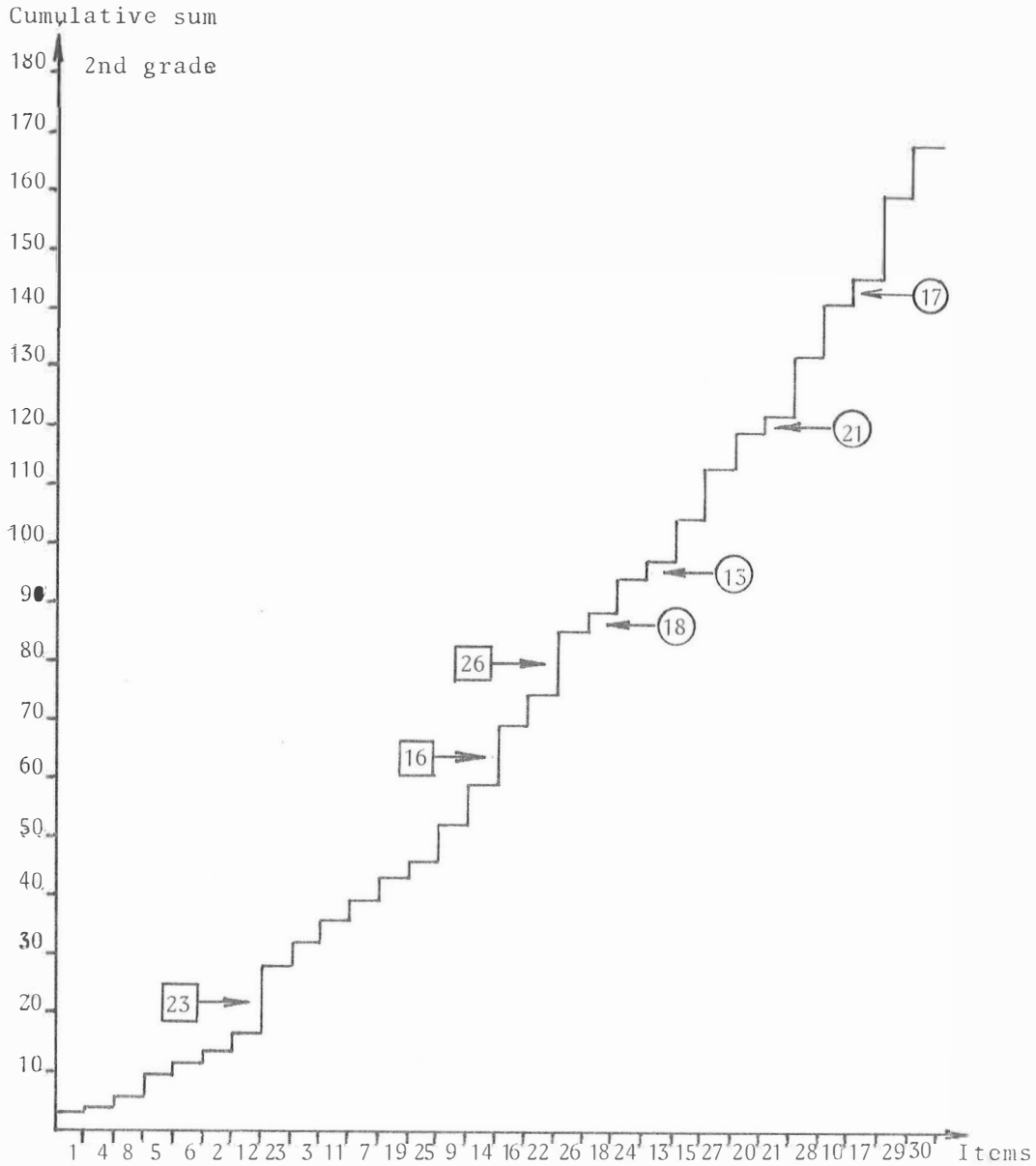


Figure 36. A test of fit based on the height of the steps in the cumulative frequency polygon

Negative estimates of basic parameters in the last two grades mean that these operations are no longer difficult for fifth and sixth graders. It does not necessarily mean that some operation makes items easier. It may also be interpreted as meaning that it is some other common property of those items in which it is included, but it has nothing to do with the difficulty of those items. For the sake of balance, negative estimates in the table must also be taken into account. Operation η_6 is also important because it is one of the two common operations for each grade. If some further analysis were made using data of 5th or 6th graders, the operations which have got negative or zero estimates could be eliminated. They give very little information about the difficulty of the item.

5.8.4. Recommendations for constructing items for LLTM

The graphical method for goodness of fit is based on the assumption that the LLTM holds if it gives the same estimates for item difficulties as the SLM does. If the test shows good fit it implies that variances of both estimates are also the same. The converse does not hold. Even if variances of estimated item difficulties are equal, items may be misfitting. If it is the case that in two sets of item difficulties variances of estimates are completely different it implies that test of fit will always end up with misfit, because at least the most extreme item difficulties cannot be equal. For getting the graphical test of fit to work properly it is a necessary starting point to standardize SLM- and LLTM-item difficulties so that both of them have mean 0 and variance 1. Misfit can be observed when studying deviations from the line (Figure 34). Two kinds of deviation can be found:

(i) The point is above the line. This means that the LLTM gives a higher estimate on item difficulty than the SLM does. If this is the case, the operational structure

of the item must be more closely examined. It is likely that the misfitting item is regarded as a more complex combination of operations than it really is. If the item is easy (δ_i in SLM is small) it may be that the subjects mostly know the answer by heart without using any operations. If the item is difficult (δ_i in SLM is big) and the item is misfitting, it means that it was not as difficult as expected. Even if it is a combination of difficult operations, it probably is the case that both numbers used in the item are so simple that some operations can be done correctly routinely. These operations do not work similarly in a misfitting item as they do in other items. To avoid this it is suggested that the whole test be constructed so that, as far as the domain of numbers is conserned, all items are approximately at the same level of difficulty. It does not help very much if number domain has been taken to be an extra η-operation, because it is always related to the computations and strongly correlated with their estimates.

(ii) The point is below the line. That is to say the LLTM gives smaller estimates to item difficulty than the SLM does. Misfit can be explained by saying that misfitting items were not regarded to be as complex as they really are. There may be something missing in their operational structure. It is likely that either at least one extra η-operation is needed for describing the structure of these items, or the numbers used in the items are more difficult than those in fitting items. If the former case is more likely, it implies that the Q-matrix must be corrected; if the latter is more probable, only the same guideline for test construction can be given as was mentioned before. For example, in the data of second graders reasons for misfit can be found in each case separately.

In the optimal case, the set of basic operations should have been selected so that the number of zeroes in each column of Q-matrix is about the same as the number of non-zero frequences in that column. For example in the data for second graders the weighted sum will reveal misfitting items as effectively as the graphical method (Figures 35 and 36).
In constructing items the next points which are essential in applying Rasch models are taken into account also in the present study. They are essential also in the case of the LLTM.

(i) Possibility of guessing has been eliminated by avoiding multi-choice items. All items for primary school are open ended. This kind of set of items also gives the opportunity to study more closely different types of errors. Analysis of errors gives a good basis for finding LLTM´s basic operations.

(ii) Items in the tests concentrate on the area of basic objectives because they have been taught to every pupil in nearly the same way. That ensures that tests are unidimensional also in the sense of subjects.

(iii) Tests have got some very easy and some very difficult items. Each pupil should be able to answer at least one item correctly and nobody should get all items right. Items in each grade have been constructed for each ability level, keeping in mind that there are enough items for average pupils in order to distinguish pupils from each other.

(iv) Two experienced teachers have checked my items and corrected some of them.

(v) Random sampling of teachers has not been performed. However, their pupils can be seen to be representatives of each grade because of the Finnish comprehensive school system.

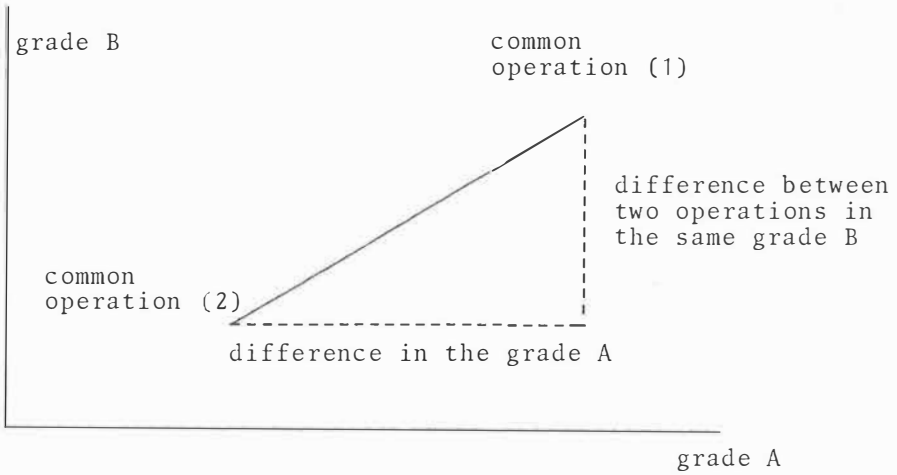
5.8.5. Linking based on basic parameters

Common basic parametets to each grade are:
- verbal problem or inequation
- tendency to solve the item incorrectly by using "short cut".

They seem to have the same significance for each grade. They can be used in linking in the sense of common operations. Values of η's are on an absolute scale, and the constant (a, b, c, d, e, f) gives the flexibility which is needed for linking. The two common operations are used in order to find values for constants a,..., f. The method used is illustrated in Table 35 and Figure 37.

