DCCUMENT RESUME

ED 154 018 TM 007 028

AUTHOR Gustafsson, Jan-Eric
TITLE The Rasch Model for Dichotomous Items: Theory, Applications and a Computer Program. No. 63.
INSTITUTION Gothenburg Univ. (Sweden). Inst. of Education.
PUB DATE Dec 77
NOTE 158p.; Best copy available

EDRS PRICE MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS *Computer Programs; Equated Scores; *Goodness of Fit; *Item Analysis; *Mathematical Models; *Reliability; *Scores; Standard Error of Measurement; Statistical Analysis; Test Construction; Test Interpretation; Test Items; True Scores
IDENTIFIERS Latent Trait Theory; *Rasch Model; Tailored Testing

ABSTRACT

The Rasch model for test analysis is described and compared with two-parameter and three-parameter latent-trait models. Conditional maximum likelihood equations for estimating item parameters are derived, and estimates of person parameters are described together with their confidence intervals. Goodness of fit tests are discussed, including a graphic test of item fit and two over-all tests. Characteristics of tests which may cause them not to fit the model are listed, together with strategies for developing tests. Applications of the Rasch model to optimizing test efficiency, test equating and linking, and to tailored testing are described. Some generalizations of the basic model and special cases are mentioned. PML, a FORTRAN IV computer program used in applying the Rasch model, is described in detail, and a number of bibliographical references are appended. (CTM)

Reports from

THE INSTITUTE OF EDUCATION
UNIVERSITY OF GÖTEBORG

THE RASCH MODEL FOR DICHOTOMOUS ITEMS: THEORY, APPLICATIONS AND A COMPUTER PROGRAM

No 63
December 1977

Jan-Eric Gustafsson

INSTITUTE OF EDUCATION, UNIVERSITY OF GÖTEBORG Fack S-431 20 MÖLNDAL SWEDEN

CONTENTS

ACKNOWLEDGEMENTS

The research presented in this report has been financially supported by the Swedish Council for Research in the Humanistic and Social Sciences and by the National Board of Education and has been carried out at the Institute of Education, University of Göteborg.

I should like to express my gratitude to some friends and colleagues, most of them at the Institute of Education and some of them at the Department of Educational Research, University of Göteborg, for their showed interest and great help. First of all I wish to thank Leif Lybeck who introduced me to the Rasch model, and throughout the work I have greatly profited from his expertise in the field. Without Jan-Gunnar Tingsell's great skill in the art of computer programming there would probably not have been any computer program to present; not only has he helped me track down errors but has also contributed parts of the program. I also owe debts of gratitude to Kjell Härnqvist, Torsten Lindblad, Berner Lindström and Inga Wernersson who read a first draft of the manuscript and all gave valuable comments. Christina Skönvall had the ardous task of typing the final manuscript; I wish to thank her for a great job.

Mölndal, December 1977 Jan-Eric Gustafsson

. ABSTRACT

The report describes the Rasch model for dichotomous items, or the one-parameter logistic model, which is the simplest of the psychometric latent trait models. In the Rasch model each item is described with only one parameter, the difficulty, and each person is described with only one parameter, the ability. In Chapter 1 the basic features of the model are spelled out and a comparison is made with other, more complex, latent traits models. It is concluded that the Rasch model has decisive advantages over the other models with respect to interpretability, estimation of parameters and possibilities of testing assumptions. In Chapter 2 is shown how conditional maximum likelihood equations for estimating the item parameters can be derived and it is explained how the numerical problems in solving these equations have been solved in a computer program so that estimates can be obtained even for large sets of items. The same chapter also deals with the estimation of person parameters and how to establish confidence intervals for the estimated parameters.

In Chapter 3 goodness of fit tests based on the conditional estimates of the item parameters are presented. A graphic test of item fit is described and two overall numerical tests are taken up: one likelihood ratio test and one chi-square test. In Chapter 4 strategies and problems in developing scales fitting the model are discussed in relation to analyses of some tests developed within the framework of the classical psychometric theory.

Chapter 5 presents some areas of applications of the Rasch model such as test optimation, test equating and linking, and tailored testing. In Chapter 6 some generalizations of the basic model are briefly taken up; it is mentioned that models can be formulated also for the case when there are more than two categories of answer and that a general linear logistic model can be used to study the sources of item difficulty. In Chapter 7, finally, the computer program is presented.

INTRODUCTION

In a discussion about prospective developments in item selection theory for the construction of mental tests Gulliksen (1950) stated that: "A significant contribution to item analysis theory would be the discovery of item parameters that remained relatively stable as the item analysis group changed..." (p. 392).

This problem has been solved, along with several others, within a class of models generally referred to as latent trait models (LT models, or modern test theory; other names sometimes applied are item response theory and item characteristic curve theory).

For different reasons, among which the mathematical and numerical complexities involved probably are the most important, LT models have not yet been widely applied in the development and use of tests, even though the last few years have shown some evidence of a change.

There is in particular one LT model, variously referred to as the Rasch model or the one-parameter logistic model, which has been applied in solving practical problems and which holds special promise for further use. This report presents the Rasch model and indicates at least a selection of all its possible uses. Also presented is a computer program for conditional maximum likelihood estimation of parameters in the model and for computing goodness of fit tests.

Chapter 1

BASIC CONCEPTS AND MODELS IN LATENT TRAIT THEORY

Although the basic tenets of LT theory can be found in early work by Lawley (1943) and Lord (1952, 1953), the breakthrough came in the sixties. (For measurement of attitudes, however, Lazarsfeld very early formulated and used the closely related latent class model, see e.g. Lazarsfeld, 1950). During this decade Rasch (1960, 1966) formulated his model and the computional problems in relation to the model began to be mastered as well (Fischer & Allerup, 1968; Wright & Panchapakesan, 1969). The sixties also saw the advent of the Lord and Novick (1968) treatise in which five chapters (four of which were contributed by A. Birnbaum) dealt with LT theory.

In the last ten years a host of papers has also appeared dealing with specific questions, and rather simple, relatively non-mathematical introductions to LT theory have appeared (e.g. Lybeck, 1974; Willmott & Fowles, 1974; Kifer, Mattson & Carlid, 1975; Baker, 1977; Hambleton et al., 1977) as well as at least one proper text book presentation (Fischer, 1974).

1.1 Three logistic models

Common to all LT models is that one set of parameters is used to describe the items in a test and that another single parameter represents ability. An underlying psychological trait or latent continuum is thus assumed on which the standing of the examinees differs. Another thing common to all LT models is that a function relating the probability of a correct answer to an item is explicitly stated (the item characteristic curve, ICC).

The differences between the models reside in the particular choice made of parameters describing the items and the kind of function used for the ICC. Two kinds of ICC's have been

tried, the normal ogive and the logistic function. However, since the logistic function is mathematically and computationally much more tractable than the normal ogive the three most commonly used models are all based on the logistic function, with the difference between the models residing in the number of parameters used to describe the items.

The one-parameter model

In the simplest case only one parameter is used for each item, its difficulty. In order to describe this model we will need the following notation:

σ_i = The difficulty parameter of item i.

ξ_v = The ability parameter of person v.

$f_i(\xi)$ = The ICC for item i.

A_{vi} = A binary response variable with the value 1 if the answer of person v to item i is correct and the value 0 if incorrect or omitted. A particular realization of this stochastic variable is given the algebraic notation a_{vi}.

k = The number of items in the test.

The one-parameter model, or the Rasch model, asserts that the probability of a correct answer by person v to item i is:

$$(1.1.1) \qquad P(A_{vi}=1|\xi_v,\sigma_i) = \frac{\exp(\xi_v-\sigma_i)}{1+\exp(\xi_v-\sigma_i)}$$

The higher the value of ξ_v the higher the probability of a correct answer and the higher the value of σ_i the lower the probability of a correct answer.

From (1.1.1) follows that the ICC for an item i in the Rasch model can be written:

$$(1.1.2) \qquad P_i(\xi) = \frac{\exp(\xi-\sigma_i)}{1+\exp(\xi-\sigma_i)}$$

Two ICC's for this model are shown in Figure 1.1. Throughout, the curve for the more difficult item is located under the curve for the easier item.

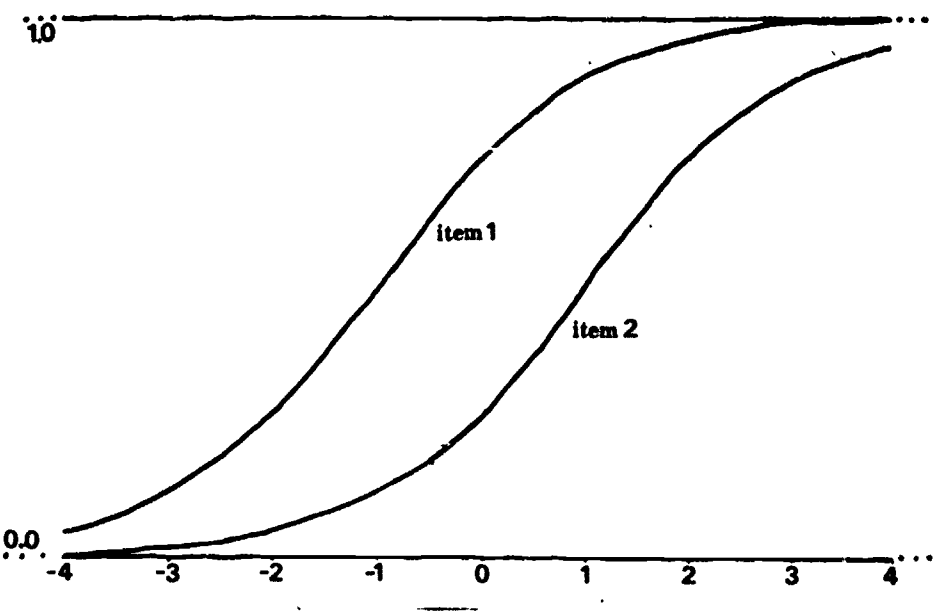


Figure 1.1. Item characteristic curves for two items (σ_1=-1,σ_2=1) in the one-parameter logistic model.

The question may of course be asked as to what the parameters in the model mean and what reality this model may present. A very concrete example which illustrates this is the Flogging Wall test invented by Lumsden (1976) as a tool for thought experiments in test theory and as a "test for test theorists" (p. 251).

> Along a wall at intervals there are k flexible canes attached at various heights. The canes flog slowly and independently up and down. In taking the test the examinee is placed on a cart which is drawn quickly along near the wall and the examinee's score, to be used as a measure of height, is the number of canes which touch him (see Figure 1.2).

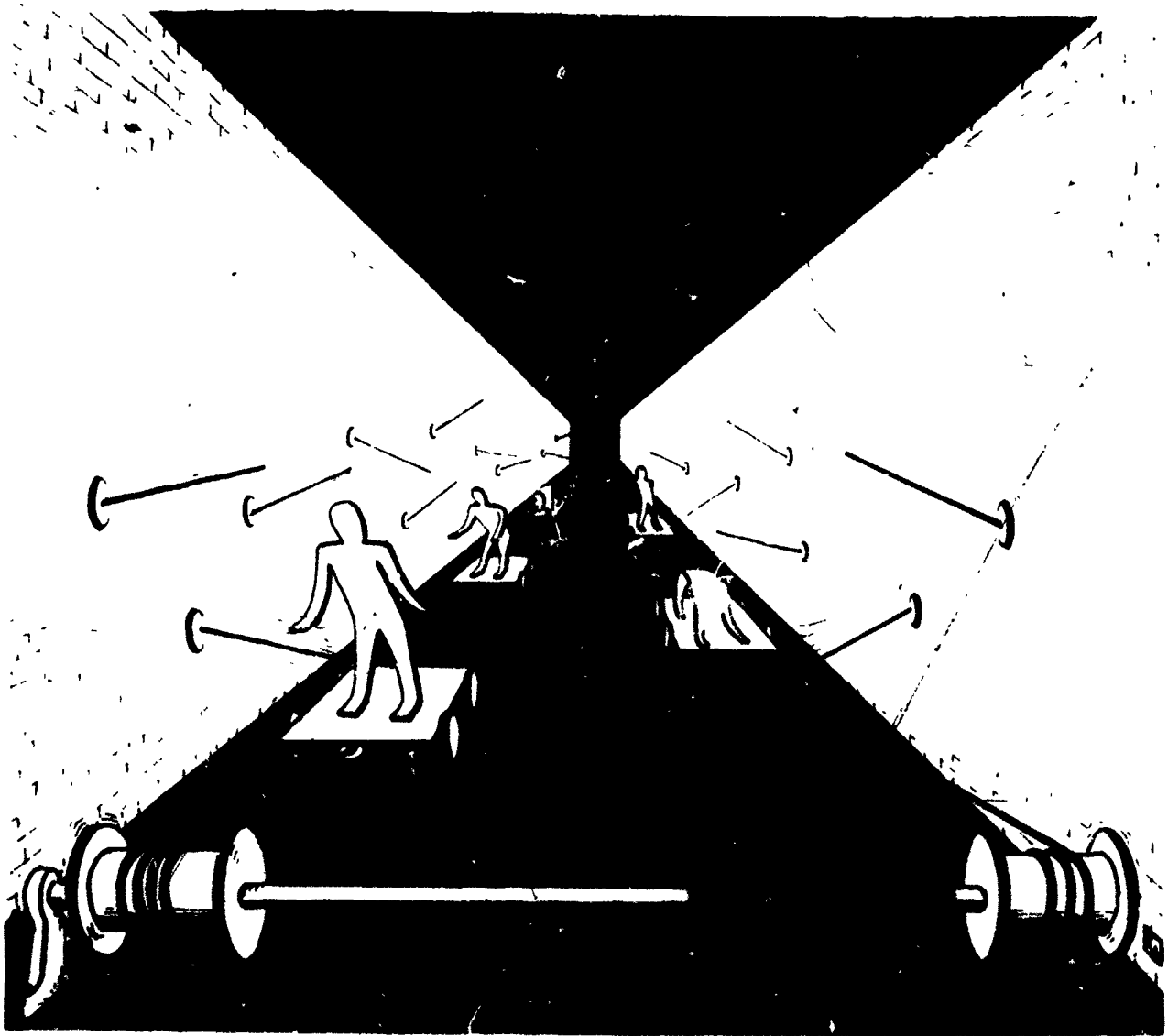


Figure 1.2. An illustration showing the Flogging Wall test (cf Lumsden, 1976, p. 252).

With one assumption, namely that the canes flog with the same amplitudes, this test would fit the Rasch model, with the height of the examinee as the ability parameter ξ_v and the heights of the canes on the wall as the item difficulties σ_i. As will be shown later the Rasch model furthermore implies that the examinee scores and the cane scores (i.e. the number of examinees which a cane touch) can be used to obtain separate estimates of the parameters.

The two- and three-parameter models

In the two-parameter model (or the Birnbaum model as it is sometimes called) another parameter (α_i, i=1,...,k), the discrimination parameter, is introduced which allows the ICC's for different items to have different slopes. The ICC for an item in this model can be written:

$$f_i(\xi) = \frac{\exp\alpha_i(\xi-\sigma_i)}{1+\exp\alpha_i(\xi-\sigma_i)} \qquad (1.1.3)$$

Two ICC's for the two-parameter model are shown in Figure 1.3. For the item with a high discrimination parameter the slope is steep, while it is much more shallow for the item with the low α_i parameter.

We can use the Flogging Wall test to illustrate the meaning of the discrimination parameter too. This parameter would reflect differences in amplitude of the flogging of the canes, i.e. with this model it would no longer be necessary to assume that all the canes have the same amplitude. But it should also be pointed out that with this model we should no longer use the number of canes touching the examinee as an estimator of his height (ability), but instead weight the score on each item with its discrimination parameter.

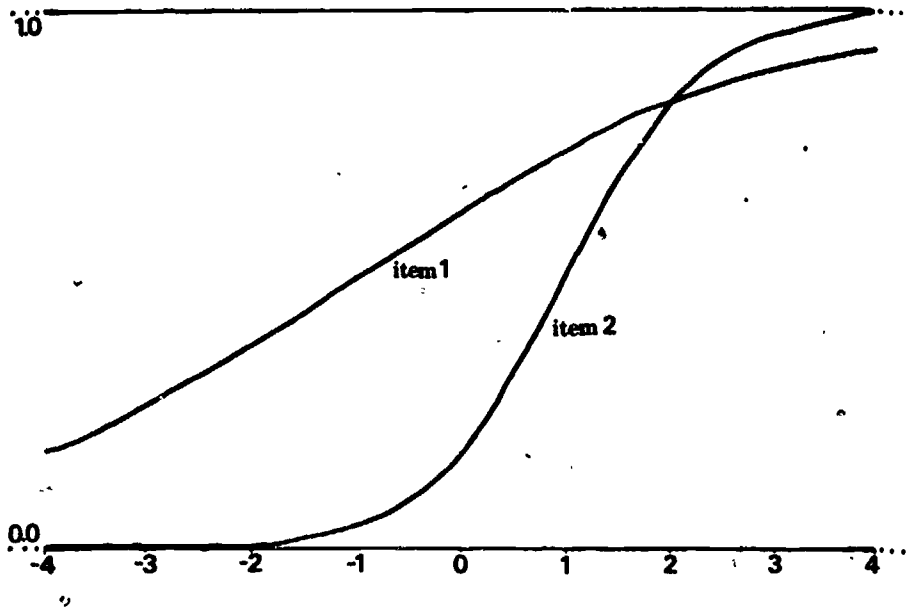


Figure 1.3. Item characteristic curves for two items (σ_1=-1,σ_2=1;σ_1=.5,σ_2=1.5) in the two-parameter logistic model.

Let's return to Figure 1.3 for a moment. Inspection of this figure (observe that only a part of the ability continuum is shown) shows that for low scores on the ability continuum the probability of a correct answer asymptotically approaches 0. This obviously implies that this model, as little as the one-parameter model, can be expected to properly represent the case when the items allow guessing.

A third model, the three-parameter model has been proposed in which another parameter (π_i,i=1,...,k) is introduced to prevent the lower asymptote of the ICC to approach zero. The ICC for an item in this model can be written:

$$(1.1.4) \qquad f_i(\xi)=\pi_i+(1-\pi_i)\frac{\exp\alpha_i(\xi-\sigma_i)}{1+\exp\alpha_i(\xi-\sigma_i)}$$

Inspection of Figure 1.4, where two ICC's for the three-parameter model are depicted, reveals that the curves approach the value of π_i as the lower asymptote (compare the graphs for item 2 in Figure 1.3 and 1.4). Since the lower asymptote can be taken as the probability of obtaining a correct answer obtained by guessing, the parameter π_i is often referred to as the guessing parameter. In can be noted, however, that the estimates of the parameter typically come out lower than the values that would result if examinees of low ability were to guess randomly. For this reason, which Lord (1974a) has attributed to there often being "too attractive" distractors, it has been argued against labelling this parameter "guessing parameter", and instead considering it as the limit of the lower asymptote of the ICC.

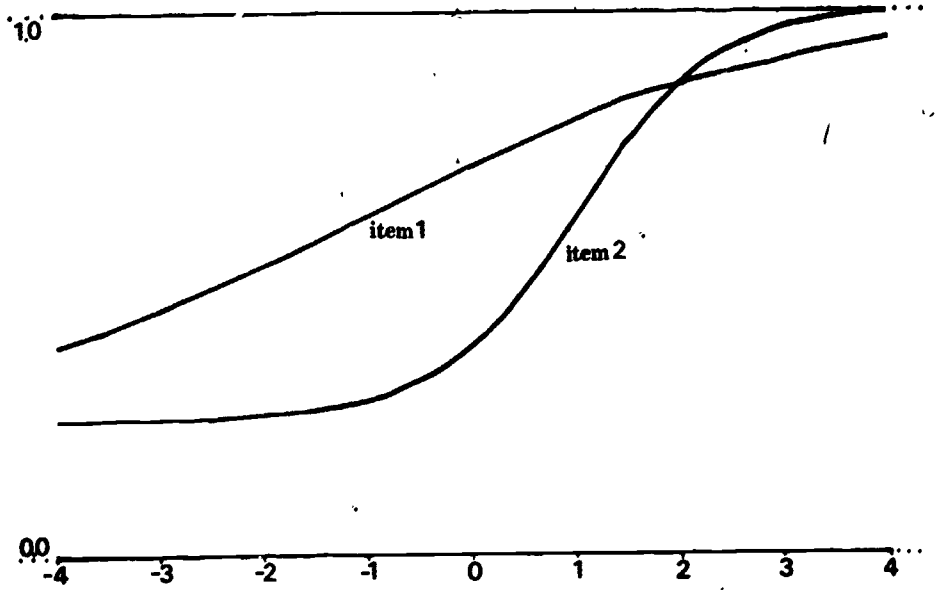


Figure 1.4. Item characteristic curves for two items (σ_1=-1,σ_2=1;α_1=.5,α_2=1.5,π_1=π_2=.25) in the three-parameter logistic model.

1.2 Assumptions underlying the LT models

All applications of LT models imply that in one step or another parameters included in the particular model chosen are estimated from the responses of a group of persons to a set of items. These parameters have a number of desirable properties and when they are at hand a number of problems can be solved which would even be difficult to formulate under the classical approach to test theory (see chapter 5).

However, there are a number of assumptions that must be fulfilled in order for any reasonable estimates of parameters to be achieved, and any sensible application to be made. The three most important assumptions are those pertaining to the dimensionality of the latent space, the principle of local statistical independence and the form of the item characteristic curve.

Unidimensionality

The three latent trait models spelled out above, and several others, are all based on the assumption that there is only one ability underlying examinee performance. The meaning of this assumption can be explained as follows (Hambleton et al. 1977): Suppose that a test of k items is to be used in r subpopulations of examinees (an example for r=2 is one group of boys and one group of girls). For any particular given ability level the conditional distributions of test scores must then be identical if the test is unidimensional. If, however, the conditional distributions vary between the subgroups this can only be because the test is measuring something more than a single ability.

With respect to certain tests in common use, the assumption of unidimensionality is certainly untenable. It can, however, be claimed that a test should be unidimensional since the resulting scores are otherwise more or less meaningless (Lumsden, 1961, 1976). McNemar (1946, also quoted in Lumsden, 1976) expressed this in the following way:

> "Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank can be quantitatively and, within limits, qualitatively similar in their attitude towards a given issue." (p. 268).

The same line of reasoning certainly also applies in the measurement of abilities.

It can be asked how one can make sure that the items intended to constitute a test are unidimensional. Factor analysis of the items is a method that has been used to investigate the number of dimensions involved in taking a test. Lumsden (1961, 1976), specifically, has argued in favor of this method when attempts are made to construct unidimensional tests and several authors have reported applications of factor analysis to assure unidimensionality before proceeding with an LT model.

The method is, however, not without its problems. One problem pertains to the choice of measure of association between the items. The phi-coefficient is commonly used but this measure has the unfortunate characteristic that there are limits on the numerical values it can attain, with the limit varying as a function of the marginal frequencies of the items. A consequence of this may be that even a strictly undimensional test may appear as multidimensional in the factor analysis (Ferguson, 1941).

The tetrachoric correlation is another measure of association that has been used in factor analyses at the item level and which is not limited as to the values it can attain. However, matrices of tetrachoric correlations are often not positive definite with breakdowns of the analyses as a common consequence.

Another problem when factor analysis is used to investigate the dimensionality of a set of items is that unless there are differences in the levels of abilities among the examinees the ratio of the first to the second principal component of the matrix of inter-item correlations will not be large, as is dictated by the assumption of unidimensionality. Since LT models can fruitfully applied even in the case when all the examinees have the same ability this restriction in the applicability of factor analysis is unfortunate.

Even though the problems mentioned above do not wholly invalidate the use of factor analysis before LT models are applied, it cannot be allowed to give the final verdict. The problem is not very serious, however, since the assumption of unidimensionality, along with the other assumptions, can be tested with the LT models themselves through goodness of fit tests.

Local statistical independence

The assumption of local independence implies that the answer of an examinee on one item must not influence his answer on another item. For any two items, i and j, this can be given the following statistical formulation:

$$(1.2.1) \qquad P(A_i=1 \text{ and } A_j=1|\xi) = P(A_i=1|\xi)P(A_j=1|\xi)$$

That is, for a given ability level the probability of getting two given items correct must be equal to the product of the probablities of getting each one of them correct.

Hambleton et al. (1977) pointed out that the assumption of local statistical independence for the case when the ability continuum is unidimensional is equivalent to the assumption of unidimensionality. They argued that, for a fixed ability level, if the responses are not statistically independent, some examinees have higher expected scores than others. Consequently more than one ability would be necessary to account

for test performance.

As a consequence of the equivalence of the two assumptions, what was said above about the testing of the assumption of unidimensionality applies to the testing of the assumption of local statistical independence as well. But it must of course be realized that the kind of action to be taken differs, depending upon which assumption has been violated.

The form of the item characteristic curve

All LT models have in common that a choice must be made as to what kind of ICC to operate with. If this was not done, it would be impossible to formulate the statistical models out of which the equations for estimation of parameters can be determined.

Of course the function relating the probability of a correct answer to an item to ability can take any form, it need not even be continuous (cf "latent class analysis", Lazarsfeld & Henry, 1968). Thus it is always necessary to test the particular assumption made, which can usually be done through applications of goodness of fit tests.

The three logistic models spelled out above differ with respect to the constraints put on the form of the characteristic curve, with the one-parameter model imposing the strongest assumptions and the three-parameter model imposing the least strong constraints. It is of course an empirical question whether, for a given set of data, a less constrained model is necessary or whether a more severely constrained one will do. But partly it is also a question of research strategy in that it is sometimes possible to select from a larger pool of items those that conform to the requirements of the more constrained model.

1.3 The Rasch model versus the other models

The different LT models all have their strengths and weaknesses and they are not all equally applicable to all types of problems. The most important differences seem to reside, however, between the Rasch model on the one hand and the two- and three-parameter models on the other.

The most important drawback of the Rasch model is that it is built on such strong assumptions that it could be argued that the opportunities for using this model are small. It has, however, been shown that it is by no means an impossible task to find existing tests that do fit the model (e.g. Rasch, 1960) and tests can of course be specifically constructed to conform to the requirements of the model. The reason that this might be a preferable strategy is that the Rasch model in many respects has decisive advantages over the other models. These advantages are discussed below.

Interpretability

The size of the item and person parameters can in the Rasch model be given simple interpretations in terms of odds of success on an item. The probability of success on item i for person v (for simplicity this probability will be called P_{vi}) is:

$$(1.3.1) \qquad P_{vi} = \frac{\exp(\xi_v - \sigma_i)}{1+\exp(\xi_v - \sigma_i)}$$

Evidently the odds of success can be written:

$$(1.3.2) \qquad \frac{P_{vi}}{1-P_{vi}} = \lambda_{vi} = \exp(\xi_v - \sigma_i)$$

If we now relate the odds of success for person v to the odds of success for person u on the same item this can be written:

$$(1.3.3) \qquad \frac{\lambda_{vi}}{\lambda_{ui}} = \frac{\exp(\xi_v - \sigma_i)}{\exp(\xi_u - \sigma_i)}$$

which can be simplified into:

$$(1.3.4) \qquad \frac{\lambda_{vi}}{\lambda_{ui}} = \exp(\xi_v - \xi_u)$$

We thus see that when two persons are compared this does not involve the item parameter at all and it can easily be shown that when two items are compared, the comparison does not involve the abilities of the persons. These possibilities for comparing persons independently of items, and items independently of persons form the core of Rasch's theory of "specific objectivity" (Rasch 1960, 1961, 1966) and it can be shown that the one-parameter logistic model is necessary and sufficient to obtain this kind of obejctivity.

Behavioral scientists are probably more conversant with the additive linear model upon which the analysis of variance and related models are built than with the exponential family of models. Framed in the language of the linear additive model, however, it can be said that the Rasch model is a model that does not allow for any interactions, i.e. the difficulty of an item must not be qualified by the conditions under which it is taken or by which person takes it. On the other hand it is of course quite difficult to imagine items the difficulties of which are immanent to such a degree that they will never be qualified by any factor. The boundary conditions for a set of items to conform to the model should thus be sought - which in a sense is done each time the model is applied and a goodness of fit test is computed.

Equation (1.3.4) above not only says that persons can be compared independent of items but can also be used to compute the relative odds of success on any item for any two persons. When the persons have the same ability the relative odds are 1 since $\exp(0)=1$. If, to take another example, person v has the person parameter 2.0 and person u the parameter -1.0 the relative odds of success on any item in favour of person v are 20 since $\exp(3)=20.1$.

The same type of calculations can of course also be applied in the comparison on items. Only the Rasch model allows this kind of simple probability statements and simple comparisons between items and persons.

Estimation of parameters

All the LT models have separable parameters which can, at least in principle, be estimated on scales that are independent of the particular sample of examinees studied. The theoretical and practical problems connected with the estimation of parameters have, however, been adequately solved only for the Rasch model.

The common approach that has been used to derive the equations for the estimation of parameter is the maximum likelihood (ML) method. However, a straightforward ML approach, resulting in equations in which the item parameters and the person parameters are estimated simultaneously, yields estimates which are not consistent, as has been shown by Andersson (1973a) and Martin-Löf (1973) (see chapter 2 below for further details).

This problem arises when structural parameters (the item parameters) are estimated in the presence of incidental parameters (the person parameters). Increasing the sample size obviously does not solve the problem since each new person brings a new incidental parameter. But it has been shown (Anderson, 1973) that if the likelihood equation can be formulated only in the item parameters, then consistency and

unbiasedness is assured, which can be done if there exists a minimal sufficient statistic for the person parameters. In the Rasch model, and only in the Rasch model, total score can be shown to be such an estimator of ability. Thus it is possible to formulate ML estimators for the item parameters through conditioning on the total score in the Rasch model but not in the other models.

In spite of the fact that the conditional maximum likelihood (CML) approach is the correct one for estimating the item parameters in the Rasch model, the unconditional (UML) approach in which item parameters and person parameters are estimated simultaneously is the one that has most commonly been used. (Wright & Panchapakesan, 1969; Wright & Mead, 1977; Wright & Douglas, 1977). The reason for this is that the CML method is computationally cumbersome and that numerical problems have prevented its use on tests with more than 20-40 items. The computer program presented in chapter 7 below does, however, present a remedy since it can be used for CML estimation of parameters for larger sets of items.

It can be pointed out parenthetically that there is some confusion concerning the use of the terms conditional and unconditional estimates in LT models. Unfortunately Bock and Lieberman (1970) used these terms in a rather peculiar sense deviating from common use in mathematical statistics. By imposing assumptions about the distribution of person parameters they were able to state the estimating equations for the item parameters in the two-parameter model without introducing the person parameters. These estimates were termed unconditional estimates while they used the term conditional estimates for those resulting when both sets of parameters are estimated simultaneously. The terms conditional and unconditional thus in a sense carry the opposite meaning in the usage of Bock and Lieberman as compared to the usage above in connection with the Rasch model. In the sequel of this paper the latter meaning of the terms will be implied.

Summarizing the discussion so far it can be concluded that only for the Rasch model are there solutions to the problem

of estimating the parameters which are theoretically completely satisfying. (To be fair, however, it must be pointed out that this is true only with respect to the estimation of the item parameters; the unbiased estimation of abilities is still a problem to be solved). But aside from these theoretical questions there are also important differences between the Rasch model and the other models with respect to the amount of practical problems met with in estimating the parameters.

Since solutions to the likelihood equations are not available in closed form, numerical methods must be resorted to. However, for the two-parameter model the iterative approach employed does not converge properly unless both the number of examinees and the number of items is large (at least 1000-3000 persons and 30-60 items seem to be required). The amount of computer time required is also very great. Hambleton et al. (1977, p. 107) report, for example, that for a test with 60 items given to 5305 examinees 40-60 minutes was required for convergence on an IBM 360/65. Practical problems alone thus make application of the two- and three-parameter models out of the reach for many researchers and for many problems.

For the Rasch model, however, the iterative procedure almost never fails and at least for well conditioned problems where the number of items is not very large (less than 50 to 80 items, say) more than a few minutes on an IBM 360/65 is seldom required, even when the CML estimates are computed.

Testing assumptions

Goodness of fit tests exist for all the different LT models (Rasch, 1960; Wright & Panchapakesan, 1969; Andersson, 1973b; Bock, 1972; Martin-Löf, 1973; Mead, 1976b). The tests generally are of the chi-square or likelihood ratio type and at least for some of the proposed test statistics it has been shown that they assume the specified distribution at least asymtotically (Martin-Löf, 1973; Anderson, 1973b).

More important, perhaps, than the statistical properties of the proposed tests are the difference between the different LT

models with respect to the possibilities of detecting important deviations from the assumptions.

It appears that the Rasch model is "safer" in this respect than the other two models. Mead (1976a) discussed factors such as guessing, carelessness, speed, practice and item bias as threats to the fit of data to the Rasch model. He concluded by saying:

> "All of the disturbances considered represent some form of multidimensionality; they would violate any model that assumes unidimensionality. Since the effect of the disturbances often appears as a change in the slope of the item characteristic curve, any model which includes item discrimination as a parameter would appear to fit the data." (Mead, 1976a, p. 11)

There is thus a risk in using the less constrained models since threats to the important assumption of unidimensionality can be "taken care of" as varying item discrimination.