Table 35. η-estimates for each grade (largest given first)

1.	2.	3.	4.	5.	6.
$\hat{\eta}_5$=.82 = 82a	$\hat{\eta}_7$= 57b	$\hat{\eta}_8$= 142c	$\hat{\eta}_8$= 135d	$\hat{\eta}_8$= 207e	$\hat{\eta}_8$= 120f
$\hat{\eta}_7$=.66 = 66a	$\hat{\eta}_1$= 45b	$\hat{\eta}_5$= 87c	$\hat{\eta}_5$= 90d	$\hat{\eta}_1$= 93e	$\hat{\eta}_7$= 94f
$\hat{\eta}_4$=.56 = 56a	$\hat{\eta}_4$= 35b	$\hat{\eta}_3$= 55c	$\hat{\eta}_7$= 75d	$\hat{\eta}_9$= 80e	$\hat{\eta}_2$= 67f
$\hat{\eta}_6$=.48 = 48a	$\hat{\eta}_3$= 31b	$\hat{\eta}_1$= 43c	$\hat{\eta}_3$= 35d	$\hat{\eta}_5$= 71e	$\hat{\eta}_9$= 67f
$\hat{\eta}_3$=.20 = 20a	$\hat{\eta}_6$= 19b	$\hat{\eta}_7$= 31c	$\hat{\eta}_2$= 31d	$\hat{\eta}_7$= 62e	$\hat{\eta}_5$= 63f
$\hat{\eta}_8$=.17 = 17a	$\hat{\eta}_8$= 16b	$\hat{\eta}_4$= 22c	$\hat{\eta}_1$= 2d	$\hat{\eta}_2$= 35e	$\hat{\eta}_3$= 53f
$\hat{\eta}_2$=.087= 9a	$\hat{\eta}_2$= -5b	$\hat{\eta}_6$= -6c	$\hat{\eta}_4$= -6d	$\hat{\eta}_4$= 21e	$\hat{\eta}_1$= 13f
$\hat{\eta}_1$=.018= 2a	$\hat{\eta}_5$= -8b	$\hat{\eta}_2$= -12c	$\hat{\eta}_6$= -8d	$\hat{\eta}_3$= 14e	$\hat{\eta}_4$= -12f
				$\hat{\eta}_6$= -42e	$\hat{\eta}_6$= -15f



$$k_{12} = \frac{57b - 16b}{66a - 17a} = \frac{41b}{49a} = .84\ \frac{b}{a} = 1.0$$

$$k_{23} = \frac{87c + 6c}{41b} = 2.27\ \frac{c}{d} = 1.0$$

$$k_{34} = \frac{90d + 8d}{93c} = 1.05\ \frac{d}{c} = 1.0$$

$$k_{45} = \frac{71e + 42e}{98d} = 1.15\ \frac{e}{d} = 1.0$$

$$k_{56} = \frac{63f + 15f}{113e} = .69\ \frac{f}{e} = 1.0$$

Figure 37. The idea of linking basic parameters

One constant can be chosen arbitrarily. We have given a the initial value of 0.5. The final linking matrix on n's will be:

Table 36. η-matrix for linking

1.	2.	3.	4.	5.	6.
41	34	37	34	45	37
33	27	23	23	20	29
28	21	14	19	17	21
24	18	11	8.8	15	21
10	11	8.1	7.8	13	20
8.5	9.5	5.7	0.5	7.5	16
4.5	-3.0	-1.6	-1.5	4.5	4.0
1.0	-4.8	-3.1	-2.0	3.0	-3.7
				-9.0	-4.6

For each grade the difference between maximum and minimum estimates is about 40 units. It can be seen from the table that operations η_7 and η_8 (in the first grade) on which the linking was based (underlined) become easier quite consistently from grade to grade. The operation η_8 (1st grade) is relatively easy from the 3rd grade onwards: that is to say that pupils have learned the idea of equation and they do not combine given numbers randomly. The operation "verbal problem or inequation" is rather difficult all the time. The only operation which is even harder is division. It is the hardest operation in the 3rd, 4th, 5th and 6th grades.

5.8.6. Summary of empirical results concerning basic operations

Interpretation of η-estimates gives the opportunity of comparing difficulties of each operation. It become clear that a new domain of numbers is a new difficulty for pupils. In the case of division this can be seen very clearly. Knowing division of positive whole numbers cannot be expected to help very much in learning division of decimals, fractions and integers.

The difficulty of problem solving cannot be analysed any more using this data because the items consisting of problem solving are very few and represent many different kinds of problems. For the first and second graders the main difficulty is likely to be in reading ability and particularly in understanding the text. It would be interesting to study more closely difficulties in problem solving. In other words it is a question of process of drawing conclusions in mathematical tasks. Especially, the logic of conclusions in each grade is an interesting problem and it could be studied using latent trait models. The levels of logical thinking could form a rating scale and estimates of thresholds would give information from probability to reach the next stage in thinking. Influence of teaching could be studied as well in comparing the estimates of threshold parameters and difficulty parameters before and after teaching period. The LLTM might also give useful information about absolute difficulties of each level of logical thinking.

Many errors in the area of negative whole numbers do not necessarily mean that these things are too difficult to learn in primary grades. It may be question of different opinions among teachers about the importance of this topic.

The analysis of errors and difficulties of the operations give teachers and curriculum planners valuable knowledge in the area of equations. First of all it is beneficial when pointing out that there are many interesting details which can be detected in analysing structures of learning tasks. The mere consciousness of the structural thinking in planning teaching and evaluating gives a new perspective to the subject.

5.9. Evalution of the properties of the LLTM as a mathematical model

Returning to the points of Cox and Hinkley (1976) mentioned at the beginning it can be seen from the previous discussion that

(i) Relationship between LLTM and other Rasch models is very clear. In constructing items for LLTM one must keep in mind the basic characteristics of a good Rasch test (Wright & Stone) and the main principles of structural learning (Scandura 1979) in addition to the experience which has been receiced from experimental studies, made mostly in German. Applications have shown that LLTM can be successfully used in different areas of psychology (e.g. Rop 1977) and education (e.g. Fischer 1972, 1973).

(ii) Analysis of log odds and comparison of the probabilities of solving the whole item or operations separately reveals that also in the case of LLTM extreme cases can be analyzed without any extra difficulties. Naturally the limitations of SLM can be taken into consideration for eliminating minimal and maximal raw scores.

(iii) Each elementary parameter has got an empirical counterpart. Estimated values of basic parameters have been interpreted as being difficulties of the corresponding tasks. Estimates give information about the learning process when the data consists of achievement tests in a relatively compact content area.

(iv) Only the most essential parameters are included in the final analysis of misfit. The first reduction in the number on parameters has been made by selecting LLTM instead of SLM to be the basis for statistical analysis in the structure of items. Instead of 30 item parameters only 8 or 9 basic parameters have been needed for each grade. In the second reduction which has been made for interpreting misfit some of the basic parameters have been combined because their estimates are close to each other. The final collection of operational parameters consists only of 5 elements for each test. They are the most essential operations and they are enough for finding misfitting items.

(v) The statistical theory behind the model should be relatively simple. In the case on LLTM it is based on the theory of other Rasch models. It may sound too idealistic to call it simple: however it is logical and concepts are so general that models will be useful in many different fields.

The main desired properties for a good mathematical model seem to occur in the case of LLTM. In this study the model has been used for clarifying the cognitive structure of mathematics tasks. Many other variants of the model can be used for measuring different behaviour and change in it (Kempf & Repp 1977). The main common feature in applications of LLTM is that they are always aimed at finding as simple a structure as possible for human reasoning. Also in physics the idea of simplicity and the use of probabilistic models are in use (Laurikainen 1973): "If we have two theories, which are equally precise for describing the same observations, we will choose the simpler theory. It is worth nothing that simplicity is not synonymous with popularity. Logical brightness and simplicity does not always mean popularity." Usually a very small number of psychologically elementary operations is needed for explaining the structure of relatively simple Mathematics and Phsysics tasks (Scheiblechner 1972; Fischer 1973; Spada 1977). In the critical discussion of the LLTM, Scheiblechner (1975) has paid particular attention to two reasons:

(1) The model is constructed of two parts: the "Rasch-part" and the "linear part". The criticism is aimed at the linear part of the model. One cannot be sure that the system of elementary parameters is item-free.

(2) Selection of elementary parameters does not usually take psychological aspects into account.

A way of overcoming these problems is to use the LLTM for analyzing only a specific group of items, not the whole item population that can be considered. From this special item group, LLTM may give relevant and interesting knowledge.

5.10. Empirical results of the LLTM from the point of view of curriculum and learning

In order to develop teaching it is necessary to evalute curriculum. Wilhelms (1971) gives five criteria for success in evaluation:

(1) Evaluation must facilitate self-evaluation. The learner needs to know what extent he has not yet achieved mastery, he needs to know what the gaps are, so that he can figure out what to do about them. It is especially important for the learner to learn about his strengths and resources, in a way that genuinely leads him to incorporate these into his self-concept.

(2) Evaluation must encompass every objective valued by the school. The best guide to curriculum improvement is evaluation; and to be an adequate guide the system of evaluation has to be as large as the purposes of the curriculum development itself.

(3) Evaluation must facilitate learning and teaching. Instructional diagnosis lies at the very heart of good teaching. After each bit of evaluative data comes in, the teacher should be a little surer of how to proceed next. A significant function of evaluation is a constant probing for the best way to move forward. The diagnosis should be so apparent in the evaluative devices used that the student will see diagnosis as an aid, not as a trap set to catch him in failure.

(4) Evaluation must produce records appropriate to the purposes for which records are essential. The records are used in a variety of ways and there probably needs to be a variety of kinds. In every case the essential thing is that the records be able to say what really counts and say it in a way that genuinely communicates.

(5) Evaluation must provide continuing feedback into the larger questions of curriculum development and educational policy. The school is full of such situations, where even the most precise measurement along one line

> may leave untouched larger questions of ultimate total effect. It is important to emphasize that evaluation must concern itself with all the important objectives of a school rather than only a narrow band out of the total spectrum. Many decisions are usually made by curriculum committees, administrators, and boards of education. This raises the problem of organizing data and channelling them so that they will be available to the persons or groups that actually make the choices. Since these are among the most important decisions made in any school, it is obviously crucial that they be based on genuinely evaluative feedback.