It is of course true that the importance of testing a more constrained model with powerful means is extremely important since otherwise all claims for superiority are invalidated. Fortunately there do exist sound statistical tests for the goodness of fit tests of the Rasch model, at least when the CML approach is used (see chapter 3 below).

Conclusion

In the comparisons made between the Rasch model on the one hand and the two- and three-parameter models on the other with respect to interpretability, estimation of parameters and testing assumptions, the Rasch model shows up more favorably in every respect. If it can empirically be shown that it is possible to make educational and psychological measurements which conform to the requirements of the model it will find a number of different uses. Some of the possible applications will be discussed in chapter 5 after a more

detailed presentation of the mathematics of the model and procedures for testing goodness of fit has been made.

Chapter 2

THE MATHEMATICS OF THE RASCH MODEL

In this chapter the structure of the Rasch model will be more formally exposed and it will be shown how the parameters in the model can be estimated. But the presentation also serves as a documentation of the computer program (called PML) presented in chapter 7; the solution of some numerical problems are presented in detail and operating characteristics of the program are presented.

2.1 Estimating item parameters

In developing the mathematics of the one-parameter model we will make use of a somewhat different notation from that used hitherto. The derivation is at points greatly simplified if an antilogarithmic transformation is made of the parameters such that $\theta_v=\exp(\xi_v)$ and $\varepsilon_i=\exp(-\sigma_i)$. The probability of success for person v on item i can then be written:

$$(2.1.1) \qquad P(A_{vi}=1|\theta_v,\varepsilon_i)=\frac{\theta_v\varepsilon_i}{1+\theta_v\varepsilon_i}$$

The usual testing situation is one in which n examinees have been given k items. As previously we assume that the response variable is of the Bernoulli type, so that in keeping with the previous notation

$$A_{vi}=\begin{cases}1 \text{ if person v is succesful on item i}\\ 0 \text{ if person v is not successful on item i}\end{cases}$$

Then we can write (2.1.1):

$$(2.1.2) \qquad P(A_{vi}=a_{vi}|\theta_v,\varepsilon_i)=\frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i}$$

The observed data can be assembled in the matrix $((a_{vi}))$ shown below:

Examinees

Items	1	...	v	...	n	
1	a_{11}	...	a_{v1}	...	a_{n1}	s_1
.	.		.		.	.
.	.		.		.	.
.	.		.		.	.
i	a_{i1}	...	a_{vi}	...	a_{in}	s_i
.	.		.		.	.
.	.		.		.	.
.	.		.		.	.
k	a_{k1}		a_{kv}		a_{kn}	s_k
	r_1	...	r_v	...	r_n	

The raw score for person r is thus:

$$r_v = \sum_{i=1}^{k} a_{vi}$$

and the total number of correct responses to item i (the item score) is:

$$s_i = \sum_{v=1}^{n} a_{vi}$$

Those persons who have 0 or k correct answers must be excluded from the $((a_{vi}))$ matrix since no estimates of their parameters are possible to obtain. Also items with 0 or n correct answers must be excluded from the matrix for the same reason.

Under the assumption of statistical independence the likelihood of the data matrix $((a_{vi}))$ is the product of the probabilities of all the answers:

$$(2.1.3) \qquad \Lambda = \prod_{v=1}^{n} \prod_{i=1}^{k} \frac{(\theta_v \varepsilon_i)^{a_{vi}}}{1+\theta_v \varepsilon_i} = \frac{\prod_v \theta_v^{r_v} \prod_i \varepsilon_i^{s_i}}{\prod_v \prod_i (1+\theta_v \varepsilon_i)}$$

Looking at the likelihood function we find that only the marginal sums of $((a_{vi}))$ are represented and not the "inner" of the matrix. Thus we need not take into account which items a certain examinee has answered correctly or which examinees answered a certain item correctly. In other words, we find that raw score is a sufficient estimator for the person parameters and that item score is a sufficient estimator for the item parameters.

The likelihood function Λ can be maximized in the usual way with respect to the parameters to yield ML estimates of the (θ_v) and (ε_i) (Wright & Panchapakesan, 1969; Martin-Löf, 1973; Fisher, 1974 p 257 ff; Wright & Douglas, 1977). Written in a simple form, although not very suitable for computations, the estimating equations are:

$$(2.1.4) \qquad \begin{cases} s_i = \sum_{v=1}^{n} \frac{\theta_v \varepsilon_i}{1+\theta_v \varepsilon_i} \\ r_v = \sum_{i=1}^{k} \frac{\theta_v \varepsilon_i}{1+\theta_v \varepsilon_i} \end{cases}$$

There is one more parameter to be estimated than there are equations, a problem that can be solved through using some kind of normation. One possibility is putting the parameter value of one item to unity and another possibility is using the product normation $\prod_i \varepsilon_i = 1$. The system of equations can only be solved iteratively but there do exist efficient computer programs for this purpose (Wright & Panchapakesan, 1969; Wright & Mead, 1977).

The approach sketched above is the unconditional maximum likelihood approach that was mentioned above on page . In that context it was also pointed out that the UML method produces estimates which are not consistent.

That ML estimators in certain situations fail to be consistent was first discovered by Neyman & Scott (1948). One class of situations in which this occurrs is the one in which the model contains incidental (or nuisance) parameters beside those structural parameters which are to be estimated. The most commonly known example of such a situation is the estimate of a population variance for a normal distribution (see Andersen, 1973a, p. 14 ff; Martin-Löf, 1973, p. 76). It is known that the deviation sum of squares is to be divided with n-1 to give the unbiased estimate. The ML estimator, however, can be shown to make use of n as the denominator and this estimate is thus biased. This occurs because the population mean must be estimated from the sample data; each new sample will thus give a new value on this (incidental) parameter .

In the Rasch model the person parameters are incidental parameters when we want to estimate the item parameters (and the item parameters are incidental parameters when we want to estimate the person parameters) and of course the number of person parameters does not stabilize when we increase the sample size since each new person brings a new parameter (this fact must not be confused with the fact that since the model is discrete we can only get a limited number of estimates of all the possible person parameters).

But it has been shown (Andersen, 1973a) that if the likelihood equation can be formulated only in the item parameters consistency is assured. This can be done if their exists a minimal sufficient statistic for the person parameters and in the Rasch model, and only in the Rasch model, raw score is such an estimator of ability. Thus it is possible to formulate ML estimators for the item parameters through conditioning on raw score. The details of this are presented below but first we shall discuss another approach to come to grips

with the inconsistency of the UML-estimates.

This approach consists in seeking corrections to rectify the UML estimates, in a similar vein with the way in which the ML estimate for the population variance is corrected for with the factor n/n-1. Simulation studies carried out within the range of 2 to 40 items (Wright & Panchapakesan, 1969; Fischer & Scheiblechner, 1970; Wright & Douglas, 1977) have indicated that for item parameter on the log scale a correction factor of (k-1)/k is suitable, and when this correction is applied the difference between the UML and CML estimates is generally not greater than one unit in the second decimal place.

Since the CML estimates are quite cumbersome computationally it could be argued that the corrected UML estimates would do for all practical purposes. There are, however, three reasons for which the UML estimates should be discarded, in spite of this, in favor of the CML estimates. The first reason is that the particular correction factor employed is empirically rather than theoretically derived and its validity hinges entirely on the range of situations studied in the simulations. My own impression is that the correction used works quite well in situations were the variance of person parameters is not to large but that it tends to become poorer when this variance increases. (These observations were made when data were generated under the two-parameter model with a high but for all items common discrimination parameter, and then analyzed under both the UML and CML approaches. The Rasch model of course does not assume that the discrimination parameter for all items is exactly unity; all that is assumed is that all the items have the same discrimination parameter. Varying item discriminations among sets of items is taken into account as a simultaneous transformation of the scales of item and person parameters and a high discrimination shows as a high variance of the person parameters.)

The second reason is that no correction has as yet been found for the bias in the person parameters. In virtually all the computer programs for UML estimation the person parameters

obtained in the simultaneous estimation of parameters are presented; somewhat better results would be obtained if the person parameters instead were estimated from the corrected item parameters.(In fact, since the problem of a strictly conditional estimation of person parameters is not solved yet, this would in many cases place the UML- and CML-approaches on an equal basis). Even without presenting any exact figures it can be claimed that the bias in the UML estimated person parameters is rather severe. It can be observed that when data are generated in accordance with the Rasch model with a standard deviation (s) of, say, unity and the s of the estimated person parameters is computed, a rather similar value is observed. But in fact the observed person parameters should have a higher s since another variance component (corresponding to the standard error of measurement) has been introduced in generating the data.

The third reason in favor of the CML approach concerns the possibilities of testing goodness of fit: under the UML approach only approximate techniques have been proposed (Wright & Panchapakesan, 1969; Mead,1976b) while under the CML approach there are test statistics which have an at least asymptotically known distribution (see chapter 3).

The most important reason for not employing the CML approach has been that numerical problems have prevented its use with more than a limited number of items. It is,however, shown below that these problems can be solved.

In developing the conditional approach let us first for simplicity consider a given examinee with the raw score r_v, corresponding to the person parameter θ_v. The probability of obtaining any score vector (a_{vi}) given the person parameter and the vector of item parameters is:

$$(2.1.5) \qquad P\{(a_{vi})|\theta_v,(\varepsilon_i)\} = \prod_{i=1}^{k} \frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i} = \frac{\theta_v^{r_v}\prod_i \varepsilon_i^{a_{vi}}}{\prod_i (1+\theta_v\varepsilon_i)}$$

To be able to express this probability as a conditional probability given score r_v we must know the probability of obtaining score r_v given θ_v. This latter probability is given by the sum of the probabilities of all possible ways of obtaining the score r_v, that is the sum of all the expressions like (2.1.5) in which the vector (a_{vi}) sums to r.

A given score r obtained on k items can of course be obtained in $\binom{k}{r}$ different ways. We will need a special notation to be able to express this in a simple way. Define:

$$(2.1.6) \qquad \gamma_r\{(\varepsilon_i)\} = \sum_{\sum_i a_{vi}=r} \prod_{i=1}^{k} \varepsilon_i^{a_{vi}}$$

In the expansion of this sum of products the summation is made over those $\binom{k}{r}$ combinations in which $\sum_i a_{vi}=r$. The $\gamma_r\{(\varepsilon_i)\}$ (or, for short, γ_r) is called the elementary symmetric function of order r in the parameters (ε_i). (On the following pages a more concrete presentation of these symmetric functions will be made).

We can now write the probability of obtaining the score r given θ_v and (ε_i):

$$(2.1.7) \qquad P\{r|\theta_v,(\varepsilon_i)\} = \sum_{\sum_i a_{vi}=r} \prod_{i=1}^{k} \frac{(\theta_v\varepsilon_i)^{a_{vi}}}{1+\theta_v\varepsilon_i} = \frac{\theta_v^{r_v}\gamma_r}{\prod_i (1+\theta_v\varepsilon_i)}$$

The conditional probability of obtaining the vector (a_{vi}) with the total score r_v, given the score r_v is thus given by equation (2.1.5) through equation (2.1.7):

$$(2.1.8) \qquad P\{(a_{vi})|r,(\varepsilon_i)\} = \frac{P\{(a_{vi})|\theta_v,(\varepsilon_i)\}}{P\{r|\theta_v,(\varepsilon_i)\}} = \frac{\prod_{i=1}^{k} \varepsilon_i^{a_{vi}}}{\gamma_r}$$

We thus see that this conditional probability is not a function of θ_v; only the item parameters appear in the expression (2.1.8).

Since the examinees are assumed to be independent we obtain the conditional likelihood of the data matrix $((a_{vi}))$ for n persons as:

$$\Lambda = \prod_{v=1}^{n} \frac{\prod_{i=1}^{k} \varepsilon_i^{a_{vi}}}{\gamma_{r_v}} \qquad (2.1.9)$$

If we use n_r to denote the number of persons with raw score r (r=1,...k-1) and recall that s_i is the score of item i (s=1,...n-1) we can simplify (2.1.9) into:

$$\Lambda = \frac{\prod_{i=1}^{k} \varepsilon_i^{s_i}}{\prod_{v=1}^{n} \gamma_{r_v}} = \frac{\prod_{i=1}^{k} \varepsilon_i^{s_i}}{\prod_{r=1}^{k-1} \gamma_r^{n_r}} \qquad (2.1.10)$$

From this conditional likelihood function the CML-estimators can be derived. If we first take the logarithm of both sides we get:

$$\log\Lambda = \sum_{i=1}^{k} s_i \log\varepsilon_i - \sum_{r=1}^{k-1} n_r \log\gamma_r \qquad (2.1.11)$$

We differentiate with respect to all the ε_i and set the derivatives equal to zero:

$$\frac{\delta \log\Lambda}{\delta \varepsilon_i} = \frac{s_i}{\varepsilon_i} - \sum_{r=1}^{k-1} \frac{\gamma_{r-1}^{(i)}}{\gamma_r} \qquad (i=1,\ldots,k) \qquad (2.1.12)$$

in which equation the symbol $\gamma_{r-1}^{(i)}$ is used for the partial derivative of γ_r with respect to ε_i. This derivative is a symmetric function of order r-1 in all parameters except ε_i. This is most easily seen in an example. Suppose that k=4 and that we are studying γ_2.

In explicit notation the symmetric function can be written:

$$\gamma_2(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4) = \varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_4+\varepsilon_2\varepsilon_3+\varepsilon_2\varepsilon_4+\varepsilon_3\varepsilon_4$$

and

$$\frac{\delta\gamma_2}{\delta\gamma_1} = \varepsilon_2+\varepsilon_3+\varepsilon_4 = \gamma_1(\varepsilon_2,\varepsilon_3,\varepsilon_4) = \gamma_{2-1}^{(1)}(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4)$$

From (2.1.12) it is seen that we end up with a set of nonlinear equations in the (ε_i) (for simplicity we will not distinguish between the parameters (ε_i) and the estimates $(\hat{\varepsilon})$ of the parameters).

$$(2.1.13) \qquad s_i = \sum_{r=1}^{k-1} \frac{n_r \varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} \qquad (i=1,\ldots,k)$$

From the fact that $\Sigma s_i = \Sigma n_r$, it follows that we must impose some constraint on the system of equations to be able to solve it. The same normations as those mentioned above (p 22) in connection with the UML approach are of course available and we can set $\varepsilon_m=1$. (It is practical to select m as the item with medium difficulty. This is done automatically in the PML program, but there is also the option to select any item.)

Even after normation it is not possible to find an explicit solution to the system of equations but there exist numerical methods (Andersen, 1972; Martin-Löf, 1973; Fischer, 1974) which can be used. In the application of these iterative methods there are two important problems to be solved: the first pertains to the computation of the symmetric functions γ_r, and the second to how a rapid convergence of the sequence of iterations can be obtained.

The computation of the symmetric functions

The symmetric function of order r consists of a sum of $\binom{k}{r}$ products, each of which consists of r terms. For example,

when k=50 and r=25 the symmetric function is a sum of about 1.26×10^{14} terms, each of which is a product of 25 terms. Obviously it is impossible to compute the γ_r and the derivatives through a process of straightforward multiplication and summation.

Fortunately there do exist recursive formulas which make a relatively rapid computation of the symmetric functions possible (Fischer, 1974, p 242 ff; Andersen, 1972). We can write:

$$\gamma_r = \varepsilon_i \gamma_{r-1}^{(i)} + \gamma_r^{(i)} \qquad (2.1.14)$$

This is true since $\gamma_r^{(i)}$ is the sum of all products of r parameters that do not contain ε_i, and $\varepsilon_i \gamma_{r-1}^{(i)}$ is the sum of all products of r parameters that contain ε_i. An example should clarify this. Suppose that k=4 and r=2. Then we want to get:

$$\gamma_2 = \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4 + \varepsilon_2\varepsilon_3 + \varepsilon_2\varepsilon_4 + \varepsilon_3\varepsilon_4 \qquad (2.1.15)$$

If we take the partial derivative of γ_2 with respect to ε_1 we get:

$$\gamma_{2-1}^{(1)} = \varepsilon_2 + \varepsilon_3 + \varepsilon_4 \qquad (2.1.16)$$

and

$$\varepsilon_1 \gamma_1^{(1)} = \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4 \qquad (2.1.17)$$

In the same way we can easily convince ourselves that the partial derivative of γ_3 with respect to ε_1 is:

$$(2.1.18) \qquad \gamma_2^{(1)} = \varepsilon_2\varepsilon_3+\varepsilon_2\varepsilon_4+\varepsilon_3\varepsilon_4$$

If we now compare the sum of (2.1.18) and (2.1.17) with (2.1.15) we find them equivalent.

Another recursive relationship of great use is the following:

$$(2.1.19) \qquad r\gamma_r = \sum_{i=1}^{k} \varepsilon_i \gamma_{r-1}^{(i)}$$

This formula can be derived from (2.1.14) but again we use the example to convince ourselves. We get:

$$(2.1.20) \qquad \begin{aligned} \varepsilon_1\gamma_1^{(1)} &= \varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_4 \\ \varepsilon_2\gamma_1^{(2)} &= \varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3+\varepsilon_2\varepsilon_4 \\ \varepsilon_3\gamma_1^{(3)} &= \varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3+\varepsilon_3\varepsilon_4 \\ \varepsilon_4\gamma_1^{(4)} &= \varepsilon_1\varepsilon_4+\varepsilon_2\varepsilon_4+\varepsilon_3\varepsilon_4 \end{aligned}$$

We thus see that in this set of equations the six product terms in (2.1.15) each appear two times.

From the two recursive formulas (2.1.14) and (2.1.19) it is possible to devise a very efficient algorithm for the computation of the symmetric functions of all orders and all the derivatives. Starting from the fact that $\gamma_0^{(i)}=1$ we get $\gamma_1=\Sigma\varepsilon_i\gamma_0^{(i)}=\Sigma\varepsilon_i$. Then we can compute $\gamma_1^{(1)}=\gamma_1-\varepsilon_1\gamma_0^{(1)}$ and all the other derivatives of the symmetric functions of order two. In the next step we get $2\gamma_2=\Sigma\varepsilon_i\gamma_1^{(i)}$ and can then obtain all the derivatives of the functions of the third order, and so on.

The algorithm has been programmed by Fischer (1974, p. 544) and this subroutine is used as one of the methods of computing the symmetric functions in the PML program. The algorithm has the virtue of being very fast: only k^2 multiplications, k^2 additions, k^2 subtractions and k divisions are performed. It has one serious drawback, however: When the number of items is large and/or there are great differences in the size of the item parameters the computations break down as a consequence of round-off errors. The problem is caused by the differences $\gamma_r - \varepsilon_i \gamma_{r-1}^{(i)}$ which, particularly for the orders around k/2, involve very large numbers, resulting in cancellation of terms. These problems are reduced if, as is done in the algorithm used, the recursive formulas are applied both from "below", starting with order one, and from "above", starting with order k and then meeting at about k/2 (which procedure also allows a test of computational accuracy). However, even with this method there is, when k is large, a virtually inescapable loss of accuracy when floating-point representation is used and even attempts to use extended precision (REAL*16 on IBM machines) have failed to appreciably increase the number of items that can be analyzed.

The breakdown of this algorithm (which will be referred to as the Difference algorithm) occurs in the range of 20-40 items. Since k for many tests is within this range, use of this algorithm is accompanied by the frustrating experience that sometimes the analysis breaks down, and that sometimes it does not, for example when different sub-groups are studied.

Fortunately, it is possible to find a recursive formula for the computation of the symmetric functions in which no use is made of subtraction. We can write (2.1.14) in a slightly different way:

(2.1.21) $$\gamma_r(\varepsilon_1,\ldots,\varepsilon_t) = \gamma_r(\varepsilon_1,\ldots,\varepsilon_{t-1}) + \varepsilon_t \gamma_{r-1}(\varepsilon_1,\ldots,\varepsilon_{t-1})$$

(Fischer, 1974, p. 250).

But this means that we can add one parameter at a time, so to speak. If we start with ε_1 then $\gamma_1(\varepsilon_1)=\varepsilon_1$ and $\gamma_0(\varepsilon_1)=1$. If we add one more parameter, ε_2, we have:

$$\gamma_1(\varepsilon_1,\varepsilon_2)=\gamma_1(\varepsilon_1)+\varepsilon_2\gamma_0(\varepsilon_1)=\varepsilon_1+\varepsilon_2$$

$$\gamma_2(\varepsilon_1,\varepsilon_2)=\gamma_2(\varepsilon_1)+\varepsilon_2\gamma_1(\varepsilon_1)=0+\varepsilon_1\varepsilon_2=\varepsilon_1\varepsilon_2$$

Adding a third parameter we get:

$$\gamma_1(\varepsilon_1,\varepsilon_2,\varepsilon_3)=\gamma_1(\varepsilon_1,\varepsilon_2)+\varepsilon_3\gamma_0(\varepsilon_1,\varepsilon_2)=\varepsilon_1+\varepsilon_2+\varepsilon_3$$

$$\gamma_2(\varepsilon_1,\varepsilon_2,\varepsilon_3)=\gamma_2(\varepsilon_1,\varepsilon_2)+\varepsilon_3\gamma_1(\varepsilon_1,\varepsilon_2)=\varepsilon_1\varepsilon_2+\varepsilon_2\varepsilon_3+\varepsilon_1\varepsilon_3$$

$$\gamma_3(\varepsilon_1,\varepsilon_2,\varepsilon_3)=\gamma_3(\varepsilon_1,\varepsilon_2)+\varepsilon_3\gamma_2(\varepsilon_1,\varepsilon_2)=\varepsilon_1\varepsilon_2\varepsilon_3$$

and so on.

After we have added the k parameters we have thus obtained the symmetric functions of all orders in the parameters. This algorithm too has been programmed by Fischer (1974, p 544) who uses it to compute the second partial derivatives of the symmetric functions, which is done through setting as equal to zero the parameter vales for all combinations of items two at a time. But it can of course also be used, with some slight alterations, to compute the symmetric functions themselves, as well as the first derivatives.

In order to obtain the symmetric functions and the derivatives the routine has to be called (k+1) times. Each call to the routine makes use of $\frac{k(k-1)}{2}$ multiplications and $\frac{k(k-1)}{2}+k-1$ additions so to obtain the needed information roughly $\frac{k(k^2-1)}{2}$ multiplications and $\frac{k(k^2-1)}{2}+k^2-1$ additions are performed. If the number of arithmetic iperations necessary for for this algorithm (which will be referred to as the Summation algorithm) is compared with the number of opera-

tions used by the Difference algorithm it is found that the Summation algorithm is slower. It can also be seen that execution time must increase rapidly as k gets larger.

Nevertheless, the Summation algorithm is not unbearingly slow: A complete iteration cycle, which involves computation of the symmetric functions of all orders and all the first derivatives, requires for 40 items about 1 second of CPU time on the IBM 360/65, and for 60 items about 4 seconds is required. For a long test containing 100 items some 20 seconds would be required for each iteration.

These estimates of computer time required are valid for the case when the computations are carried out in double precision. However, since the Summation algorithm is very accurate numerically there is in the PML program an option of using single precision arithmetic in this algorithm. When this option is used, somewhat less computer time (a reduction of some 10 per cent is a reasonable estimate) is required.

In addition to the fact that the amount of computer time required may become prohibitive when very long tests (k>100, say) are analyzed there is one more problem that may appear. The problem is that the symmetric functions, and especially those of orders around k/2 assume very large values and sooner or later the limit set by the size of the floating point numbers which can be represented in the particular computer used will be reached. This problem could be solved through scaling down the parameter values, but since the product normation is used in the PML program after the parameters have been estimated this method is not immediately available in the program.

The amount of computer time required for each iteration is one factor affecting the cost of the analysis. Another important factor is of course the number of iterations required. How to obtain a rapid convergence is discussed next.

The convergence of the iterations

Several different methods have been proposed for the solution of the system on non-linear equations (2.1.13). Andersen (1972) suggested Fisher's Method of Scoring for the equations to be solved in the polychotomous model, which has been programmed by Allerup and Sorber (1977). This method requires only few iterations but on the other hand the computation of improvements makes use of the second derivatives of the symmetric functions so each iteration cycle is very time consuming. For example, a test with 40 dichotomous items required about 3 minutes on the IBM 360/65 with this program.

Another method, suggested by among others Martin-Löf (1973) and also presented by Fischer (1974) makes use of a simple switching between the right hand side and the left hand side of the equations (2.1.13). This is the method used in the PML program and it is presented in greater detail below.

A first problem is how to choose start values for the iterations. One simple solution is to put all the $(\hat{\varepsilon}_i)$ equal to unity. Martin-Löf (1973) suggested that start values can be obtained through an approximate solution to the equations (2.1.13) using a linearization in the parameters.

$$(2.1.22) \qquad \log \varepsilon_i = \frac{s_i - \bar{s}}{\sum_{r=1}^{k-1} n_r \frac{r(k-r)}{k(k-1)}} \qquad (i=1,\ldots,k)$$

Both methods of selecting start values are available in the program. Sometimes use of the approximation effects a considerable saving of iterations in comparison with when unities are used as start values, sometimes the approximations has no appreciable effect. There should be no risks involved in using it, however, so it can be regularly applied.

In each new iteration cycle, t+1, new values for the parameters are computed from the previous cycle t as:

$$(2.1.23) \qquad \varepsilon_i^{(t+1)} = \frac{s_i}{\sum_{r=1}^{k-1} \frac{n_r \gamma_{r-1}^{(i)}\{(\varepsilon_i^{(t)})\}}{\gamma_r^{(t)}\{(\varepsilon_i)\}}} \qquad (i=1,\ldots,k)$$

When the absolute difference between $\varepsilon_i^{(t+1)}$ and $\varepsilon_i^{(t)}$ is less than a specified value (in the program it is taken to be .001 but it can be changed at will) for all items the iteration sequence is stopped.

The number of iterations required to a very high degree depends upon the range of parameter values, but for the most part a rather large number of iterations is required (often not less than 100). It has,however, been noted by Fischer (1974 , p. 245; se also Fischer & Allerup, 1968) th the sequence of improvements tends to form terms in a ge etric series and as soon as three iteration steps have been performed several iteration cycles can be saved through extrapolation. (In numerical analysis this extrapolation is known as the Aitken extrapolation, see Dahlquist & Björck, 1974, p. 235).

If we call the estimated parameter values for item i from three succesive iteration cycles $\varepsilon_i^{(t)}$, $\varepsilon_i^{(t+1)}$ and $\varepsilon_i^{(t+2)}$ we get the extrapolated value as:

$$(2.1.24) \qquad \varepsilon_i^{\infty} = \varepsilon_i^{(t+2)} - \frac{(\varepsilon_i^{(t+2)} - \varepsilon_i^{(t+1)})^2}{\varepsilon_i^{(t)} + \varepsilon_i^{(t+2)} - 2\varepsilon_i^{(t+1)}} \qquad (i=1,\ldots k)$$

Two new iteration cycles can then be performed, whereafter the extrapolation can be applied anew.

This method is available in the program and generally it effects a very considerable saving of iterations. However, it is reported by Fischer (1974 , p 245) that the extrapolation may also cause the iterations to diverge. To prevent this

from happening two precautions also mentioned by Fischer have been taken. The first precaution is not to apply the extrapolation on the basis of the results in the first few iteration cycles and the second to set an upper limit as to the amount of extrapolation. In no case have I observed that the Aitken extrapolation using these precautions should cause divergence, só it can probably be regularly applied.

It is impossible to give any generally valid estimate of the number of iterations required for convergence since this to some degree varies from problem to problem. It can be observed, however, that the range of item parameters is of critical importance -- as soon as one or more of the parameters assume high values, a larger number of iterations is required. For those problems in which the proportions of correct answers on the items vary between, say, .10 and .90 convergence is, however, generally obtained within 8-20 iterations. For a test with 40 items, the item parameters can thus often be estimated in less than 20 seconds and for a test with 60 items in a minute or so.

2.2 Estimating person parameters

In estimating the person parameters we could in principle proceed in a similar way as when estimating the item parameters, i.e. through conditioning on item score a conditional likelihood function expressed only in the person parameters can be developed. It can be shown that the equations to be solved can be written:

$$r_v = \sum_{i=1}^{k} \frac{\theta_v \gamma_{s-1}^{(v)}\{(\theta_v)\}}{\gamma_s\{(\theta_v)\}} \qquad (v=1,\ldots,n) \tag{2.2.1}$$

(cf. Fischer, 1974, p. 240)

Unfortunately it is an impossible task to compute the symmetric functions in the θ_v parameters so it is not possible to solve this system of equations.

If, however, it is assumed that the number of persons is large in comparison with the number of items, we can treat the estimates of the item parameters as fixed and estimate the person parameters under this assumption. We then get the following set of equations to solve:

$$(2.2.2) \qquad r = \sum_{i=1}^{k} \frac{\theta_r \varepsilon_i}{1+\theta_r \varepsilon_i} \qquad (r=1,\ldots,k-1)$$

We find that these equations are the same as those appearing in (2.1.4) for UML estimates of the person parameters, except that here the subscript v has been changed to the subscript r, which is possible since all persons having the same raw score must get the same estimated ability.

The equations can easily be solved iteratively using the Newton-Raphson method. In the PML program a routine presented by Fischer (1974, p. 525) is used to do this using the item parameters resulting under the product normation $\Pi\varepsilon_i=1$.

Looking more closely at (2.2.2) we find that in one special case the equations can be solved explicitly and this is when all items are of equal difficulty (i.e. all item parameters are 1). Then (2.2.2) reduces to

$$(2.2.3) \qquad r = \frac{k\theta_r}{1+\theta_r} \qquad (r=1,\ldots,k-1)$$

so

$$(2.2.4) \qquad \theta_r = \frac{r}{k-r}$$

When the range of item parameters is not too great (2.2.4) is used to compute start values for the iterations. This approximation of course gets poorer the more the item parameters vary, so when the difference between the largest and the smallest item parameter on the log scale is greater than 2.0

another approximation presented by Wright & Douglas (1975, p. 22) is used to compute start values. This approximation is based on an assumption of equally spaced item parameters:

$$(2.2.5) \qquad \theta_r = \frac{(1-\exp-\frac{wr}{k})\exp w(\frac{r}{k}-\frac{1}{2})}{1-\exp-w(1-\frac{r}{2})}$$

where $w = \log\varepsilon_{max} - \log\varepsilon_{min}$

From the presentation above it is obvious that with this method of estimating person parameters the only thing that influences the estimates is the distribution of item parameters. The resulting estimates are known to be slightly biased but it should be pointed out that one advantage is gained: as soon as the item parameters are in hand, an ability scale corresponding to the different raw scores is easily constructed (see chapter 5 below).

2.3 The information function and confidence intervals for the parameters

It is possible to determine standard errors for the estimates of the parameters, which are based on the information function with respect to each parameter. The statistical information in the sample with respect to any parameter Π is defined as:

$$(2.3.1) \qquad I(\Pi) = E\{(\frac{\delta \log\Lambda}{\delta\Pi})^2\}$$

where Λ is the likelihood function.

It can easily be shown (see e.g. Fischer, 1974, p. 294 ff) that the information of item i with respect to the person parameter ξ_v is:

$$(2.3.2) \qquad I_i(\xi_v) = \frac{\exp(\xi_v-\sigma_i)}{\{1+\exp(\xi_v-\sigma_i)\}^2}$$

The information of a test (I_t) with respect to the person parameter ξ_v is the sum of the information of each of the k items:

$$(2.3.3) \qquad I_t(\xi_v) = \sum_{i=1}^{k} \frac{\exp(\xi_v - \sigma_i)}{\{1+\exp(\xi_v - \sigma_i)\}^2}$$

Analogously we get the information in the sample with respect to the item parameters $\{I_p(\sigma_i)\}$ to:

$$(2.3.4) \qquad I_p(\sigma_i) = \sum_{v=1}^{n} \frac{\exp(\xi_v - \sigma_i)}{\{1+\exp(\xi_v - \sigma_i)\}^2}$$

Confidence intervals for the item parameters

In the theory of ML estimation it is known that the estimates are asymptotically normally distributed with the standard error $\frac{1}{\sqrt{I}}$. We can thus construct confidence intervals for the item parameters in the usual way:

$$(2.3.5) \qquad \hat{\sigma}_i - z_\alpha \sqrt{I_p(\sigma_i)^{-1}} \leq \sigma_i \leq \hat{\sigma}_i + z_\alpha \sqrt{I_p(\sigma_i)^{-1}}$$

where z_α is the critical value from the normal distribution. In most cases the asymptotic properties of these confidence intervals should be assured since n is usually large.

Confidence intervals for the person parameters

At least when the number of items is larger than, say, 20-30 items, it is possible to determine useful confidence intervals for the person parameters:

$$(2.3.6) \qquad \hat{\xi}_v - z_\alpha \sqrt{I_t(\xi_v)^{-1}} \leq \xi_v \leq \hat{\xi}_v + z_\alpha \sqrt{I_t(\xi_v)^{-1}}$$

From the fact that the information function is a function of ability it is clear that the confidence intervals will be different for different person parameters. Thus, in contrast with the classical psychometric theory, the LT models make no assumption about homoscedastic standard errors of measurement. Some details on how the functions for these standard errors look for different tests are presented in chapter 5.1.