The main idea of Wilhelms seems to be the flexible use of the information from schools to decision making and developing the curriculum. In Finland we need more co-operation between researchers, school boards and teachers in the area of curriculum evaluation. We need more scientific knowledge for organizing more successful teaching and learning. Not only theories but also experimental knowledge from schools for continuous evaluation process of curriculum is needed. All the evaluation is aimed at the best of the pupils. However, it could make work of teachers easier if they were able to consult with researchers and sometimes get their own tests analysed and results interpreted in terms of elementary parameters.

Latent trait theory would be useful also in the broader evaluation of curriculum. Not only achievements but also attitudes can be studied using rating model and NEWRATE-program. The next step in this study could probably be to analyse motivation of pupils to learn mathematics and to try to find out what are the reasons connected to low motivation. If the reasons are basically related to the contents of curriculum it is necessary to study the areas in question more closely and to find more effective teaching methods and materials or to change objectives if needed.

The evaluation process should be continuous. Information about learning difficulties in each content area should be collected quickly and used as soon as possible for overcoming more serious repercussions and lack of school motivation. This study gives one model for collecting feedback and analysing it. The whole study is based on probabilities. Even in error analysis we can imagine that some errors are more probable than others. By the time pupils have finished making errors in some items, their errors have become very similar. Only some types of common errors can be detected. This can be considered to be a sign of learning. In the SLM the probability of giving the correct response to an item is a function of the person´s ability and the difficulty of the particular item. Having combined the results of error analysis we can compute difficulty estimates for each elementary parameter (representing the components of items which remain difficult in a certain grade) by means of the LLTM.

Most of the difficulties in primary grades are connected to the expansion in the domain of numbers. Particularly division does not become easier from grade to grade because of a new domain of numbers. Difficulties with negative integers cause errors in very many items. It seems that pupils try to remember the rules for calculations by heart without any deeper understanding of concept of numbers. It usually takes a couple of years before pupils have really understood the idea of a new calculation (Malinen 1980). More detailed studies about level of understanding could be done by interviewing pupils and asking them to explain what they were thinking when they made their mistakes. Presuming that some common levels for thinking process could be fixed, it would be possible to study how easy or difficult it will be to move from one level to the next. The suitable model for analysing threshold parameters is the rating model mentioned before.

As a summary of the whole study, Figure 38 gives the stages of the analysis as well as the aims for which the stage in question has been used. As a result of linking η´s

we have got the difficulty order of elementary parameter estimates. Seven of the most difficult things are:

(1) Division in the 5th grade.
(2) Many big numbers (≧ 10) in the same item in the 1st grade.
(3) Division in the 3th grade.
(4) Division in the 6th grade.
(5) Problem or inequation in the 2nd grade.
(6) Division in the 4th grade.
(7) Problem or inequation in the 1st grade.

All division parameters are included in these most difficult things. In addition, problem type items in the beginning of the learning process belong to the same group. The difficulty order of η-estimates in mentioned in Table 36.

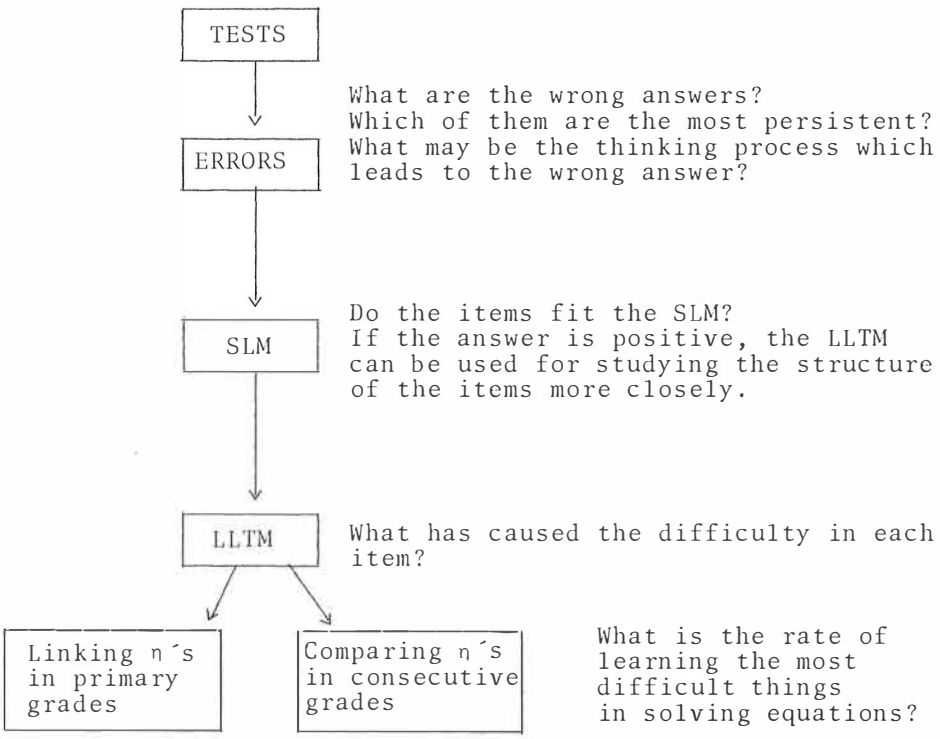


Figure 38. Stages in analysing tests of equations in this study.

Table 36. The difficulty order of η-estimates for primary grades after linking

1st grade $\hat{\eta} \rightarrow$ order	2nd grade $\hat{\eta} \rightarrow$ order	3rd grade $\hat{\eta} \rightarrow$ order	4th grade $\hat{\eta} \rightarrow$ order	5th grade $\hat{\eta} \rightarrow$ order	6th grade $\hat{\eta} \rightarrow$ order
$\eta_5 \rightarrow$ 2.	$\eta_7 \rightarrow$ 5.5	$\eta_8 \rightarrow$ 3.5	$\eta_8 \rightarrow$ 5.5	$\eta_8 \rightarrow$ 1.	$\eta_8 \rightarrow$ 3.5
$\eta_7 \rightarrow$ 7.	$\eta_1 \rightarrow$ 10.	$\eta_5 \rightarrow$12.5	$\eta_5 \rightarrow$12.5	$\eta_1 \rightarrow$17.5	$\eta_7 \rightarrow$ 8.
$\eta_4 \rightarrow$ 9.	$\eta_4 \rightarrow$ 15.	$\eta_3 \rightarrow$ 24.	$\eta_7 \rightarrow$ 19.	$\eta_9 \rightarrow$ 21.	$\eta_2 \rightarrow$ 15.
$\eta_6 \rightarrow$ 11.	$\eta_3 \rightarrow$ 20.	$\eta_1 \rightarrow$26.5	$\eta_3 \rightarrow$ 30.	$\eta_5 \rightarrow$ 23.	$\eta_9 \rightarrow$ 15.
$\eta_3 \rightarrow$ 28.	$\eta_6 \rightarrow$26.5	$\eta_7 \rightarrow$ 32.	$\eta_2 \rightarrow$ 33.	$\eta_7 \rightarrow$ 25.	$\eta_5 \rightarrow$17.5.
$\eta_8 \rightarrow$ 31.	$\eta_8 \rightarrow$ 29.	$\eta_4 \rightarrow$ 35.	$\eta_1 \rightarrow$ 41.	$\eta_2 \rightarrow$ 34.	$\eta_3 \rightarrow$ 22.
$\eta_2 \rightarrow$36.5	$\eta_2 \rightarrow$ 45.	$\eta_6 \rightarrow$ 43.	$\eta_6 \rightarrow$ 42.	$\eta_4 \rightarrow$36.5	$\eta_1 \rightarrow$ 38.
$\eta_1 \rightarrow$ 40.	$\eta_5 \rightarrow$ 49.	$\eta_2 \rightarrow$ 46.	$\eta_8 \rightarrow$ 44.	$\eta_3 \rightarrow$ 39.	$\eta_4 \rightarrow$ 47.
				$\eta_6 \rightarrow$ 50.	$\eta_6 \rightarrow$ 48.

The greater the accuracy a science attains, the greater the need for mathematical models. For example social sciences, education and linguistics have started to use mathematical models which traditionally have satisfied only needs on natural sciences. This study is one example of applying mathematical models to ordinary set of tests measuring achievements. It (hopefully) encourages researchers to use probabilistic models more extensively than they have usually been applied. Even if we must be critical when changing human characteristics in to numbers, it makes sense to analyse more closely the information which teachers collect in any case, trying to find as much information as possible from the results of achievement tests concerning on one hand pupils and on the other hand tests. The information given by the LLTM is useful for everyday teaching situations. If teachers are aware of the difficulties, they will probably have more motivation for finding better methods for difficult areas and positive feedback from learning is also more

satisfying. Objectivity of measurement is the main reason for using logistic test models. All the time teachers must give marks and grades. If we have some methods for getting the process of marking more accurate, why should we not use them?

6. DISCUSSION

6.1. General viewpoints

The latent trait theory was used in approaching the main purpose of the study: to find the qualities of the items which cause learning difficulties. The question "Why does a pupil not succeed in an item?" could have been studied also in many other ways. In the present study we wanted to concentrate on the structure of the task, knowledge of it is important for example in preparing learning matherials. When it is clear what the difficult points are it is possible either to construct easier items or to teach with sufficient emphasis the point were errors were made. If it is reasonable to change the curriculum, the innovation can be based on the structural analysis mentioned above.

The latent trait theories do not imply random sampling from subjects, they are sample-free. It was more important that teachers in the study were interested in performing the task as well as possible and that teachers could motivate their pupils to do the work seriously. Their classes were, however, normal classes in comprehensive school and the pupils can be considered to represent their age group well enough if one bears this in mind when interpreting the results from the tests.

The data was composed of the tests of equations and inequations which were constructed for the primary and junior secondary level of the comprehensive school. There were three parts in the study. The first was error analysis concerning tests of the primary level of comprehensive school. The errors were classidied according to their commonness and a basis for the later choose of elementary parameters were sought. In the second part the simple logistic model was applied and simultaneously the structure of the model was examined from the statistical viewpoint. In the third part the data was treated using the linear logistic model; also this model was analysed. The advantage of this model was that it was possible to find out the learning difficulties in primary school.

So the error analysis was done primarily in order to get an empirical basis for the use of the linear logistic test model. One purpose of applying the simple logistic model was to make sure that items fitted this model and the linear logistic model could be applied later. In these circumstances the two main parts at the beginning of the study were needed for performing the third part. That is why the major research problems are in the latter part of the report.

6.2. Viewing the results

The Mathematics syllabus in the comprehensive school has been changed as this study was being done. For example, part of the four basic computations on integers has been moved to the seventh grade. Also this study showed that negative integers have not yet been mastered by the sixth grade. Nevertheless the conslusion need not necessarily have been to move the main part of the area of negative integers to upper grades. The other alternative could have been to present these things in two grades so that complete mastering of multiplication and dicision were not expected in primary level.

The most difficult calculation turned out to be division. This can be understood because long division includes also subtraction and multiplication. When a pupil starts to learn division, the domain of number becomes larger and for example long division does not benefit division of fractions very much. Also in other calculations the same feature can be seen: when we look at the achievements they tend to become worse when the pupil goes to the upper grade, this happens because the domain of numbers has became larger. The way of teaching calculations in connection with the new domain of numbers is worthy of consideration. If learning is based on rules learned by heart, the collection of rules starts to be so big by sixth grade that it is probable the new rules be confused with the old ones.

Also the problem solving tasks remained difficult in all tests, as may be expected. The questions of problem-solving have been under consideration in journals of Mathematics teaching. Didactics of problem-solving obviously needs more research. When the "new Mathematics" was in vogue the number of problem-solving tasks in textbooks decreased. Nowadays their number has increased and teaching of things like problem-solving is considered to be essential in developing mathematical thinking.

Part of the results is associated with the theoretical qualities of the statistical models used. Some of them are associated with the test statistics measuring items and person fit and comparing information functions, some others are associated with the operational structure of items and with linking of tests in the linear logistic test model using estimates of elementary parameters. It is not possible to get information on the benefits and weaknesses of the theoretical portion of this study until researchers in this field have got acquainted with them.

6.3. Suggestions for further research

A further study has already been started in Jyväskylä University in the Department of Teacher Education in a project of student teachers in the area of Mathematics education. They will for example be applying the latent trait theory to the research problems concentrating on developing problem-solving, using mastery learning and developing verbal items in such manner that they become interesting.

The rating model could be a good tool in studying attitudes of pupils to Mathematics especially in the situations when the difficulty of the task is at the upper limit of their capacity of performing.

It would be interesting to make the present study more complete by interviewing pupils. It could be possible to find immediately the reasons for lack of success in some tasks.

The study could be extended to include areas of school Mathematics other than only equations and inequations; specially problem-solving could be an area of its own. In this way we could come to a better understanding of learning difficulties in Mathematics and it would be possible to apply remendial teaching effectively in advance. Valuable knowledge could be got for developing the curriculum further.

SUMMARY

The aim of this thesis was to apply logistic test models to tests in Mathematics for the purpose of analysing learning difficulties. At the same time the aim was to come to a deeper understanding of the structures of the models used and to develop the characteristics of the models from the point of view of theory. Since this is an area that has not been studied for any period longer than the last few decades, the theory behind it is also still in a state of flux.

The thesis may be divided into three parts, each dealing with a different aspect of the tests. In the first part, all errors made by the pupils are analysed, a flow chart being made of the most prevalent errors. Pupils were found to make a large number of error types at the outset of tuition. This high error frequency persisted at later stages as well in the case of weaker pupils. At later stages of instruction, with more learning having occured, errors tend to decrease, both from the quantitative point of view and from the point of view of the number of types of error made.

The second part concentrates on the application on the simple logistic model to the corpus. In addition, the results were compared to results obtained by traditional test theory. Initially items were tested for sompatibility with the model. In this test the two test statistics T_1 and T_2, presented in the NEWRATE programme were compared with each other, It was found that the test statistic obtained

by logarithmic transformation was the better alternative for its symmetricity of skewness and kurtosis. In addition, T_3, a new test statistic was developed, with characteristics quite similar to those of the test statistic T_2.

A formula was developed for the variance of the residual z^2_{vi}, a formula which clearly demonstrated the relationship between the variance and the term $\beta_v - \delta_i$. Suitability of testees to the model was examined briefly, using the same methods as for the test items. Reasons for misfit were examined separately, firstly for items and secondly for persons. The biserial correlation coefficient and the test statistic T_2 were compared to each other in order to measure the goodness of the test items.

Information provided by the item and by the test was approached by comparing Birnbaum and Fisher´s consepts to the information-theoretic approach. Information functions were drawn for each test on the basis of the empirical material. In addition a method was presented for combining tests, once it has been ascertained that both consecutive tests have a sufficient number of items in common.

The third part used the linear logistic test model to examine the structure of the items. As a result, the operations were arrived at which were difficult for pupils, and failure in which resulted in the wrong answer for the whole item. An estimate for the difficulty parameter was calculated for each operation with its confidence limits. The consecutive combination of tests was based on these parameters of difficulty. It was considered essential to standardize the difficulty parameters of the items before application of the conformity test between the item difficulties of the simple logistic model and the linear logistic model respectively. Theoretical treatment was given, amongst other things, to the question under what conditions completion of the whole task was easier than the error-free completion of the individual test operations.

TIIVISTELMÄ

Tämän tutkimuksen tavoitteena oli soveltaa logistisia testimalleja matematiikan testeihin oppimisvaikeuksien analysoimiseksi. Samalla oli tarkoitus perehtyä käytettyjen mallien rakenteeseen ja kehittää myös mallien ominaisuuksia teoreettisista lähtökohdista käsin. Koska kyseessä on vasta parikymmentä vuotta tutkittu alue, on teoriakin vielä kehittelyn alaisena.

Tutkimus voidaan jakaa kolmeen osaan, joista jokainen käsitteli testejä eri näkökulmasta. Ensimmäisessä osassa kaikki oppilaiden tekemät virheet analysoitiin ja muodostettiin yleisimmistä virheistä kulkukaavio. Havaittiin, että oppimisvaiheen alussa (ja heikoilla laskijoilla myöhemminkin) virhetyyppejä esiintyi runsaasti. Myöhemmin, kun oppimista on enemmän tapahtunut, virheillä on taipumusta sekä vähentyä että samalla luokittua vain muutamaksi virhetyypiksi.

Toisessa osassa keskityttiin Raschin perusmallin soveltamiseen tutkimusaineistoon. Sen lisäksi verrattiin tuloksia traditionaalisen testiteorian antamiin tuloksiin. Aluksi analysoitiin osioiden sopivuutta malliin. Siinä verrattiin NEWRATE-ohjelmassa esitettyjä kahta testisuuretta T_1 ja T_2 toisiinsa, todettiin logaritmimuunnoksella tehdyn testisuureen paremmuus sen jakauman vinous- ja huipukkuustulosten symmetrisyyden perusteella. Kehitettiin lisäksi uusi testisuure T_3, joka noudattelee hyvin paljon testisuureen T_2 piirteitä.

Jäännöstermin z^2_{vi} varianssille kehiteltiin lauseke, josta selvästi näkyy varianssin riippuvuus erotuksesta $\beta_v - \delta_i$. Henkilöiden sopivuutta malliin tarkasteltiin lyhyesti samoilla menetelmillä kuin osioiden. Syitä yhteensopimattomuuteen pohdittiin erikseen osioiden ja henkilöiden tapauksissa. Osioiden hyvyyden mittoina verrattiin biseriaalista korrelaatiokerrointa ja testisuuretta T_2 toisiinsa.

Osion ja testin informaatiota lähestyttiin vertaamalla Birnbaumin ja Fisherin esittämiä käsitteitä sekä informaatioteoreettista tapaa toisiinsa. Empiirisen aineiston pohjalta piirrettiin informaatiofunktioiden kuvaajia kutakin testiä varten. Lisäksi esitettiin tapa, jolla testejä voidaan yhdistää toisiinsa, kun on huolehdittu siitä, että peräkkäisillä testeillä on riittävä määrä yhteisiä osioita.

Kolmannessa osassa käytettiin lineaarista logistista testimallia osioiden rakenteelliseen tarkasteluun. Sen avulla saatiin selville ne operaatiot, joiden suorittamisessa oppilailla oli vaikeuksia ja joissa epäonnistumisesta johtui väärä vastaus koko osioon. Kullekin operaatiolle laskettiin vaikeusparametrin estimaatti ja sille luottamusväli. Näihin vaikeusparametreihin perustettiin testien peräkkäinen yhdistäminen toisiinsa. Osioiden vaikeusparametrien standardoimista pidettiin välttämättömänä ennen yhteensopivuustestiä Raschin perusmallin ja lineaarisen logistisen mallin osiovaikeuksien välillä. Teoreettisena kysymyksenä tarkasteltiin mm. sitä, millä ehdolla koko tehtävästä suoriutuminen on helpompaa kuin sen sisältämien yksittäisten osatehtävien suorittaminen virheettömästi.

BIBLIOGRAPHY

Andersen, E.B. (1977) Sufficient statistic and latent trait models. Psychometrica, 42, 69-81.

Andrich, D. (1975) The Rasch multiplicative binomial model: applications to attitude data. Research Report number 1, Measurement and Statistics Laboratory, Department of Education. The University of Western Australia.

Andrich, D. (1978) A raiting formulation for ordered response categories. Psychometrica, 43, 561-573.

Andrich, D. & Godfrey, J.R. (1977) Hierarchies in the skills of Davis´ reading comprehension test form D: an empirical investigation. Research Report number 3. Measurement and Statistics Laboratory. Department of Education. The University of Western Australia.