It must be observed that the standard errors are normally distributed only when k is large, so only then can these confidence intervals be trusted. But we have already derived an expression (2.1.7) for the probability of observing a certain raw score, given a person parameter and the item parameters. This expression can of course be used for a straightforward computation of the probabilities of observing each different raw score (including 0 and k) for each estimated person parameter. Such a matrix of probabilities is included in the output from the PML program for $k \leq 30$.

Some comments on how to interpret confidence intervals around person parameters might be in place. Lord and Novick (1968, p. 511-512) stressed that any confidence statement about which region a persons ability falls into can be made with the specified probability only for a randomly chosen person. We can in fact make no confidence statements "about a particular, nonrandomly chosen examinee in whom we happen to be interested. Nor can any confidence statements be made about those examinees who have some specified observed score." (Lord & Novick, 1968, p. 512).

It is a distressing fact that we can have no confidence in confidence statements relating to specified observed scores; for a particularly illuminating discussion of the problems involved the reader is referred to Cronbach, Gleser, Nanda and Rajaratnam (1972, p. 132-134).

The index of subject separation

In some cases there is a need, when the Rasch model is applied, to have a counterpart to the coefficient of reliability in the classical theory, i.e. a measure of the accuracy with which the relative positions of the subjects on the latent trait can be discriminated. Such a measure has been introduced by Andrich and Douglas (1977).

The traditional concept of reliability can be defined:

$$(2.3.7) \qquad r_{xx'} = \frac{\sigma^2_\xi}{\sigma^2_\xi + \sigma^2_\varepsilon}$$

where σ^2_ξ is the variance of true scores and σ^2_ε is the variance of the errors of measurement. From the assumptions that the observed score x_v can be written $x_v = \xi_v + \varepsilon$ and that true scores and errors are uncorrelated, it follows that we can also write the reliability:

$$(2.3.8) \qquad r_{xx'} = \frac{\sigma^2_x - \sigma^2_\varepsilon}{\sigma^2_x}$$

The measure introduced by Andrich and Douglas (1977), called the index of subject separation (ISS), serves as a counterpart to the coefficient of reliability in those cases in which we can obtain direct estimates of the variance of the errors of measurement.

They argued that even though the variance of the errors of measurement varies as a function of ability, the average of the estimated error variances, $\bar{\hat{\sigma}}^2_\varepsilon = \sum_{v=1}^{n} \frac{\hat{\sigma}^2_{\varepsilon_v}}{n}$, can be taken as a reasonable estimate of σ^2_ε in (2.3.8) above. Since the variance of the estimated person parameters (i.e. the counterpart to σ^2_x) is easily computed we have estimates

of all the quantities in (2.3.8) and can directly compute the ISS according to this formula.

This measure tends to give estimates that are highly similar to estimates of the coefficient of reliability with KR_{20} (both measures are given in the PML program) but there are sometimes differences between them (when the sample is severly skewed, for example, the ISS tends to be considerably lower than KR_{20}).

The ISS of course shares with the coefficient of realibility the characteristic of being sample specific but it appears that the ISS has a conceptual advantage. The coefficient of reliability can be low for two reasons; either because the items are heterogenous or because each of the levels of ability is not measured with enough precision because too few items are used. If, however, the ISS is low for a test fitting the Rasch model we can rule out item heterogeneity as a cause and instead concentrate on getting better estimates of each level of ability through adding more items.

Chapter 3

TESTING GOODNESS OF FIT TO THE RASCH MODEL

It has been stressed above thas as a consequence of the rather strong assumptions underlying the Rasch model it is very important that sound procedures for testing goodness of fit are applied.

Several different procedures for testing goodness of fit to the Rasch model have been suggested. Here, some methods based on the CML approach for estimation of item parameters are presented in detail; one graphic method for assessing item fit (Allerup & Sorber, 1977) and two overall numerical tests (Andersen, 1973b; Martin-Löf, 1973). There do exist other more primitive methods for assessing goodness of fit, and some of these are briefly mentioned first.

Since the item parameters, if the model holds, should show no systematic differences if estimated from different subgroups of the sample it is possible to plot such estimates against each other and look for systematic deviations (Fischer, 1974, p. 281 ff). Since the standard errors are estimated too it is also possible to test for each item the difference between the estimates obtained in any two groups. (Fischer, 1974, p. 297-298 and chapter 5.2 below).

Another approach to testing model fit for the Rasch model has been developed by Mead (1976a, 1976b) and Wright and Mead (1977). In this method, estimated item and person parameters are used to predict scores at the item level and from the residuals between observed and predicted scores chi-square-like tests of item fit, person fit and overall fit are developed. However, these tests have unknown asymptotic distributions and simulation studies (Mead, 1976b) indicate that even though the means of the distribution conform to the expected the variances may depart substantially.

Before the tests based on the conditional approach mentioned above are presented, it should be pointed out that a sound application of statistical tests for evaluating goodness of fit implies much more than the choice of a test statistic with known properties. Any inferential method is strongly dependent upon the number of observations made: when the sample is too small even gross departures from the model will be accepted and when the sample is very large even the slightest deviation will cause us to reject the model. The first problem reduces down to one of making enough observations to obtain a reasonable power in the test. Unfortunately the power characteristics of the overall tests are unknown but some simulation studies of this problem will be presented below. The problem that since no model ever holds perfectly true all models will be rejected granted that enough observations are collected has, however, been solved. Martin-Löf (1974a) has introduced a measure call redundancy which on an absolute scale gives a measure of the degree to which the data deviate from the model, which gives a basis for accepting in some cases the model even though the test statistic yields a significant value. This measure is described below in section 3.3.

There are several other questions relating to strategic applications of goodness of fit tests, such as trading relationships between assumptions, item selection procedures, cross validation problems and so on. This type of problems will, however, be discussed at length in chapter 4.

3.1 Testing item fit

Before the overall tests of goodness of fit are presented, methods for evaluating goodness of fit at the item level will be considered. Under the CML approach there exists no statistical test that yields a p-value for the probability of fit of each item. Instead graphic methods are employed. The disadvantage of the graphic methods is that they involve an inescapable element of judgement which, especially until

experience has been accumulated may be quite difficult. But the graphic methods have the important advantage that they are not so strongly influenced as the inferential methods by the sample size: thus deviations not detected by a powerless statistical test may be possible to detect by a graphic method and a statistically significant departure from the model may be judged practically insignificant on the basis of a graphic test.

In investigating fit we do not work with the $((a_{vl}))$ matrix introduced above on page 21 but reorganize it into the item by scoregroup frequency matrix of correct answers, $((n_{ir}))$, in the following way:

		Score group			
		1 ... r ... k-1			
	1	n_{11} ...	n_{1r} ...	$n_{1,k-1}$	s_1
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
Item	i	n_{i1}	n_{ir}	$n_{i,k-1}$	s_i
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	k	n_{k1}	n_{kr}	$n_{k,k-1}$	s_k
		n_1 ...	rn_r ...	$(k-1)n_{k-1}$	

It is obvious that:

$$(3.1.1) \qquad \sum_{r=1}^{k-1} n_{ir} = s_i$$

and recalling that n_r is the number of persons with raw score r we see that:

$$(3.1.2) \qquad \sum_{i=1}^{k} n_{ir} = rn_r$$

The observed proportion of correct answers to item i within score group r is n_{ir}/n_r. We can also compute the predicted proportion of corrects answers to item i for score group r. The conditional probability that a person with raw score r answers item i correctly is the number of answers vectors in which item i is answered correctly divided by the total number of possible answer vectors which add up to r, i.e.:

$$(3.1.3) \qquad P\{A_{vi}=1|r,(\varepsilon_i)\} = \pi_{vi} = \frac{\varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r}$$

Thus, if the model holds true for the data the relation

$$(3.1.4) \qquad \frac{n_{ir}}{n_r} = \frac{\varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r}$$

should hold for all score groups. If we, for a fixed item, plot the observed proportion against the predicted proportion the points should fall along a straight line with a slope of unity. As a function of sampling error the points will of course be spread around the line with unit slope, thus systematic deviations from the predicted proportions along different regions of the abscissa is what is to be looked for.

In the PML program this graphic test is produced as one printer plot for each item with each plot requiring about 1 second of CPU time on the IBM 360/65. Each plot uses one page (or, to be more exact, 54 lines) of printed output.

Even though no statistical test yielding p-values for the fit of each item has as yet been found within the conditional approach it is possible to compute for each score group the probability that an observed frequency of correct answers deviates from what would be expected on the basis of the mo-

del. Under the null hypothesis of model fit the n_{ir} should be distributed binomially $B(n_r, \pi_{ri})$. A two sided test can thus be performed such that when $n_{ir} \leq n_r \pi_{ri}$ the probability of observing n_{ir} or fewer correct answers is computed and when $n_{ir} > n_r \pi_{ri}$ the probability of obtaining n_{ir} or more correct answers is computed, in both cases under the assumption that the null hypothesis holds true.

These tests, too, are available in the PML program but it should be pointed out that the power of these tests is lower than the "power" of the graphic test in the sense that systematic deviations from the model which can be detected with the graphic test are often not detected with the binomial test.

A slightly different version of the binomial test has been presented by Allerup and Sorber (1977) and for computing the cumulative binomial probability distribution a subroutine written be these authors is used.

3.2 Overall tests of goodness of fit

It has been shown by Rasch (1960) that it is possible to devise a test of the model which is completely free from estimated parameters. This test, which is a generalization of the Fisher exact test for a 2x2 matrix, is, however, so computationally cumbersome that it is impossible to put it into practical use.

Thus methods based on estimated item parameters have to be used. This, however, is no great sacrifice since it has been shown by Martin-Löf (1973, 1974b) that certain tests based on ML-estimates are parametric counterparts to generalizations of the Fischer exact test.

There do exist two overall numerical tests of goodness of fit for the Rasch model which are both asymptotically chi-square

distributed. One is a conditional likelihood ratio test independently suggested by Martin-Löf (1973) and Andersen (1973b). Since this test has come to be called the Andersen test the same label will be used here. The other test is a chi-square test computed from a quadratic form suggested by Martin-Löf (1973). This test will be referred to as the Martin-Löf test.

The Andersen conditional likelihood ratio test

Likelihood ratio tests are intimately associated with ML estimation and stated verbally in simple terms the general principle of such tests is to compare values of the likelihood function resulting from parameters estimated under competing hypotheses.

The logarithm of the conditional likelihood function was derived as formula (2.1.11) above and we repeat it here:

$$(3.2.1) \qquad \log\Lambda = \sum_{i=1}^{k} s_i \log\varepsilon_i - \sum_{r=1}^{k=1} n_r \log\gamma_r$$

After having estimated the item parameters for the total sample we can insert the estimated parameter values in (3.2.1) to get the maximum value of the logarithm of the likelihood function. We call the resulting value H_t.

Under the null hypothesis of model fit we should expect essentially the same estimated values of the item parameters whichever subgroup in the sample the estimates are based upon. In the limit we can estimate the item parameters whithin each of the k-1 score groups and still expect the same estimates (within the limits of stochastic variation, of course). If we compute the value of the logarithm of the likelihood function for each of the score groups and call these H_r (r=1,...,k-1) we can form the statistic:

$$(3.2.2) \qquad \log\lambda = H_t - \sum_{r=1}^{k-1} H_r$$

It can be shown that $-2\log\lambda$ is asymptotically chi-square distributed when each $n_r \to \infty$ with $(k-1)(k-2)$ degrees of freedom.

This particular form of the test can, however, seldom be used. Only rarely is the sample size so large that sufficently stable estimates can be obtained within each score group and when there are differences among the item difficulties the simple items tend to be answered correctly by all persons in the higher score groups and the difficult items tend to be answered correctly by no person in the lower score groups, under which conditions it is not possible to estimate the parameters.

However, Andersen (1973b) has shown that the test can be computed also when adjacent score groups are pooled. Thus, if we pool the k-1 score groups into g disjoint groups we can estimate the parameters within each group, compute the $H_j (j=1,\ldots,g)$ and form the statistic:

$$\log\lambda = H_t - \sum_{j=1}^{g} H_j \qquad (3.2.3)$$

Here too $-2\log\lambda$ is asymptotically chi-square distributed when $n_j \longrightarrow \infty$, now with $(g-1)(k-1)$ degrees of freedom.

This test is available in the PML program with an automatic grouping of the score groups. The grouping is carried out under the constraints that there must be a minimum number (m) of examinees within each group (this number can be specified, with the default taken to be m=100) and that there must be no zero or perfect item scores within any group.

The grouping process may fail either as a consequence of choice of too high a value of m or as a consequence of there being items answered correctly by all or no examinees in most of the score groups (or as a consequence of a combination of these two problems). The first problem can of course be easily solved through the choice of a lower m but the second problem can only be solved if those items causing the disturbance are excluded.

The amount of computer time required for computing the test depends upon three factors: the number of items in the test, the number of groups in which the parameters are estimated and the number of iterations required for convergence within each of the groups. There are two reasons for which it is necessary to choice an m so large that the grouping results in only a few groups when k is large. The first reason is that it may be quite time consuming just to estimate the item parameters within the total group when the test consists of many items; if this is to be repeated for a large number of subgroups as well, the costs may become prohibitive. The second reason is that a large number of iterations is often required in groups composed of just the highest score groups. The reason for this is that the proportion of correct answers on the easiest item tends to be very high in these groups in which case the convergence is slow.

Thus, before testing goodness of fit of a long test it is strongly recommended that the $((n_{ir}))$ matric be inspected for a suitable choice of m. In fact, for very long tests it may even be impossible under a strict budget for computer time to apply this test for overall goodness of fit. It should be pointed out, however, that in the first steps of an item selection procedure with the purpose of constructing a undimensional test conforming to the Rasch model, the graphic tests give all the information needed. Only when the final test is to be composed of very many items may an overall test be required. In such a case, however, there is the possibility of constructing the test in parts and then testing whether the parts can be fitted together into one long test, using the procedure described in chapter 5.2 below.

It is also possible to compute the conditional likelihood ratio test for the equality of item parameters between subgroups defined in other ways than through differing raw scores. Each analysis with the PML program namely results in the value of the maximum of the logarithm of the likelihood function being printed, and these values can be used for simple hand calculations. Thus, if separate analyses are made within each disjoint subgroup (boys and girls, for

example) and one analysis is made with all groups merged into one, all ingredients necessary for computing the test statistic (3.2.3) are at hand and only a few arithmetic operations are required. (For an example see chapter 4.2 below).

The Martin-Löf chi-square test

Martin-Löf (1973) has suggested an alternative test for assessing overall goodness of fit to the Rasch model in which a chi-square sum is built up from deviations between observed and predicted frequencies of corrects answers within each score group.

From (3.1.4) above follows that if the model holds true:

$$(3.2.3) \qquad n_{ir} \approx \frac{n_r \varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r}$$

If we label the vector $\begin{pmatrix} n_{1r} \\ \vdots \\ n_{kr} \end{pmatrix} = (q_r)$ and call the corresponding vector of predicted frequencies $\begin{pmatrix} \frac{n_r \varepsilon_1 \gamma_{r-1}^{(1)}}{\gamma_r} \\ \vdots \\ \frac{n_r \varepsilon_i \gamma_{r-1}^{(k)}}{\gamma_r} \end{pmatrix} = (t_r)$ the test statistic can be written:

$$(3.2.4) \qquad T = \sum_{r=1}^{k-1} \{(q_r)-(t_r)\}' \{((V_r))\}^{-1} \{(q_r)-(t_r)\}$$

in which quadratic form $((V_r))$ is a variance- covariance matrix of order kxk with elements defined as follows:

$$(3.2.5) \qquad \begin{cases} \dfrac{n_r \varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} & \text{in the diagonal} \\ \dfrac{n_r \varepsilon_i \varepsilon_j \gamma_{r-2}^{(i,j)}}{\gamma_r} & \text{for } i \neq j \end{cases}$$

Martin-Löf (1973) has shown that the test statistic is asymptotically chi-square distributed with (k-1)(k-2) degrees of freedom when each $n_r \to \infty$.

In (3.2.4) the summation is made over all score groups. If, however, some $n_r=0$ we have to restrict the summation to those R groups in which $n_r>0$. The degrees of freedom then are (k-1)(R-1).

This test requires computation of the second derivatives of the symmetric functions, $((\gamma_{r-2}^{(i,j)}))$. In the PML program this is effected with the Summation algorithm, through repeated calls to this routine with the parameter values for two items at a time put equal to zero.