Andrich, D. & Koponen, R.L. (1980) Consistency of ratings in the selection of students into a teacher education program in Finland. Research Report number 8. Measurement and Statistics Laboratory. Department of Education. The University of Western Australia.

Andrich, D. & Sheridan, B.E. (1980) RATE: A Fortran IV program for analysing rated data according to a Rasch model. Research Report number 5. Measurement and Statistics Laboratory. Department of Education. The University of Western Australia.

Arter, J.A. & Clinton, R. (1974) Time and error consequences of irrelevant data and question placement in arithmetic word problems II: fourth graders. The Journal of Educational Research, 68, 28-31.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee´s ability. In F.M. Lord & M. Novick: Statistical theories of mental test scores. Reading, Mass.: Addison Wesley, 397-479.

Caldwell, J.H. & Goldin, G. A. (1979) Variables affecting word problem difficulty in elementary school mathematics. Journal for Research in Mathematics Education, 10, 323-336.

Casey, D. P. (1978) Failing students: a strategy of error analysis. In Castello (ed.) Aspects of motivation, Melbourne: Mathematical Association of Victoria, 295-306.

Cifarelli, V.V. & Wheatley, G.H. (1979) Formal thinking strategies: A prerequisite for learning basic facts? Journal for Research in Mathematics Education, 10.

Clements, M.A. (1979) Analyzing childrens errors on written mathematical tasks. Unpublished.

Cox, D.R. (1970) The analysis of binary data. London: Methuen.

Cox, L.S. (1975) Systematic errors in the four vertical algorithms in normal and handicapped population. Journal for Research in Mathematics Education 6, 202-220.

Cox, D.R. & Hinkely, D.V. (1974) Theoretical statistics. London: Chapman and Hall.

Dienes, Z. (1979) A round table discussion. How does learning take place? Journal of Structural Learning, 6, 97-114.

Douglas, G.A. (1980) Conditional inference in a generic Rasch model. Melbourne: Invitational Seminar on the Improvement of Measurement in Education and Psychology. Unpublished paper.

Fischer, G.H. (1972) A measurement model for the effect of mass-media. Acta Psychologica, 36, 207-220.

Fischer, G.H. (1973) The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Fischer, G.H. (1974) Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen. Bern: Verlag Hans Huber.

Fischer, G.H. (1976) Some probabilistic models for measuring change. In de Gruijter & van der Kamp (Ed.) Advances in Psychological and Educational Measurement, London: Wiley & Sons, 97-110.

Fischer, G.H. (1977a) Linear logistic test models: Theory and applications. In H. Spada & W.F. Kempf. (Eds.) Structural models of thinking and learning. Bern: Hans Huber Publishers.

Fischer, G.H. (1977b) Some probabilistic models for the description of attitudinal and behavioral changes under the influence of mass communication. In W.F. Kempf & B.H. Repp (Eds.) Mathematical models for social psychology, 102-151.

Fisher, R.A. (1959) Statistical methods and scientific inference. Second edition. London: Oliver and Boyd.

Gerholm, T.R. (1977) The meaning of scientific objectivity. Danish Yearbook of Philosophy, 14, 97-105.

Gustafsson, J-E. (1977) The Rasch model for dichotomous items: theory, applications and computer program. Reports from the Institute of Education. University of Göteborg. Number 63.

Gustafsson, J-E. (1979) Testing and obtaining fit of data to the Rasch model. Reports from the Institute of Education. University of Göteborg. Number 83.

Gustafsson, J-E. & Lindblad, T. (1978) The Rasch model for dichotomous items: A solution of the conditional estimation problem for long tests and some thoughts about item screening procedures. Reports from the Institute of Education. University of Göteborg. Number 67.

Guttman, L. (1954) A new approach to factor analysis: The Radex. In Lazarsfeld (Ed.) Mathematical thinking in the social sciences.

Hambleton, R.K. & Cook. L.L. (1977) Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R.K. & Traub, R. E. (1973) Analysis of empirical data using two logistic latent trait models. Br. J. Math. statist. Psychol., 26, 195-211.

Hendrickson, A.D. (1979) An inventory of mathematical thinking done by incoming first-grade children. Journal for Research in Mathematics Education, 10, 7-23.

Hilke, R., Kempf, W. & Scandura, J. (1977) Deterministic and probabilistic theorizing in structural learning. In H. Spada & W.F. Kempf (Eds.) Structural models of thinking and learning. 415-436.

Hogg, R.V. & Craig, A. T. (1978) Introduction to mathematical statistics. Fourth edition. New York: Macmillan.

Hollander, S.K. (1978) A literature review: Thought processes employed in the solution of verbal arithmetic problems. School Science and Mathematics, 4, 327-335.

Häussler, P. (1978) Evaluation of two teaching programs based on structural learning principles. Studies in Educational Evaluation, 4, 145-161.

Izard, J.F. & White, J.D. (1980) The use of latent trait models in the development and analysis of classroom tests. Melbourne: Invitational Seminar on the Improvement of Measurement in Education and Psychology. Unpublished paper.

Jenks, S. & Peck, D. (1976) Missing addend problems. School Science and Mathematics, 76, 647-661.

Kaila, M. (1971) Aritmeettisten ja algebran alkeiden virheiden analysoimista. Laudaturtyö kasvatustieteessä. Helsingin yliopisto.

Kempf, W.F. (1975) A test-teoretical approach to structural learning. Kiel: IPN, reprint 14.

Kempf, W.F. (1977a) Dynamic models for the measurement of "traits" in social behavior. In Kempf & Repp (Eds.) Mathematical models for social psychology, 14-58.

Kempf, W.F. (1977b) A dynamic test model and its use in the microevaluation of instructional material. In Spada & Kempf (Eds.) Structural models of thinking and learning, 295-318.

Kempf, W.F. & Repp,B.H. (1977) (Eds.) Mathematical models for social psychology. Bern: Hans Huber.

Knifong, J.D. & Holtan, B. (1976) An analysis of children´s written solutions to word problems. Journal for Research in Mathematics Education, 7, 106-112.

Konttinen, R. (1979) Mitä uutta moderni testiteoria antaa kasvatustieteelliselle tutkimukselle? Kasvatus, 10, 389-393.

Konttinen, R. (1981) Testiteoria. Johdatus kasvatus- ja käyttäytymistieteellisen mittauksen teoriaan. Helsinki: Gaudeamus.

Konttinen, R. & Kortelainen, S. (1979) Osioanalyysi OSANA-ohjelmalla. Jyväskylän yliopisto. Laskentakeskuksen tiedonantoja no. 2.

Koponen, R. (1976) Peruskoulun yläasteen matematiikan opetuksen sisältö- ja tavoiteanalyysi. Jyväskylän yliopisto. Opettajankoulutuslaitoksen julkaisuja 3.

Koskenniemi, M. (1968) Opetuksen teorian perusaineksia. Helsinki: Otava.

Krutetskii, V. A. (1976) The psychology of mathematical abilities in schoolchildren. Chicago: The University of Chicago Press.

Lahti, A. (1949) Virheet opetusopillisena ongelmana. Kasvatustieteellinen kirjasto no. 4. Kuopio: Suomen kasvatusopillinen yhdistys.

Lankford, F.G. (1974) What can teacher learn about pupil´s thinking through oral interviews? The Arithmetic Teacher, 21, 26-32.

Laurikainen, K.V. (1973) Atomistiikan aatemaailma ja sen heijastumia aikamme ideologioissa. Helsinki: WSOY.

Lazarsfeld, P.F. (1954) (Ed.) Mathematical thinking in the social sciences. Illinois: The Free Press.

Leino, J. (1978) Matemaattisten kykyjen ja ajatteluprosessien kehittäminen kouluopetuksessa 2. Helsingin Yliopiston Kasvatustieteen laitoksen julkaisuja 66/1978.

Lindvall, C.M. & Ibarra, C.G. (1980) Open addition and subtraction sentences. Journal for Research in Mathematics Education, 11, 50-62.

Linville, W.J. (1976) Syntax, vocabulary, and the verbal arithmetic problem. School Science and Mathematics, 76, 152-158.

Lord, F.M. (1977) Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lumsden, J. (1957) A factorial approach to unidimensionality. Australian Journal of Psychology, 9, 105-111.

Lumsden, J. (1961) The construction of unidimensional tests. Psychological Bulletin, 58, 122-131.

Lumsden, J. (1976) Test theory. Annual Review of Psychology, 27, 254-280.

Lumsden, J. (1977) Person reliability. Applied Psychological Measurement, 1, 477-482.

Lumsden, J. (1978) Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 31, 19-26.

Malinen, P. (1969) The learning of elementary algebra. An empirical investigation of the results of learning in a simplified school learning system. Research Bulletin. Institute of Education. University of Helsinki.

Malinen, P. (1974) Kasvatustieteen metodologia. Opetusmoniste 43. Kasvatustieteen laitos. Jyväskylän yliopisto.

Malinen, P. (1980) Matemaattisen ajattelun kehittyminen peruskoulun ala-asteen oppilailla. Opettajankoulutuslaitos. Jyväskylän yliopisto. Tutkimuksia 4.

Masters, G. (1980) A Rasch model for rating scales. University of Chicago, Department of Education. Unpublished doctoral dissertation.

Mc Donald, R. (1980) The dimensionality of tests and items. The Ontario Institute for Studies in Education. Unpublished paper.

Morgan, G. (1980) The use of the Rasch latent trait measurement model in the equating of scholastic aptitude tests. Melbourne: Invitational seminar of the Measurement in Education and Psychology. Unpublished paper.

Newman, M.A. (1977) An analysis of sixth-grade pupils´ errors on written mathematical tasks. Research in Mathematics education in Australia, 1, 239-258.

Niehusen, B., Mach, G., Hansen, K-H., Rost, J. & Kempf, W.F. (1978) Manual der IPN-Programmbibliothek, Band 2. IPN-Arbeitsberichte 24.

Ord, J.K. (1972) Families of frequency distributions. London: Griffin.

Pelka, R.B. (1975) Nicht-Markoff´sche stochastisce Lernprozesse in Moore-Automaten mit Konstanten und Differentiellen. In Tack (1975) Bericht über den 29. Kongress der Deutschen Gesellschaft für Psychologie in Salzburg 1974, 330-333.

Perline, R., Wright, B.D. & Wainer, H. (1979) The Rasch model as additive conjoint measurement. Applied Psychological Measurement, 3, 237-255.

Pincus, M. (1975) If you don´t know how children think, how can you help them? The Arithmetic Teacher, 22, 580-585.

Puro, J. (1974) Tukiopetuksessa tarvittavan materiaalin kehittäminen I. Kuvaus matematiikan tukiopetuksessa tarvittavan materiaalin kehittämisestä peruskoulun ala-asteella sekä joitakin tuloksia materiaalin ensimmäisistä käyttökokeiluista. KTL:n julkaisuja 217/1974.

Puro, J. (1977a) Tukiopetuksessa tarvittavan materiaalin kehittäminen II. Peruskoulun ala-asteelle kehitettyjen matematiikan tukiopetusmateriaalien käyttökelpoisuuden evaluointi. KTL:n julkaisuja 269/1977.

Puro, J. (1977b) Tukiopetuksessa tarvittavan materiaalin kehittäminen III. Peruskoulun ala-asteelle kehitettyjen matematiikan tukiopetusmateriaalien käyttökelpoisuuden evaluointi. Liiteraportti. Tutkimuksenmetodi ja laskentatulokset. KTL:n julkaisuja 270/1977.

Radatz, H. (1979) Error analysis in mathematics education. Journal for Research in Mathematics Education, 10, 163-172.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.

Rasch, G. (1968) A mathematical theory of objectivity and its consequences for model construction. Paper read at European Meeting on Statistics, Econometrics and Management Science, Amsterdam 2.-7. September.

Rasch, G. (1977) On specific objectivity an attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy, 14, 58-94.

Rasch, G. (1980) Probabilistic models for some intelligence and attainment tests. Second edition with a foreword and afterword by Benjamin D. Wright. Chicago: The University of Chicago Press.

Rathmell, E.C. (1979) A reply to "Formal Thinking Strategies: a prerequisite for learning basic facts?". Journal for Research in Mathematics Education, 10, 374-377.

Roberts, G.H. (1968) The failure strategies of third grade arithmetic pupils. The Arithmetic Teacher, 15, 442-446.

Rop, I. (1977) The applications of a linear logistic model describing the effects of pre-school curricula on cognitive growth. In H. Spada & W.F. Kempf (Eds.) Structural models in thinking and learning, 281-293.

Rosenthal, D.J.A. & Resnick, L.B. (1974) Children´s solution processes in arithmetic word problems. Journal of Educational Psychology, 66, 817-225.

Samejima, F. (1977) Weakly parallel tests in latent trait theory with some criticism of classical test theory. Psychometrica, 42, 193-198.

Scandura, J. (1979) Integrated systems of human behavior. Journal of Structural Learning, 6. 137-138.

Scheiblechner, H. (1972) Das Lernen und Lösen komplexer Denkaufgaben. Z. Exp. Angew. Psychol., 19, 476-506.

Scheiblechner, H. (1975) Kritik der Anwendung des linearen logistischen Models in der Psychologie. In Tack (Ed.) Bericht über den 29. Kongress der Deutschen Gesellschaft für Psychologie in Salzburg 1974, 324-326.

Spada, H. (1977) Logistic models of learning and thought. In H. Spada & W.F. Kempf (Eds.) Structural models of thinking and learning, 227-262.

Spada, H. (1980) The linear logistic test model and its application in educational evaluation. Melbourne: Invitational Seminar on the Improvement of Measurement in Education and Psychology. Unpublished paper.

Spada, H.F. & Fischer, G.H. (1973a) Die psychometrischen Grundlagen des Rorschach Tests und der Holtzman Inkblot Technique. Huber: Bern/Stuttgart.

Spada, H.F. & Fischer, G.H. (1973b) Latent trait models and the problem of measurement in projective techniques (Rorschach, Holtzman). Huber: Bern/Stuttgart.

Spada, H. & Kempf, W.F. (1977) Structural models of thinking and learning. Proceedings of the 7th IPN-Symposium on Formalized Theories of Thinking and Learning and their Implications for Science Instruction. Bern: Hans Huber Publishers.

Steffe, L.P. (1979) A reply to "Formal Thinking Strategies: a prerequisite for learning basic facts?". Journal for Research in Mathematics Education, 10, 370-373.

Tack, W.H. (1975) (Ed.) Bericht über den 29. Kongress der Deutschen Gesellschaft für Psychologie in Salzburg 1974, Band 1, Göttingen: Verlag für Psychologie. Dr. C.J. Hogrefe.

Wilhelms, F.T. (1971) Evaluation as feedback. In R. Hooper (Ed.) The Curriculum: Context, Design & Development. Edinburg: Oliver & Boyd, 320-335.

Wilmut, J. (1975) Objective test analysis: some criteria for item selection. Research in Education, No. 13, 27-56.

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

Wright, B.D. (1980) Afterword in the second edition of the Rasch´s book: Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

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

Wright, B.D. & Mead, R.J. (1977) BICAL: Calibrating items and scales with the Rasch model. Research Memorandum no. 23. Statistical Laboratory. Department of Education. The University of Chicago.

Wright, B.D. & Stone, M.H. (1979) Best test design: Rasch measurement. Chicago: MESA Press.

Young, J.F. (1971) Information theory. London: Butterworth.

WRITE THE MISSING NUMBER
IN THE SQUARE

1. $5 + 3 = \square$
2. $4 + \square = 9$
3. $\square + 6 = 9$
4. $7 + 9 = \square$
5. $\square + 8 = 15$
6. $9 + \square = 17$
7. $8 - 3 = \square$
8. $6 - \square = 2$
9. $\square - 4 = 3$
10. $\square - 4 = 16$
11. $17 - \square = 11$
12. $\square - 8 = 15$
13. $35 - \square = 24$
14.
$$\begin{array}{r} 1\ 4 \\ +\ \square\square \\ \hline 2\ 6 \end{array}$$

TEST 1.

NAME: _____

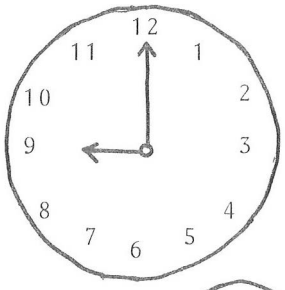
SCHOOL: _____

DATE: _____

DATE OF BIRTH: _____

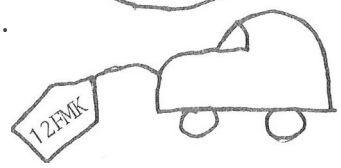
15.
$$\begin{array}{r} 2\ 8 \\ -\ \square\square \\ \hline 1\ 3 \end{array}$$
16. $\square < 7 - 5$
17. $\square > 3 + 1$
18. $\square - 7 < 2$
19. $3 + \square + 2 = 9$
20. $10 - 3 - \square = 7$
21. $64 + \square + 16 = 100$
22. $85 - 15 - \square = 5$
23.
$$\begin{array}{r} 2\ 4 \\ +\ 1\ 8 \\ \hline \square\square \end{array}$$
24.
$$\begin{array}{r} 4\ 8 \\ -\ 1\ 6 \\ \hline \square\square \end{array}$$

25. HOW MUCH TIME IS LEFT BEFORE IT WILL BE 12 O´CLOCK?



ANSWER: _____ HOURS

26. THE PRICE OF A CAR IS 14 FMK. IT WAS SOLD AT 12 FMK. CALCULATE THE LOSS.



ANSWER: _____ FMK

27. THERE ARE 8 PIECES IN EVERY ORANGE. HOW MANY PIECES ARE THERE IN THE 3 ORANGES?

ANSWER: _____ PIECES

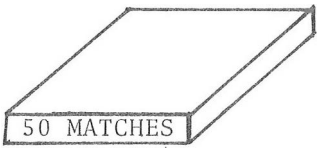
28. HOW MUCH CHANGE WILL JUSSI HAVE FROM 10 FMK IF HIS SHOPPING COST 7 FMK 50 P ?

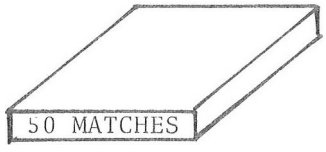
ANSWER: _____ FMK _____ P

29. MATTI HAS 24 FMK. THE PRICE OF AN ICE-HOCKEY STICK IS 43 FMK. HOW MUCH MORE DOES HE NEED?

ANSWER: _____ FMK

30. YOU HAVE GOT TWO BOXES FULL OF MATCHES. YOU TAKE 78 MATCHES AWAY. ONE OF THE BOXES WILL BECOME EMPTY. HOW MANY MATCHES ARE LEFT IN THE OTHER BOX?





WRITE THE MISSING NUMBER IN THE SQUARE

1. $23 + 12 = \square$

2. $\square + 54 = 60$

3. $35 + 17 = \square$

4. $\square\square + 12 = 73$

5. $24 + \square\square = \square\square$

6. $48 - 16 = \square\square$

7. $728 - 709 = \square\square\square$

8. $\square\square - \square = 13$

9. $18 + 14 = 14 + \square$

10. $25 + 17 + 5 = 17 + \square$

11. $\square \cdot 3 = 24$

12. $7 \cdot 3 = \square$

13. $7 \cdot 8 = \square$

14. $7 \cdot \square < 10$

15. $12 + 14 + \square < 27$

16. $4 \cdot 3 > \square \cdot 5$

TEST 2

NAME: _____

SCHOOL: _____

DATE: _____

DATE OF BIRTH: _____

17. Which of the following numbers can go in the square: 5, 6, 7, 8, 9, 10, 11?

$17 - \square > 10$

Answer: Numbers _____

18. $102 - 81 = \square\square\square$

19. $\square - 4 = 16$

20. $6 \cdot 4 - \square = 5$

21. $3 \cdot 124 = \square\square\square$

22. $789 + 977 = \square\square\square\square$

23. There are 8 pieces in every orange. How many pieces are there in the 3 oranges?

Answer: _____ pieces

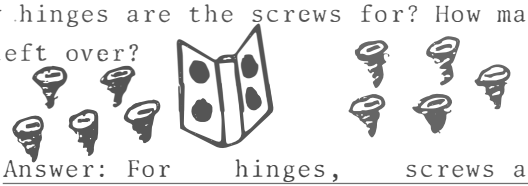
24. How many exercise books can you buy with 28 Fmk if each of them costs 4 Fmk?