From (3.2.4) it is seen that at any step in the computations only the $((\gamma_{r-2}^{(i,j)}))$ of one order (i.e. for one score group) are required. With the Summation algorithm, however, the derivatives for all the score groups are obtained, which makes it necessary first to compute the off-diagonal values in the variance-covariance matrices for all the score groups and store these. Since the total number of off-diagonal elements in the variance-covariance matrices is given by the formula $k(k-1)^2/2$ it is easily seen that a vast amount of storage space is needed when the number of items is large. For example, when k=60 816K bytes would be necessary to store

these elements as REAL*8 numbers.

For larger problems as sequential scratch-file is thus used to store the elements. Since it would be rather time consuming to read this file as many times as there are score groups, an array is used in which the information for several score groups is stored. The number of matrices which can be stored in this array depends upon how large it is; it must, however, be dimensioned at least for k(k-1) elements and the larger it is the better. Both in this array and on the scratch file the second derivatives are stored as single precision numbers even though the precision used in the computations is dependent upon whether single- or double-precision arithmetic is chosen.

When the number of items is large the Martin-Löf test tends to be quite time consuming to compute; not only must the second derivatives be computed but the test requires inversion of (at worst) k-1 matrices of the order kxk as well. For example, for k=60 and with all $n_r>0$ the test requires about 7 minutes of CPU-time on the IBM 360/65. When the number of items is moderately large, however, the amount of computer time required is no obstacle against using the test. For k=40 somewhat more than a minute is required and when k=20 the test is computed in less than 20 seconds. In most cases when the number of items is moderate this test is faster to compute than the Andersen test.

The Martin-Löf test vs the Andersen test

Both the overall numerical tests are asymptotically chi-square distributed (they are in fact related through a Taylor expansion), but there may be differences in the power characteristics of the tests and as well as in their asymptotic properties. It should also be noted that while the computation of the Andersen test may fail at times, especially when the sample is small, the computation of the Martin-Löf test almost never fails. But even though the Martin-Löf test can almost always be computed this does not imply that the results

of the test can always be trusted; when a small sample is used and the number of items is large, quite a few score groups will necessarily consist of only a few persons with the consequence that the test statistic may be far from chi-square distributed. In order to cast at least some light on the characteristics of the two overall goodness of fit tests some simulation studies have been performed.

To obtain some information about the difference in the behavior of these tests for smaller sample sizes, data were generated so that they would conform to the model. For generating the scores, a modified version of the routine presented by Allerup and Sorber (1977) was used, with a version of the the feedback shift register random number generator (Lewis & Payne, 1973) as the basic generator[1)]. (It should be pointed out parenthetically that great demands are put on the basic random number generator in these simulations since the tests, and especially the graphic tests, are so sensitive as to be able to pinpoint generators with less than optimal qualities). Data were generated only for k=15 with the size of the item parameters chosen to vary in equal steps between -2 and 2 with the person parameters randomly sampled from a normal distribution with zero mean and unit standard deviation.

Data were generated for two sample sizes, n=150 and n=300, each with 50 replications. The number of observed p-values less than .05 ($N_{.05}$) and the means of the p-values ($\bar{x}_p$) for these analyses are presented in Table 3.1.

---

1) I wish to thank Dr. Philip Ramsey at Hofstra University for putting into my hands an easy-to-use version of this excellent random number generator.

Table 3.1. Results from the two overall goodness of fit tests for data generated to fit the model.

	Sample size			
	150		300	
	$N_{.05}$	$\bar{x}_p$	$N_{.05}$	$\bar{x}_p$
The Martin-Löf test	5	.57	7	.47
The Andersen test	3	.48	2	.48

With 50 replications we should not expect more than 2 or 3 significancies at the 5 per cent level, and this is also what is found for the Andersen test (in all replications two groups were used in computing the Andersen test). But we also find that the Martin-Löf test discards the model at too high a rate for both the sample sizes.

The reason for the difference between the tests is quite obvious when a look is taken at how they are computed. In the Martin-Löf test all score groups are treated regardless of their size (except when $n_r=0$) while in the Andersen test small score groups are pooled to form larger groups. In the present simulations there were of course score groups which contained only one or a few persons.

In the presentation of the results from the Martin-Löf test in the PML program all the independent contributions to the chi-square sum from each score group are, however, printed out and it was noted that in all the cases when this test resulted in a highly significant chi-square sum a very large part was contributed by one or two score groups consisting of only a few persons. It is thus strongly recommended that when this test is applied in situations where the sample is small relative to the number of items, the contributions from the small score groups are investigated, and that the results of this test are put aside as soon as there is a large contribution from any score group consisting of less than, say, 10 persons.

In investigating the power of the tests sets of data were generated under the two-parameter model, with varying values of the discrimination parameter for the items. As previosly, only the case with k=15 was considered, with the item parameters taken to be three each with the values -2,-1,0,1 and 2 and the person parameters chosen in the same way as above. Data were generated to reflect three degrees of deviation from the one-parameter: small, with one third of the parameters 0.9, one third 1.0 and one third 1.1; moderate, with the discrimination parameters chosen to be 0.7, 1.0 and 1.3; and finally large, with the corresponding discrimination parameters chosen as 0.5, 1.0 and 1.5. In all cases the three discrimination parameters were represented at all the five levels of item difficulty.

Three different sample sizes were used; 150, 300 and 1 000 and 10 replications were made. The results are presented in Table 3.2.

Table 3.2. Results from the two overall tests for data generated to deviate from the

	Amount of deviation					
	Small		Moderate		Large	
	$N_{.05}$	$\bar{x}_p$	$N_{.05}$	$\bar{x}_p$	$N_{.05}$	$\bar{x}_p$
n=150						
The Martin-Löf test	2	.41	3	.36	5	.15
The Andersen test	1	.42	2[1)]	.27	8[2)]	.01
n=300						
The Martin-Löf test	0	.66	4	.20	7	.07
The Andersen test	0	.47	6	.10	10	.00
n=1 000						
The Martin-Löf test	0	.50	9	.01	10	.00
The Andersen test	1	.34	10	.00	10	.00

1) The Andersen test could be computed in only 9 cases.

2) The Andersen test could be computed in only 8 cases.

We find that when there are only small deviations from the one-parameter model there is no possibility with the sample sizes used here to detect any deviation from the model (it will be shown below that even though highly significant values of the test statistics are obtained when the sample size is heavily increased there would still be reason to accept the model with this amount of deviation in the data).

With large deviations from the model we find that the Andersen test in all successful analyses, for all the sample sizes, discards the model, while the Martin-Löf test discards the model only for the sample size 1 000 in all analyses. At least for deviations from the model caused by varying discrimination among the items the power of the Andersen test thus appears to be greater than the power of the Martin-Löf test.

For the intermediate case with medium deviations we do find indications, too, that the Andersen test is more powerful than the Martin-Löf test but it can also be noted that only for the largest sample does the former test consistently discard the model.

Even though these simulations are merely some examples it does seem as if the conclusion can be drawn that the likelihood ratio test has somewhat better properties than the chi-square test both with respect to the number of observations needed to claim that the test has the assumed distribution and with respect to power.

3.3 Redundancy

No model is ever completely true in describing a set of data, which means that with a sufficient number of observations any goodness of fit test would discard the model. In discussing this problem Martin-Löf (1974a) stated:

> "This indicates that for large sets of data it is too destructive to let an ordinary significance test decide

whether or not to accept a proposed statistical model, because, with few exceptions, we know that we shall have to reject it even without looking at the data simply because the number of observations is so large. In such cases we need instead a quantitative measure of the size of the discrepancy between the statistical model and the observed set of data... " (p. 3).

Martin-Löf derived such a measure called redundancy (R) from concepts in the statistical information theory, which on an absolute scale measures the deviation between a statistical model and a set of data. The redundancy exists in two forms: the micro-canonical redundancy corresponding to non-parametric formulations of the test and the canonical redundancy corresponding to parametric formulations. The canonical redundancy, which is of course the only one that is accesible in tests of the Rasch model, should be regarded as an approximation to the microcanonical redundancy and both can be given the same interpretation:

"it is the relative decrease in the number of binary units needed to specify the given set of data when we take into account the regularities that we detect by means of the exact test"(Martin-Löf, 1974, p. 10).

Since the measure reflects a relative decrease it assumes values between 0 and 1 and low values indicate a good fit between the model and the data.

The canonical redundancy can easily be computed from the likelihood ratio quotients (3.2.2) or (3.2.3) above together with the maximum of the logarithm of the likelihood function (3.2.1):

$$(3.3.1) \qquad R = \frac{\log\lambda}{H_t}$$

It is also possible to compute R from the Martin-Löf chi-square test, which gives an approximation for R in the formula above:

$$(3.3.2) \qquad R = -\frac{\chi^2}{2H_t}$$

Since the scale upon which R is expressed is in a sense absolute it is possible to use case studies for calibrating it. Martin-Löf (1974a) computed the values of R for different values of the binomial probability p with respect to the hypothesis p=.5, with the results presented below:

p		R	Fit
.000	1.000	1.	Worst possible
.216	.684	.1	Very bad
.441	.559	.01	Bad
.482	.518	.001	Good
.494	.506	.0001	Very good

It might be of some interest to compare this calibration of the redundancy scale with the results which can be observed for R when very large sets of data with known deviations from the model are generated. Data have thus been generated under the two-parameter model with different values of the discrimination parameter for the items. In all analyses 15 items were used with the same 5 levels of difficulty parameters as in the simulations investigating power presented above. The sample size was 50 000 persons (N(0,1)) and three different discrimination parameters all represented at all levels of difficulty were used.

Case	Discrimination parameters			The Andersen test			The Martin-Löf test		
				χ^2	df	R	χ^2	df	R
1	1.00	1.00	1.00	185.2	182	.0003	188.0	182	.0003
2	.95	1.00	1.05	247.8	168	.0005	254.2	182	.0005
3	.90	1.00	1.10	455.4	154	.0008	494.7	182	.0009
4	.85	1.00	1.15	933.2	182	.0017	933.6	182	.0017
5	.80	1.00	1.20	1504.7	182	.0028	1498.7	182	.0028

In case 1, using data fitting the model, we find non-significant values of the test statistics and the redundancy indicates a "very good" fit. In all the other cases the statistical tests are very highly significant but at least for some of them the value of R is low enough to indicate an acceptable fit.

For case 2 the value of R is .0005, which on the scale established by Martin-Löf corresponds to a fit that is "good" to "very good". The graphic and binomial tests of the items in this analysis showed no signs of systematic deviations from the model and would thus have been useless to improve the fit. (Had a plotting method yielding greater accuracy been used, such as the one in the Allerup & Sorber, 1977, program it might of course have been possible).

For case 3 the valus of R indicate a "good" fit. Here, however, the graphic tests could be used to identify all the deviating items. Here there is thus a choice of whether to improve the fit through selecting items, or to accept the fit as satisfactory.

The other cases all show a fit which is worse than "good" and in all these analyses both the graphic and the binomial tests could clearly be used to identify the deviating items.

The results obtained in case 2 show that it is possible to observe a highly significant deviation from the model with an inferential test while at the same time it is impossible to find any deviations with descriptive methods. If in such a case the redundancy is suffiently low, less than .001 say, we have a good basis for accepting the model in spite of the significant test statistic.

If the redundancy is low and it is possible to use the results from the graphic tests to improve the fit we have the choice of doing so or to accept the model as showing a good fit to the data. In making this decision it does seem necessary to invoke other than statistical criteria, such as content related considerations.

In order to prevent any misunderstanding to occur it should finally be pointed out that the redundancy statistic is of any interest only when the number of observations is large; a high redundancy observed for a smaller sample is not necessarily a sign of a poor fit.

Chapter 4

CONSTRUCTING RASCH SCALES

It has repeatedly been stressed that the Rasch model is the LT model which entails the strongest assumptions, and even though no model is ever wholly valid for describing a set of data, serious deviations from the assumptions will invalidate most attempts to capitalize on the great potentialities for applications in the model. Thus, whatever eventual application is intended, one inevitable first step is to make sure that the data do show a reasonable fit to the model, and if they don't, take the necessary precautions to make sure that they do.

In the introduction it was mentioned that the Rasch model has already been applied to some extent and surely some experience has accumulated as to possible sources of threats to the model. But it must also be stressed that in the applications carried out on the European continent as well as in North America the problems of testing goodness of fit have been taken rather lightly, which is almost surely a consequence of the fact that the procedures employed for testing goodness of fit have less than optimal properties. In fact, there are very few studies where the test procedures developed on the basis of the conditional approach have been used for other than illustrative purposes.

In this context neither will it be possible to present much more than illustrations of applications but the important point to note is that there is still much research to be carried out on the sources of deviations from the model and how to remedy them.

Before the possible sources of deviations from the model are discussed, analyses of two tests of PMA-type (Primary Mental Abilities), develop within the framework of classical test theory will be presented.

4.1 Analyses of two tests of PMA-type

The two tests to be analyzed are Number Series and Opposites constructed to measure inductive (or non-verbal reasoning) and verbal ability respectively. The tests were constructed by Svensson (1964, 1971) and the only reason for chosing these tests was simple access to data which consist of a sample of 566 fifth-graders (see Gustafsson, 1976, for a detailed account of why and how the data were collected).

Each of the 40 items in Number Series consist of a series of six numbers and the task is to add the two following numbers. The time limit of the test is 18 minutes.

In Opposites, which test also consists of 40 items, the task is to select from among four given words the one which is the opposite of a given word. This test too is timed, with the limit being 10 minutes

Opposites is thus a multiple-choice test which allows guessing and can for this reason alone be supposed to show a poor fit to the model. But it is of course of some interest to investigate in what ways this kind of violation of model assumptions expresses itself in the model tests. Number Series, in contrast, requires constructed responses which means at least that guessing is minimized as a source of deviation from the model. Rasch (1960) who also investigated the fit of some previously existing tests to the model in fact found, with graphic methods, a good fit for a test highly similar to Number Series.

Number Series

With a sample of 566 persons and a test with 40 items it is obvious that quite a few score groups will be very small; an attempt was thus made to use the Andersen test to investigate the overall fit of the test. This test could, however, not be computed for the original set of items since easy items were solved by all persons in almost all the score groups, excepting only some of the lowest while two of the most difficult

items were solved only by a few persons in some of the highest score groups. When the five easiest and the two most difficult items were excluded, however, the score groups could successfully be grouped into four groups, with the value of the test statistic being 349.2 with 96 degrees of freedom, which is of course highly significant.

Thus, if we hade hoped to find a good fit for the Number Series test to the Rasch model, there is reason for disappointment. But on the other hand it will be instructive to find out the reasons for the poor fit of this test.

Several factors may, singly or in combination, be responsible for the poor fit: item heterogeneity, speededness of the test, learning effects from one item to another, varying item discriminations, just to mention a few. In searching for the cause or causes to the deviations, the information which is of most help is the graphic test of each item, along with, of course, the content of each item and every piece of information about the testing situation which can be found.

The items in the test have been analyzed and the recursive formulas defining the series have been determined. These algorithms are presented in Table 4.1 along with the proportions of correct answers and rough summaries of the graphic tests in which for the lower and the higher score groups + and - signs have been used to indicate whether the observed proportion of correct answers is higher or lower than the predicted proportion.

Table 4.1. The recursive formula defining the items in the Number Series test.