Aswer: _____ exercise books

25. The petrol tank of a car has a capacity of 36 liters. How many liters were in the tank already if it takes 21 litres to fill it up?

Answer: _____ liters

26. For how many hinges are the screws for? How many screws are left over?

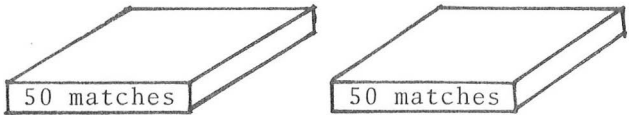


Answer: For _____ hinges, _____ screws are left over

27. Mother baked 54 buns. One baking tray takes 20 buns. How many trays did she need?

Answer: _____ trays

28. You have got two boxes full of matches. You take 78 matches away. One of the boxes will become empty. How many matches are left in the other box?



Answer: _____ matches

29. How much is left from 100 Fmk after having done the following shopping: a blouse 48 Fmk, a scard 15 Fmk and a knitted hat 15 Fmk?

Answer: _____ Fmk

30. You have 20 liters of juice. How many half litre bottles are needed for bottling?

Answer: _____ bottles

WRITE THE MISSING NUMBER
IN THE SQUARE

1. $\square \cdot 3 = 24$

2. $7 \cdot 8 = \square$

3.
$$\begin{array}{r} 1\,0\,2 \\ -\quad 8\,1 \\ \hline \square\,\square\,\square \end{array}$$

4. $3 \cdot 124 = \square$

FIND AT LEAST ONE NUMBER
WHICH CAN SUBSTITUTE FOR x

5. $3 + x = 15$ x = ____

6. $x - 4 = 16$ x = ____

7. $x \cdot 8 = 64$ x = ____

8. $x : 4 = 7$ x = ____

9. $72 : x = 8$ x = ____

10. $240 - x = 100$ x = ____

11. $\frac{x}{10} = 5$ x = ____

12. $x \cdot 3 + 7 \cdot 3 = 9 \cdot 3$ x = ____

13. $x \cdot 40 = 160$ x = ____

14. $214 + x = 400$ x = ____

15. $850 - x = 150$ x = ____

16. $x-(2400+3600)=4000$ x = ____

17. $20 - 2 + x = 20$ x = ____

18. $5 \cdot x < 14$ x = ____

19. $2 \cdot 2 + 8 \cdot x = 20$ x = ____

20. $3 \cdot 458 = x$ x = ____

21. $48 = 5 \cdot 8 + x$ x = ____

22. $3 \cdot 5 > x \cdot 6$ x = ____

TEST 3

NAME: _____

SCHOOL: _____

DATE: _____

DATE OF BIRTH: _____

GRADE IN MATHS IN YOUR
LAST REPORT: _____

23. $x : 6 + 2 < 4$ x = ____

24. $40 : x > 10$ x = ____

25. If the temperature rose 4^{o} higher, it would be -20^{o}. What is the temperature now?
Answer: _____

26. The total length of the sides of a triangle is 28 cm and there are two sides in the triangle which are 11 cm in length. What is the length of the third side?
Answer: _____

27. A quarter kilogram of coffee costs 8Fmk 20p. How much does 2 kg of this coffee cost?
Answer: _____

28. A season ticket for thirty bus trips costs 90Fmk. The price of one single trip is 3Fmk 20p. How much has been saved per trip if season tickets are used?
Answer: _____

29. You have 20 litres of juice. How many half litre bottles are needed for bottling?
Answer: _____

30. You buy an equal number of tickets for adults and for children. The price of an adult ticket is 12Fmk and a child´s ticket 6Fmk. How many tickets can you buy for 54Fmk?
Answer: _____

FIND AT LEAST ONE NUMBER
WHICH CAN SUBSTITUTE FOR x

1. $10 \cdot 14 = x$ x =____
2. $x - 4 = 16$ x =____
3. $x : 4 = 7$ x =____
4. $18 : x < 2$ x =____
5. $8700-3888 = x$ x =____
6. $x \cdot 8 = 88$ x =____
7. $214 + x = 400$ x =____
8. $x+28:4+3\cdot5 = 40$ x =____
9. $1,3 - x = 1,0$ x =____
10. $840 : 8 = x$ x =____
11. $0,8 + x = 8,0$ x =____
12. $(12+8)\cdot8-x = 105$ x =____
13. $(432-388):x = 4$ x =____
14. $3\cdot8 + x\cdot8 = 32$ x =____
15. $(300+60+x):3 = 123$ x =____
16. $3 \cdot (1+x) < 30$ x =____
17. $(1+2+3)\cdot5-18:3 = x$ x =____
18. $150:(x+3) = 150:5$ x =____
19. $2 \cdot x \cdot 5 = 400$ x =____
20. $0,7+1,8+x = 4,0$ x =____
21. $x-14 = 3\cdot(7+5)$ x =____
22. $2 \cdot x - 5 = 15$ x =____
23. $3\cdot450+x = 2000$ x =____
24. $168:12 = x$ x =____

TEST 4

NAME:_____

SCHOOL:_____

DATE:_____

DATE OF BIRTH:_____

GRADE IN MATHS IN YOUR
LAST REPORT:_____

25. You have 20 liters of juice. How many half litre bottles are needed for bottling?
Answer:_____

26. A season ticket for thirty bus trips costs 90 Fmk. The price of one single trip is 3Fmk 20p. How much has been saved per trip if seasontickets are used?
Answer:_____

27. A total of 18Fmk has been divided equally between 4 people. How much will each of them receive?
Answer:_____

28. The reduced price of a jacket was only one third of its original price. What was the original price when the reduced price was 40Fmk?
Answer:_____

29. When you divide the number x by 8 you will get 4 and 2 will be left over. What is the result if the number x is divided by two?
Answer:_____

30. How many times will you have to ski around a circular skitrack of 800 meters in order to ski more than 7 km?
Answer:_____

FIND AT LEAST ONE NUMBER WHICH CAN SUBSTITUTE FOR x

1. $7 - x = 0$ x = ___
2. $x - 4 = 16$ x = ___
3. $5 - x = {}^{-}2$ x = ___
4. ${}^{-}5 + x = {}^{-}1$ x = ___
5. $730 : 35 = x$ x = ___
6. ${}^{-}5 \cdot x = {}^{-}40$ x = ___
7. $x \cdot {}^{-}3 = 36$ x = ___
8. $x : {}^{-}4 = 8$ x = ___
9. $8700-3888 = x$ x = ___
10. $x - 8 < {}^{-}2$ x = ___
11. ${}^{-}8 \cdot x > 16$ x = ___
12. $x - {}^{-}5 = {}^{-}3$ x = ___
13. $\frac{2}{7} < \frac{2}{x}$ x = ___
14. $\frac{3}{8} + x = \frac{5}{8}$ x = ___
15. $\frac{7}{9} - x = \frac{1}{9}$ x = ___
16. $\frac{4}{7} = \frac{12}{x}$ x = ___
17. $\frac{3}{4} : \frac{1}{4} = x$ x = ___
18. $0,85-0,7+x = 1,25$ x = ___
19. $4\frac{5}{6} - (1\frac{1}{3} + 2\frac{2}{3}) = x$ x = ___
20. $x - 14 = 3 \cdot (7+5)$ x = ___
21. $2 \cdot x - 5 = 15$ x = ___
22. $2\frac{3}{4} = \frac{x}{4}$ x = ___
23. $0,6 : x = 0,1$ x = ___
24. $3\frac{3}{8} : x = 1\frac{1}{8}$ x = ___

TEST 5

NAME: _____

SCHOOL: _____

DATE: _____

DATE OF BIRTH: _____

GRADE IN MATHS IN YOUR LAST REPORT: _____

25. You have 20 litres of juice. How many half litre bottles are needed for bottling?
Answer _____

26. The reduced price of a jacket was only one third of its original price. What was the original price when the reduced price was 40 Fmk?
Answer: _____

27. The temperature in the morning was ${}^{-}21^{o}$ and at noon ${}^{-}13\frac{1}{2}^{o}$. How much was the increase between the two temperatures?
Answer: _____

28. Calculate on fifth of the sum of $2\frac{1}{3}$ and $7\frac{2}{3}$.
Answer: _____

29. How many 7,5 litre pails do you need to get a container of 150 litres full?
Answer: _____

30. Having travelled 15 km on a trip, one third of it was still left. How long was the whole trip?
Answer: _____

FIND AT LEAST ONE NUMBER WHICH CAN SUBSTITUTE FOR x

1. $^{-}5 + x = 0$ x = ___
2. $^{-}7 - x = 0$ x = ___
3. $x - 4 = 16$ x = ___
4. $1000 \cdot x = 4{,}5$ x = ___
5. $x : 4{,}5 = 900$ x = ___
6. $4{,}85 - x = 3{,}28$ x = ___
7. $2^3 + 5^2 = x$ x = ___
8. $18 - {}^{-}5 + x = 6$ x = ___
9. $5 : x = 2\frac{1}{2}$ x = ___
10. $5 \cdot x = {}^{-}40$ x = ___
11. $8700 - 3888 = x$ x = ___
12. $x : {}^{-}4 = 8$ x = ___
13. $2 \cdot x - 5 = 15$ x = ___
14. $x - 9 = 3\frac{3}{4}$ x = ___
15. $\frac{x}{-3} = {}^{-}5$ x = ___
16. $x : 3{,}2 = 3{,}0$ x = ___
17. $x : 12 = 144$ x = ___
18. $x - 5 < {}^{-}4$ x = ___
19. $5 - x > 2$ x = ___
20. $3\frac{1}{3} - x = \frac{{}^{-}1}{3}$ x = ___
21. $2{,}4 \cdot x = 10{,}8$ x = ___
22. $({}^{-}4)^3 - 3^2 = x$ x = ___
23. $1500 : x = 120$ x = ___

TEST 6

NAME: _____

SCHOOL: _____

DATE: _____

DATE OF BIRTH: _____

GRADE IN MATHS IN YOUR LAST REPORT: _____

24. How much is 5 % of 120?
    Answer: _____
25. What percentage is 5 of 20?
    Answer: _____
26. Of which number is 4 10 percent?
    Answer: _____
27. Calculate one fifth of the sum of $2\frac{1}{3}$ and $7\frac{2}{3}$.
    Answer: _____
28. A side of a square is 8 cm in length, 15 % of the area of the square has been cut out. What is the area of the rest of the square?
    Answer: _____
29. What number must be subtracted from 4,2 to get $^{-}1{,}2$?
    Answer: _____
30. What is the height of the prism if the length of it is 5 cm, the width 4 cm and the volume 120 cm^3?
    Answer: _____

The value of x belongs to the set of rational numbers.