Item	Prop. corr.	Algorithm a_{n+1}=		Low score groups	High score groups
2	.97	$a_{n-1}+1$	$a_1=1,\ a_2=1$		
3	.98	a_n-1	$a_1=9$		
4	.98	a_{n-1}	$a_1=1,\ a_2=4$		
5	.97	a_n-2	$a_1=18$		
6	.93	a_n+2	$a_1=3$		
7	.91	a_n-3	$a_1=24$	-	
8	.87	a_n-4	$a_1=29$	-	
9	.85	a_n+7	$a_1=10$	+	-
10	.77	a_n-7	$a_1=51$	+	-
11	.63	$a_{n-1}+2$	$a_1=2$, n=1,3,5...		
		a_{n-1}	$a_2=2$ N=2,4,6...	+	-
12	.56	$a_n \cdot 2$	$a_1=2$		
13	.67	$a_{n-1}+3$	$a_1=7,\ a_2=8$	-	-
14	.57	$a_{n-1}+5$	$a_1=5,\ a_2=7$	-	-
15	.64	$a_{n-1}+1$	$a_1=11,\ a_2=8$	-	
16	.54	$a_n \cdot 2$	$a_1=5$	-	
17	.50	a_n+n-1	$a_1=2$		-
18	.51	$a_{n-1}-4$	$a_1=22,\ a_2=21$		
19	.46	a_n+n+2	$a_1=3$		
20	.49	$a_{n-1}-5$	$a_1=19,\ a_2=17$		
21	.43	$a_n \cdot 2$	$a_1=3$		-
22	.41	$a_{n-1}-1$	$a_1=12,\ a_2=13$		-
23	.41	a_n-10+n	$a_1=43$	-	
24	.45	$a_{n-1}+9$	$a_1=5,\ a_2=11$	-	+

Table 4.1 Continued

Item	Prop. corr.	Algorithm $a_{n+1}=$		Low score groups	High score groups
25	.28	$a_n+2(n-1)$	$a_1=5$	+	-
26	.35	$a_{n-1}-2$	$a_1=34,\ a_2=29$	-	
27	.31	$a_{n-1}+3$	$a_1=17,\ a_2=15$	-	+
28	.40	$a_{n-1}\cdot 2$	$a_1=6,\ a_2=12$	-	+
29	.35	$a_{n-1}/2$	$a_1=128, a_2=64$	-	+
30	.25	$a_{n-2}-5$	$a_1=20,\ a_2=18, a_3=16$	-	+
31	.28	$a_{n-1}+2$	$a_1=1,\ a_2=4$	-	+
32	.29	$a_{n-2}+5$	$a_1=1,\ a_2=3,\ a_3=5$		
33	.32	$a_{n-1}+1$	$a_1=1,\ n=1,3,5$		
		$a_{n-1}-1$	$a_2=2,\ n=2,4,6$		
34	.20	a_n+1	$n=2,5,8\ldots\ a_1=1$		
		$a_{n-1}+1$	$n=4,7,10\ldots$	+	
		9	$n=3,6,9\ldots$		
35	.13	$a_{n-2}+16$	$a_1=13,\ a_2=15, a_3=22$	-	+
36	.17	$a_{n-1}+2$	$a_1=1,\ a_2=2,\ a_3=3$		+
37	.09	$a_{n-3}+2$	$a_1=3\ \ n=4,7,10\ldots$		
		a_n+1	$n=2,5,8\ldots$		
		$a_{n-1}+a_n$	$n=3,6,9\ldots$		+

Looking at the pattern of deviations from the model as evidenced by the graphic tests we find that for most of the items late in the test the observed proportion is too high for the higher score groups and too low for the lower score

groups. If the items appearing late in the test have a higher discrimination parameter such a pattern of results would be found, but there are other explanations as well of which speededness of the test appears to be most reasonable. For the items with order numbers around 30 almost half the sample did in fact not attempt any answer, correct or incorrect, which is a strong indication that a large proportion of the sample did not even attempt to solve the items appearing later in the test. Additional evidence in favor of this interpretation is obtained from the algorithms for the items. The recursive formula for items 27 and 31 are in fact essentially the same as those for items 13 to 15, for example, and still the items appearing early have proportions of correct answers which are almost twice as large as those for the items appearing later in the test. This must be regarded as a very strong indication that the test is speeded in the sense that, if given additional time, some persons would get additional items correct (or for that matter, that there may be some other reason, such as boredom, accounting for why some of the examinees did not attempt the items later in the test).

If speededness or some other factor with equivalent effects, is the only reason for the poor fit of the whole test, we should expect a good fit for items placed early in the test. Since omitted responses were coded in a special way it has been possible to determine the proportion of omitted responses for each item and this proportion was found to be fairly low, never exceeding 20%, for item 22 and earlier items, while there was a rather rapid increase in the proportion of omitted responses for the items from number 23 to the end of the test.

A new analysis was thus performed including only items 2-22. This analysis too resulted in a highly significant χ^2-value of 57.6 with 20 degrees of freedom (the Andersen test with the score groups grouped into two groups). Again the graphic tests of the items were resorted to and these indicated a poor fit for items 9, 10 and 11, with the fit being worst for item 11. For the higher score groups there was for this item a too low observed proportion of correct answers and for

the lower score groups the observed proportion was to high. Just a glance at the recursive formula for this item (see Table 4.1) is sufficient to show that it deviates from those for the other items early in the test in that it defines two intertvined series defined by different rules. Obviously this item measures at least partly an ability which is different from the ability measured by the other items in the early part of the test.

The graphic tests for items 9 and 10 gave a pattern very much like that found for item 11, but less pronounced. The algorithms for these two items are the same as for those four items immediately preceeding them. What obviously makes items 9 and 10 more difficult and also showing a poor fit is that they pose requirements for arithmetical ability: they require computation of expressions like 45-38, which is a task which pupils in the fifth grade have a high probability of failing (Kilborn & Johanson, 1976. It can parenthetically be mentioned that when the second author above was asked to identify those items in the early part of the test posing exceptional demands for arithmetic skill, items 9 and 10 were clearly identified and a few more with some doubt). Thus we can draw the conclusion that the reason why items 9 and 10 do not fit together with the other items is multidimensionality of the latent space, i.e. performance on these items is affected by arithmetical skill in addition to the ability measured by the other items.

A new analysis was performed in which these three items were excluded with the result that the Andersen test gave $\chi^2=28.4$ with 17 degrees of freedom, with a corresponding p-value of .04, which will here quite arbitrarily be regarded an acceptable fit.

In passing it can be mentioned that the Martin-Löf test for the same items resulted in a very highly significant value of the test statistic ($\chi^2=763.3$, df=272). A very large part of the χ^2-sum (457.6) was, however, contributed by score group 2, consisting of one single examinee who had answered items 15 and 20 correctly. The results from this test must

thus obviously be set aside (cf page 55 above).

Even though the overall test indicates that an acceptable fit was finally obtained, the graphic tests could be used to select a still more homogenous item set or perhaps several item sets. There were, for example, some indications that ascending and descending series gave slightly different results. However, since we in this case are restricted to a very limited set of items there is but little to be gained from pursuing such analyses.

In conclusion, we have thus learned that unless a reasonable number of examinees have attempted the items and unless influence from other abilities is not controlled for, the data will not fit the model. But it should also be pointed out that we have made a heavy selection among the items and have thus to some degree capitalized on chance effects. Thus, the fit of a set of items selected from a larger pool on the basis of the results in one sample should be tested in another sample, for purposes of crossvalidation.

Opposites

Analysis of the items in Opposites with the Andersen test resulted in $\chi^2=333.4$ with 117 df, which is of course highly significant.

Table 4.2 presents gross summaries of the graphic tests of the items (the first two items have been excluded since they were answered correctly by almost all persons; thus little information is gained by keeping them, but as soon they are included in an analysis a large number of iterations is required for convergence). As before the method of marking too high and too low observed proportions of correct answers for lower and higher score groups with + and - signs has been used. When looking at the pattern of signs it should, however, be kept in mind that they represent a very simplified description, there sometimes being important differences between the plots for items with the same pattern of signs.

Table 4.2. Summary of the graphic tests of the items in Opposites.

Item	Prop. corr	Score groups low	high	Item	Prop. corr.	Score groups low	high
3	.98			22	.60	-	+
4	.97			23	.51	-	+
5	.89	-	+	24	.37	+	-
6	.90	-	+	25	.44	+	-
7	.71			26	.31	+	-
8	.71	-	+	27	.32	+	-
9	.75	-	+	28	.19		
10	.80	-	+	29	.39		
11	.58			30	.39	+	-
12	.69	-	+	31	.16	+	-
13	.72	-	+	32	.33		
14	.69	-	+	33	.25	+	-
15	.66	-		34	.27	+	-
16	.56			35	.22	-	+
17	.60			36	.19	+	-
18	.47	-	+	37	.22	-	+
19	.69	+	-	38	.11	-	+
20	.53	-	+	39	.25	+	-
21	.38			40	.22	+	-

Nevertheless the deviations form quite a clear pattern: for the items late in the test, which are also the more difficult ones, there tends to be a too high observed proportion of correct answers for the lower score groups and a too low proportion for the higher score groups, while the reverse pattern of deviations is found for the easier items. This is exactly the pattern to be expected when a test permits guessing: on the difficult items the examinees with low ability will get scores which are too high by guessing, with the consequence that their ability is overestimated which in turn implies that on the easier items where the proportion of the sample which guesses is smaller, the low ability examinees will appear to perform too poorly.

What is perhaps more interesting than this general pattern is that there, nevertheless, are items which do not conform to it: Some of the most difficult items do not appear to be affected by random guessing and there are in fact a few items (35,37 and 38) with a very low proportion correct (lower actually than would be expected if all the examinees guessed randomly) on which the - + rather than the + - pattern is observed. It may be interesting to take a closer look at these items which are "good" items in the sense that if it were possible to estimate the discrimination parameters, this parameter would be found to be high for these items.

The three items are presented in Table 4.3 together with the percentage of subjects marking each alternative.

Table 4.3. Three difficult, highly discriminating items in Opposites.

Item	Stem	Choices 1	2	3	4	No resp
35	Foolhardy	Cautious(22)	Attentive(35)	Foolish(18)	Daring(16)	(9)
37	Significant	Unclear(14)	Unimportant(22)	Despised(13)	Meaningless(41)	(10)
38	Ample	Poor(63)	Impoverished(9)	Magnificent(8)	Scanty(11)	(9)

What is especially striking, particularly for items 37 and 38, is the high percentage of examinees marking one of the distractors. If the content of the items is looked at, it does become obvious, however, why one of the incorrect alternatives is so attractive. In item 37 the majority of the examinees have chosen "meaningless" the opposite of "significant". I suspect that even in English this distractor would be quite attractive, but it must be so to an even higher degree in Swedish since the Swedish counterpart of significant can be literally translated as "meaningful". Obviously, many of the examinees, not knowing the exact meaning of the words, were fooled by their literal appearances to choose this particular distractor.

The same explanation holds true for item 38, even though this is here less clear from the translation into English. The literal translation from Swedish into English of ample is, however, "richlike" which makes it understandable why more than 60 per cent of the sample chose "poor" as the opposite to ample.

These examples provide an explanation of why some multiple choice items don't show evidence of any guessing effects: if one (or more) of the distractors is so attractive that almost all of those who don't know the correct answer chose it, of course little or no random guessing will take place (cf. Lord, 1974a and page 8 above). From this, it follows that it is at least in principle possible to construct multiple choice tests where guessing will only to a small degree be another factor affecting performance. Whether it is possible to construct such a multiple-choice test in practice is of course more doubtful, and is probably not worth the attempt.

We will now embark on an excersise intended to serve above all as a warning: The usual practice in item screening to obtain fit to the model is to try out a larger set of items on a sample and select those that appear to fit the model. We will investigate whether this is possible here, in which case we know that such a procedure can yield only essentially meaningless results as a consequence of the fact that all the items are of multiple choice type and thus are influenced by guessing.

There are essentially three types of items to be found in Table 4.2: those with no signs marked, those with the + - pattern and those with the - + pattern, corresponding to items with intermediate, low and high discrimination, respectively (the items will be referred to as MD, LD and HD items). It could be argued that those items without any signs marked are those that should be selected since they do not show any deviation from the model. Not much thought is required, however, to detect that this is incorrect: items do not show fit or lack of fit to the model in themselves, the model

assumption instead says that the items should be homogeneous, i.e. that each item should fit together with the other items.

This implies that if we analyze the three groups of items separately, we should expect to find three sets of items which each form a scale conforming to the requirements of the Rasch model.

Such analyses have been performed using every item listed in Table 4.2 except item 15 since the results of the graphic test for this item did not conform to the results of any other item. The results from the goodness of fit tests (the Andersen test) are presented below:

Type of item	Number of items	χ^2	df	p	ISS
LD	12	33.0	33	.47	.27
MD	10	19.8	27	.84	.41
HD	15	49.6	42	.20	.72

We thus find that the reasoning was correct; for each set of items a good fit is found. It can also be observed that the ISS (see page 41) is considerably higher for the HD than for the LD items.

Two conclusions can be drawn from this exercise. First, the question of item fit is wrongly stated if it is asked whether an item does or does not fit the model, the correct question to ask is whether any given item fits together with the other items. This implies in turn that in most cases analysis of an item pool should not result in the selection of a subset of items which are "good" in relation to the requirements of the model, instead grouping of items into internally homogenous scales is the result to be sought, and throughout, of course, attempts should be made to clarify what each scale is measuring.

The second conclusion to be drawn is purely negative: Obviously it is very easy to select items from a pool so as to form scales conforming to the model, but in this case it is almost equally obvious that the result is nonsensical, since if these

groups of items were administered to a new sample with only a slightly different distribution of person parameters a poor fit would be found. What has been done can probably best be described as a capitalization on trading relationships between assumptions; for example the amount of guessing is different on the items and the discrimination can be supposed to vary. These two factors can blend and balance in different ways for different items, with the net result being that items which are very different in both these respects can be found to fit together.

4.2 Item bias in Opposites

The overall numerical tests as well as the graphic tests are constructed from the starting point that the item parameters should remain the same for all levels of score groups and evidently these tests are powerful means of guarding against violation of certain kinds of model assumptions such as varying item discrimination. The Rasch model, however, states that the item parameters shall be the same whichever subdivision of a sample is made, and the tests based on the results for groups with different levels of performance need not be powerful when some items are too easy for one subgroup and too difficult for another if the overall level of performance of the groups is equivalent.

This problem of analysis of what has been termed item bias will be illustrated through analyses of sex differences in Opposites.

As was mentioned above on page 50 the Andersen test (equation 3.2.3) can be used to test differences between the estimates of item parameters obtained in any disjoint grouping of the sample through performing some simple hand calculations of figures found in the computer printout, i.e. the maximum of the log likelihood.

The item parameters were first estimated separately for boys and girls for items 3-40 in Opposites with the resulting va-

lues of H_j being -4 905.31 and -4 834.16 respectively. Those values, together with the value of H_t of -9 795.40 found with both groups pooled were put into formula (3.2.3) with a resulting χ^2=111.86 with 37 df, which is highly significant.

Had there been large differences in the level of performance of boys and girls it could have been argued that the significance does not reflect anything except the kind of deviations already detected with the overall goodness of fit test. Since this is not the case (even though the mean for boys is slightly higher) we can go on to study which items tend to favor boys and girls respectively.

Since estimates of the normally distributed standard errors are obtained along with the estimates of the item parameters in each analysis a z-test for the difference between the parameters for each item can easily be computed (Fischer, 1974, p. 297):

$$(4.2.1) \qquad z = \frac{\sigma_{i1} - \sigma_{i2}}{\sqrt{SEM_{i1}^2 + SEM_{i2}^2}} \qquad (i=1,\ldots,k)$$

where the subscripts 1 and 2 refer to the groups.

The item parameters, along with the results from the statistical test, are presented in Table 4.4. A negative sign of z indicates a lower value of the item parameter for boys, i.e. that the item is easier for boys. There are four items for which a significant difference in favor of boys is found and for four items a significant difference is found in favor of girls.

The words which are "too easy" for boys are "spurt", "attack", "noble", and "foolhardy" and the words which are "too easy" for girls are "smooth", "desert", "merry" and "anonymous". This is not the place to venture into discussion why certain items are biased in a certain way, but at least for the "boys items" it does appear as if they are related to what is regarded as boys' activities (cf. Wernersson, 1977).

Table 4.4. Tests of equality of estimated item parameters for boys and girls for item 3-40 in Onposites.

Item	Item parameters Boys	Girls	z	Item	Item parameters Boys	Girls	z
3	-4.30	-4.07	-.36	22	-.16	-.58	2.23[X]
4	-3.16	-3.78	1.36	23	.03	.07	-.24
5	-2.26	-2.22	-.16	24	.39	1.03	-3.40[X]
6	-2.14	-2.50	1.25	25	.47	.26	1.15
7	-1.05	-.75	-1.56	26	.94	1.00	-.29
8	-1.34	-.52	-4.03[X]	27	.92	.90	.10
9	-1.01	-1.26	1.21	28	1.60	1.69	-.41
10	-1.27	-1.62	1.62	29	.73	.44	1.59
11	-.20	-.33	.72	30	.63	.57	.33
12	-.72	-.93	1.04	31	2.04	1.82	.92
13	-.91	-1.00	.45	32	.98	.74	1.22
14	-.60	-1.06	2.37[X]	33	1.20	1.45	-1.21
15	-.48	-.87	2.07[X]	34	1.10	1.34	-1.16
16	-.21	-.14	-.42	35	1.18	1.85	-3.07[X]
17	-.51	-.20	-1.70	36	1.53	1.91	-1.66
18	.49	-.04	2.91[X]	37	1.44	1.59	-.68
19	-1.27	-.44	-4.15[X]	38	2.49	2.22	.96
20	-.02	-.04	.11	39	1.32	1.36	-.20
21	.75	.51	1.32	40	1.40	1.59	-.88

It will be recalled that in the analyses performed on Opposites in the previous section, it was possible to divide the items into three groups, within each of which a good fit to the model was observed. It can be asked how this grouping of the items is related to sex bias. According to the signs of the z-test presented in Table 4.4 each item in the three groups was classified according to whether it tended to be biased in favor of boys or girls. The results are presented below:

Item type	Tendency towards bias in favor of Boys	Girls	Total
LD	8	4	12
MD	5	5	10
HD	5	10	15
	18	19	37

There is a correlation: the items which were identified as having a low discrimination tend to be biased against girls, while the HD items tend to be biased in favor of girls. But on the other hand a closer scrutiny of Tables 4.2 and 4.4 reveals that two of the items significantly favouring boys are of the LD type, while the other two are of the HD type. There is thus a considerable heterogeneity within the groups of items with respect to sex bias, which in turn implies that the three scales previously found to fit the model may well have to be discarded on the basis of an analysis of sex differences.