GRADE 7, LOWER LEVEL GROUP

NAME: _____

SCHOOL: _____

DATE: _____

DATE OF BIRTH: _____

GRADE IN MATHS IN YOUR LAST REPORT: _____

1. $x - 4 = 16$ x = ___
2. $5x = -40$ x = ___
3. $2x - 6 = 14$ x = ___
4. $8 - 4x = 2x + 14$ x = ___
5. $-x + 5 = 14$ x = ___
6. $x - 5 < 9$ x ___
7. $5 - x > 2$ x ___
8. $x^2 = 49$ ___
9. $6x-8-(x-5) = 12$ x = ___
10. $x+3\ (-x-4) = 0$ x = ___
11. $\frac{x-4}{6} = 0$ x = ___
12. $\frac{x+2}{5} = 1$ x = ___

PUT A RING ROUND THE CORRECT ALTERNATIVE		a	b	c	d	e
13. $x \cdot x \cdot x = 8$	x =	-2	2	$\frac{8}{3}$	$\frac{3}{8}$	no solution
14. $\|4-x\| = 13,\ x>0$	x =	-9	9	4	17	13
15. $x-(4-2) = 6$	x =	0	8	12	-4	6
16. $-x = 5\frac{1}{2}$	x =	$-5\frac{1}{2}$	5,05	$\frac{2}{11}$	$5\frac{1}{2}$	5
17. A kilogram of apples costs 3 Fmk 40p. How many kilos of these apples can you buy for 17 Fmk?		4kg	5kg	6kg	4,5kg	5,5kg
18. On a trip 3 km was travelled on foot, 15 km by bike and 7km by boat. What proportion was travelled by bike?		$\frac{1}{3}$	$\frac{3}{20}$	$\frac{3}{5}$	$\frac{7}{20}$	$\frac{1}{15}$
19. The average speed of a car is 90 km/h. How long will a trip of 450 km take when you are assumed to rest 1 hour?		5h	4h	6h	9h	2h
20. $\frac{x}{0,6} = 3$	x =	0,2	3,6	5	18	1,8
21. $1,7x = 0,017$	x =	0,01	0,1	10	100	0,17
22. $-x = \|-6\|+\|+4\|-\|7\|$	x =	-3	-9	3	9	17

GRADE 7, UPPER LEVEL GROUP

Items 1.-12. are exactly the same as in the test for lower level group.

Items 13.-22. are like in the test for lower level group but without alternative for answering.

All items in these tests are open-ended.

GRADE 8, MIDDLE LEVEL GROUP

Items 1.-12. as before (common items)

Items 13.-22. as before

GRADE 8, HIGHEST LEVEL GROUP

Items 1.-22. like in the test for the middle level group.

Item 23. $\frac{2}{3} - \frac{4}{5}x = \frac{1}{4}$

Item 24. $0,4x - 2,9x < 0,8$

GRADE 9, MIDDLE LEVEL GROUP

Items 1.-12. as before (common items)

Items 13.-22. are like in the test for lowest level group but without alternatives for answering.

GRADE 9, HIGHEST LEVEL GROUP

Items 1.-22. like in the test for the middle level group.

Item 23. $2x^2 - 7x + 3 = 0$

Item 24. $x^2 - 4 \leq 12$

1. $x - 4 = 16$ $x =$___
2. $5x = -40$ $x =$___
3. $2x - 6 = 14$ $x =$___
4. $8 - 4x = 2x + 14$ $x =$___
5. $-x + 3 = 14$ $x =$___
6. $x - 3 < 9$ x ___
7. $5 - x > 2$ x ___
8. $x^2 = 49$ ___
9. $6x-8-(x-5) = 12$ $x =$___
10. $x+3(-x-4) = 0$ $x =$___
11. $\frac{x-4}{6} = 0$ $x =$___

GRADE 8, LOWEST LEVEL GROUP

NAME:_____

SCHOOL:_____

DATE:_____

DATE OF BIRTH:_____

GRADE IN MATHS IN YOUR LAST REPORT:_____

12. $\frac{x+2}{3} = 1$ $x=$___

PUT A RING ROUND THE CORRECT ALTERNATIVE

		a	b	c	d	e
13. $\frac{x}{7} - \frac{2x}{7} = \frac{4}{7}$	$x =$	2	$\frac{5}{7}$	$-\frac{5}{7}$	-4	4
14. $\frac{3}{4}x = -2$	$x =$	$-\frac{6}{4}$	$-2\frac{2}{3}$	3	$\frac{3}{8}$	$2\frac{2}{3}$
15. $-3x + 4 = 8 + 7x$	$x =$	-0,4	-4	$\frac{4}{10}$	4	1,2
16. $\frac{x}{4} = \frac{4}{16}$	$x =$	0,4	-4	4	16	1
17. $2x-1-(x-3) = 5+x$	$x =$	what ever	no number	2	3	5
18. $200-x = x-200$	$x =$	0	400	no number	200	100
19. The price of a deep-freezer was 2800 Fmk. There was a 10% reduction and it was paid in five equal installments. How much was each installment?		280 Fmk	560 Fmk	504 Fmk	540 Fmk	616 Fmk
20. What number must be added to the difference of 3/4 and 1/8 in order to get a sum equal to be quotient of the same numbers?		$\frac{5}{8}$	$\frac{3}{24}$	$5\frac{5}{8}$	$5\frac{3}{8}$	$8\frac{3}{8}$
21. By what number must 5,8 be divided in order to get a quotient > 2?		>11,6	<2,9	=11,6	<11,6	>2,9
22. A rectangle is 20 cm in lenght and 15 cm in height. The lenght remains the same. How much must the height be increased to make a 20% increase in area?		3cm	2cm	4cm	1cm	5cm

1. $x - 4 = 16$ x = ___
2. $5x = -40$ x = ___
3. $2x - 6 = 14$ x = ___
4. $8 - 4x = 2x + 14$ x = ___
5. $-x + 3 = 14$ x = ___
6. $x - 3 < 9$ x ___
7. $5 - x > 2$ x ___
8. $x^2 = 49$ ___
9. $6x-8-(x-5) = 12$ x = ___
10. $x+3(-x-4) = 0$ x = ___
11. $\frac{x-4}{6} = 0$ x = ___

GRADE 9, LOWEST LEVEL GROUP

NAME: _____

SCHOOL: _____

DATE: _____

DATE OF BIRTH: _____

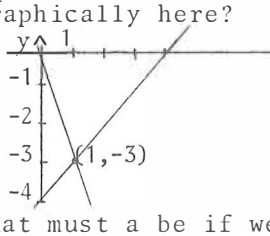
GRADE IN MATHS IN YOUR LAST REPORT: _____

12. $\frac{x+2}{3} = 1$ x = ___

PUT A RING ROUND THE CORRET ALTERNATIVE

		a	b	c	d	e
13. $\|x\| \leq 2,\ x \in Z$	set of solutions	{-2,-1, 0,1,2}	{2}	ø	{-2,2}	{±2,±1}
14. $x^2 + 3x = 0$	set of solutions	{1,3}	{0,3}	{-3}	{0,1}	{0,-3}
15. $-2(3x-4)-(x-6) = 4$	x =	$\frac{6}{7}$	$1\frac{3}{7}$	2	1	$\frac{7}{10}$
16. $(x+4)^2 = 0$	set of solutions	{-4}	{4}	{-4,4}	{0}	{-2}
17. $x^2+2x-3 = 0$	set of solutions	{1,-3}	{-1,3}	{0,-2}	{2,-3}	{3,-2}
18. $\frac{x}{7} = \frac{2}{3}$	x =	$\frac{3}{2}$	$\frac{2}{21}$	$\frac{21}{2}$	$4\frac{2}{3}$	3
19. 0,45x-1,25=2,35+0,25x	x =	$5\frac{1}{7}$	5,5	1,8	18	0,5
20. There are cars and bicycles at a parking place. The total number of them is 21 and they have 58 wheels (possible spare tyres are not taken into account). How many bicycles are there?		12	13	8	11	16

21. What pair of simultaneous equations has been solved graphically here?



$\begin{cases} y=-3x \\ y=-x-4 \end{cases}$ $\begin{cases} y=x \\ y=x-4 \end{cases}$ $\begin{cases} y=x-4 \\ x=-3 \end{cases}$ $\begin{cases} x+1=0 \\ y-3=0 \end{cases}$ $\begin{cases} y=-3x \\ y=x-4 \end{cases}$

22. What must a be if we know that the pair of equations

$\begin{cases} 3x-4y = 12 \\ x-ay = 4 \end{cases}$

has infinite number of solutions?

$\frac{4}{3}$ $\frac{3}{4}$ 3 4 12