Three separate analyses have thus been performed in which the Andersen test was used to test sex differences for each of the three scales:

	χ^2	df	p
LD	28.1	11	$p<.01$
MD	14.0	9	ns
HD	57.7	14	$p<.001$

For two of the scales there are significant differences between boys and girls with respect to the item parameters in spite of the fact that the overall goodness of fit tests did not indicate any reason for discarding the model. Thus, even though one test shows a good fit another can show a poor fit, which is of course due to the fact that the tests have differential power of detecting different deviations from the model.

It is of some interest to study the distribution of person

parameters for boys and girls on the three sub-scales. Martin-Löf (1973) has presented two tests for comparison of the distributions of the person parameters for two groups. One of the tests is a likelihood ratio test and the other is a χ^2-sum. If we use the subscript e (e=1,2) to denote the groups we can write the likelihood ratio test:

$$(4.2.2) \qquad \log\lambda = -\sum_{e=1}^{2}\sum_{r=0}^{k} n_{re} \log \frac{n_{re}}{n_e} + \sum_{r=0}^{k} n_r \log \frac{n_r}{n}$$

Since the index r here varies from 0 to k the test has k degrees of freedom, and as before $-2\log\lambda$ is asymptotically chi-square distributed as $n_r \to \infty$.

The χ^2-test is computed according to the following formula:

$$(4.2.3) \qquad \chi^2 = \sum_{r=0}^{k} \frac{n_1 n_2}{n_r}\left(\frac{n_{r1}}{n_1} - \frac{n_{r2}}{n_2}\right)^2$$

This test, too, has k degrees of freedom and is asymptotically chi-square distributed under the same condition as the likelihood ratio test.

Application of these tests require that there be no difference between the groups among the item parameters so in this case they can be strictly used only for the MD items. Nevertheless, the tests were used on all three scales and the results are presented below:

Item type	Likelihood ratio test	χ^2-test	df	Means of raw scores Boys	Girls
LD	34.9*	33.2*	12	4.16	3.55
MD	12.0	11.9	10	5.77	5.61
HD	14.5	8.2	15	8.89	8.76

The two tests give highly similar results except for the HD items and they both agree that only for the LD items is there a significant difference among the distributions of

item parameters. On this scale the boys have a higher mean than the girls, which is also the case for the other two scales even though the differences on the latter are considerably smaller.

According to the finding reported above that there is a correlation between the sex bias and item discrimination, we might perhaps have expected to find a difference in favor of the girls on the HD items, which is obviously not the case. Three explanations can be put forth to account for this. First it should be noted that even though some items may be found to be biased against one group there is nothing that says that this group will have a lower mean on these items since the mean is also affected by the distribution of person parameters. Second it must be observed that if there are in the test a few items which are severly biased against one group, several of the others will appear to have at least a small bias in favor of this group which follows from the fact that the constraint must be imposed that the parameters shall sum to zero. As the third explanation, it can be pointed out that even though we in this case have interpreted the differences between the groups in terms of the sex variable, there is evidently in this sample a correlation between sex and ability. This implies that the division of the sample according to sex to some extent is confounded with level of performance which in turn implies that the correlation observed between "sex bias" and item discrimination may also be accounted for by differences in level of performance.

Let us now summarize some of the conclusions which can be drawn from these analyses. First of all it can be concluded that the Rasch model can be used to study item bias, both as nuisance in measuring devices and as a substantive area of research. (In order to prevent any misunderstanding from arising it should perhaps be pointed out that if all items in a test to the same degree favor one special group this will not be detected as any deviation from the model assumptions. Here the problem rather is one of definition of the ability being measured.) The model adds two aspects to the study of item

bias which are quite unique: the first that there exists an overall test of item bias, which is also supplemented by tests at the item level and the second that differences between the groups with respect to the person parameters do not influence the results, at least not when there is no differences among the item parameters as a function of ability.

A second important conclusion to be drawn is that the overall numerical test of goodness of fit presented in chapter 3.2 have a low power of detecting certain threats to the model; in this case multidimensionality of the latent space since sex is an additional factor to ability which systematically affects performance on some items. The implication is of course that even though the overall test of goodness of fit is not significant there may be a need to carry the investigation further by division of the sample along other lines than level of performance.

4.3 Discussion

It must be stressed that the analyses of Number Series and Opposites presented above are nothing but examples, which can serve to highlight a few of the characteristics of the Rasch model. In the analyses several scources of threats to the model have been pin-pointed and we will first discuss these and a few more common violations of model assumptions. After that strategies and problem in the development of scales fitting the model are discussed.

Scources of threat against the model

Item heterogeniety is a violation of the assumption of unidimensionality and, as was pointed out in chapter 1.2 above, there is no entirely satisfactory method with which one can make sure that this assumption is not violated before the Rasch model is applied. But it does appear as if the goodness of fit tests (and especially the graphic tests of the

items) are powerful means with which item heterogeneity can be detected. In the analysis of Number Series, for example, the heterogeneity caused by some items requiring more arithmetical ability than others was easily detected. It thus appears that the model in itself is a very useful tool for studying unidimensionality of measurements.

It must be strongly emphasized that the question of item homogeneity is a question of finding items measuring the same ability, and not a question of excluding items not fitting the model. This implies among other things that purely statistical criteria cannot be used in selecting items and that a very clear grasp of the content and the processes required of each items is demanded.

Speededness of the test is obviously a violation of the model assumptions since if a person does not have time to read an item any statement about the probability of a correct answer as a function of the person parameter will be meaningless. None of the LT-models considered here can thus be supposed to properly represent the case when there is any amount of speededness involved. It can be pointed out, however, that Rasch (1960) has proposed a Poisson process model for one particular case where speed is involved, namely tests of oral reading speed.

For almost all group tests of ability there is a fixed time limit (this also holds true for some achievement tests), which makes them at least in principle speeded. But knowledge about the time limit under which a test is administered does not say much about whether some persons would answer additional items correctly if given unlimited time; the omitted items may all have been so difficult as to make the probability of a correct answer very close to zero.

Thus whether a test with a time limit is speeded or not is an empirical question (in a sense this applies to tests given without time limits too since there may be self-imposed "time limits" which are consequences of boredom and tiredness) and

it does appear as if it is possible to investigate this question with the Rasch model. Not only is the analysis of Number Series presented above an example of this; Rasch (1960) was also able to identify lack of fit for a test as being a consequence of the test being speeded.

But it should also be pointed out that if we know that speed is the only violation of the assumptions, the Rasch model can be used to "partial out" the speed factor. This is effected through estimating the person parameter for the "power" from the scores obtained only on the attempted items. (A study using such procedures has been presented by Allerup, Mylov and Spelling, 1977).

Guessing can probably never be completely avoided but certain kinds of items, i.e. multiple-choice items, are of course especially likely to be affected by this extranous factor. Unless active attempts have been made to minimize guessing, the Rasch model (or the Birnbaum model) should be used only with great caution when the items are of multiple-choice type, keeping in mind that the item parameters cannot be expected to remain invariant over samples differing in levels of ability.

Varying item discrimination is a kind of threat to the validity of the Rasch model which is quite difficult to discuss since its implications are hard to identify at a more concrete level. In chapter 1.1 the Flogging Wall test was used as an example, and it was pointed out that the discrimination parameter of the canes corresponds to the amplitude of the flogging, i.e. to item reliability.

It certainly is possible to imagine that different kinds of item have different reliability; items requiring a constructed response are for example usually more reliable than multiple choice items. To take another example, T Lindblad (1977, personal communication) has pointed out that tests of listening comprehension for measuring foreign language achivement tend to be less reliable than tests of reading comprehension and

that the reason for this is probably that listening comprehension tests are more susceptible to chance influences than reading comprehension tests.

In one sense it could be argued that those items in the Number series test which were found to be influenced by arithmetical ability do have lower discrimination parameters, but since this can be explained with reference to a systematically working factor, it is better described as representing multidimensionality. Also in the analyses of Opposites we found that the items had different discriminative abilities which were among other things related to how much random guessing took place. But again, of course, it is basically a kind of multidimensionality which causes this to occur.

It may of course be possible to find items which measure the same unidimensional ability and which are, to different degrees, affected by chance factors (i.e. have different discrimination parameters). I do suspect, however, that in most cases when what appears to be varying item discrimination is found, a closer look will reveal that some kind of multidimensionality is involved.

Item bias is also a kind of multidimensionality since in this case variables associated with a particular group make the items systematically too easy or too difficult. A good knowledge of the items as well as of the sample is of course essential to produce a fruitful approach to this problem.

Constrained responses and learning effects from one item to another are threats to the model as well. If, for example, four responses are derived from a question requiring the pairing with respect to meaning of four given English words with four given Swedish words those of the examinees who know three of the answers will automatically get the fourth pair correct too, which obviously is a violation of the assumption of local statistical independence.

Learning effects from one item to another are violations of model assumptions of somewhat the same kind. Such effects may be very difficult to identify but it can be mentioned that it is also possible to generalize the Rasch model so that it can be used to study this problem specifically (see chapter 6.2).

Person fit may be a problem too: idiosyncratic working methods, cheating and carelessness (person reliability does appear to be at least as fruitful a concept as item reliability) are a few such threats to the model. Some of these factors can be controlled out in the administration of the items, others only through excluding persons.

Wright and Mead (1977) have presented a test of person fit, based on analysis of residuals. There is also the possibility of constructing a theoretically satisfying test of person fit under the conditional approach. Since the probability of each observed score vector can easily be computed (equation 2.1.8), all that is needed to obtain a p-value is to sum the probabilities of all more extreme score-vectors (i.e. those with lower probabilities) than the observed. The test does, however, appear to be computationally complex, so it has not been implemented in the present version of PML.

Strategies and problems in the development of Rasch scales

The usual procedure in attempting to find a set of items fitting the model is to select, on the basis of a tryout, out of a larger set of items those which appear to fit the model and then, at best, cross-validate the set on a new sample. Such a procedure is reasonable enough (at least when the reasons for misfit are not to be found in factors other then item heterogeneity) but there are some risks involved which need to be discussed. It has already been pointed out that when item heterogeneity is at issue the items do not fit the model but may fit each other. If a vast majority of the items measure the same ability, there being just a few deviating ones, the latter can easily be identified in the

graphic tests. But if there is more of item heterogeneity, the graphic tests are useless if used to select items which have a plot where the points fall close to the diagonal (and a statistical test is even worse); in this case it is necessary to keep an eye on the pattern of deviations common to several items and if such a group of items shows similarities also in other respects such as content, it is reasonable to select those and investigate if they form a Rasch scale.

But this can at time be a risky strategy. Sometimes several threats to the model are in operation and the problem is that these can combine in different ways for different items and even cancel out. It is quite easy to imagine, for example, what would happen if guessing is allowed in a set of heterogenous items; it would almost surely be impossible to get anything meaningful out of such an analysis.

This indicates that when attempts are made to maximize item homogeneity it is essential that all or most of the other sources of threat to the model assumptions are controlled for, which is reasonably easy with respect to factors such as guessing and speededness but which may be more difficult with respect to others.

Another conclusion which is inevitable in this light is that a very clear conception of the content of the items in the tryout is required if a meaningful selection and/or classification of items is to be made.

Degree of fit and inferential tests

There are problems involved in using statistical tests to decide whether the data show an acceptable fit or not. One problem is that even though one test may indicate a good fit, another may indicate a very poor fit. Another problem is related to sample size; when large samples are used very small deviations will result in significant values on the test statistic and when small samples are used even gross deviations may remain undetected. The first problem can to some degree

be solved by use of the measure of redundancy (chapter 3.3) but how large a sample is required to obtain a reasonable power in the statistical tests is as yet an unresolved problem.

These problems indicate that not too much weight should be placed on the inferential tests of goodness of fit and especially not when the sample sizes are extreme in either direction. Less formalized approaches thus appear to be necessary complements in evaluating fit. The graphic tests of items are here valuable and content-related considerations are indispensible.

But it must also be pointed out that the degree of fit which is necessary to some degree depends upon the applications intended. For some applications, such as the study of unidimensionality, we can accept only small deviations, but for others, such as perhaps more technologically oriented applications, we might expect to get useful results even when the fit is not the best possible. It does in fact appear to be a very important area of research to study how much the model assumptions can be violated without jeopardizing different kinds of applications of the model.

The concept of unidimensionality

The notion of unidimensionality is essential in all the IT models but particularly so in the Rasch model since there is no possibility of treating different kinds of multidimensionality as varying item discrimination in this model. There is thus reason to take up the notion of unidimensionality to special discussion.

In my opinion it is as yet an unanswered question what properties those scales fitting the Rasch model have from a psychological perspective. Are they for example so narrow and specific that they will be impractical to use? My own impression which is, however, only based on analyses of tests

originally constructed within the framework of classical test theory is that the Rasch model is extremely sensitive to any kind of multidimensionality and that the scales thus tend to be quite narrow. It does appear to be a research question of the highest priority to investigate the "psychological width" of item sets which do fit the model and to study how one should proceed if it is found that they in fact tend to be very narrow.

But be as it may with this question; the notion of unidimensionality is nevertheless of utmost importance in any attempt to make measurements. Some arguments in favour of this view have already been presented (page 9) but there is reason to emphasize once again the central importance and great use of the concept of unidimensionality.

As has been pointed out by Lumsden (1976) the requirement that tests should be unidimensional has been seriously neglected in classical test theory, which is probably partly due to the fact that there has existed no satisfactory method for studying unidimensionality but probably also to the fact that reasonable degrees of success in practical applications have been obtained without imposing this requirement.

But whenever anything more than some degree of correlation with an extranous measure is to be achieved, the assumption of unidimensionality is essential. Lumsden (1976) stressed that measurement is always measurement of an attribute or a property (a latent trait) so it may be asked: "How can we make any claims to measure if our measuring instrument has a number of different sets of items based presumably on different attribute conceptions?" (p. 266).

To construct a test intended to measure an attribute we of course need a conception of the attribute at once when the work is begun. But this conception is likely to be vague and there will be little basis for deciding whether an item or an item type does reflect the attribute. But through a continuing process of revision of the conception of the attribute and revision of the items used to measure the attribute we are likely

to obtain a better understanding both of the attribute and the measuring device. In such a process of revision the Rasch model can be supposed to contribute greatly, even though it is of course not the only method to be used in such work.

The notion of unidimensionality implies that only one attribute should be measured with the same test, but it does not imply that the latent trait in itself is unidimensional; it may well be functionally (and factorially) complex and we can certainly not claim that there is one unitary process underlying test performance. (But there are in fact developments of the Rasch model which are well suited to the study of what kind of processes contribute to the difficulty of items, see chapter 6.2).

Let me give one more example showing the importance of unidimensionality. In experimental educational research it is common practice to administer to groups given different treatments the same post-test, and then compare the outcomes in the treatments in terms of the means of raw scores obtained on the post-test. But if there are interactions between treatment and outcomes so that the difficulties of the items in the post-test vary as a function of treatment such a comparison can only produce more or less meaningless results. In such a case we would want to reorganise the items in the post-test into internally homogenous scales which measure the same thing in all treatments, the Rasch model can easily be applied to accomplish this.

Chapter 5

SOME AREAS OF APPLICATION

In the preceeding chapter it was pointed out that one very important area of application of the model is to study the internal workings of a test. But it is also true that once scales fitting the model have been developed it is possible to solve within the framework of the Rasch model a number of measurement problems. We will briefly indicate some of these possibilities.

5.1 Test optimation

The problem of how a test should be organized in terms of number of items, level of difficulty and spread of item difficulty in order to obtain a suitable precision of measurement can rather easily be solved using the information function with respect to the person parameters (chapter 2.3).

In the Rasch model the information with respect to a person parameter (and the item parameter) contained in the response to an item is a function only of the probability of a correct answer which is easily seen if we rewrite (2.3.2) slightly:

$$(5.1.1) \qquad I_i(\xi_v) = p_{vi}(1-p_{vi})$$

In Figure 5.1 $I_i(\xi_v)$ for any item is shown as a function of the probability of a correct answer. The maximum of the curve is where $p_{vi}=.5$ but we can also note that the information obtained is relatively constant within the range $.20 \leq p_{vi} \leq .80$.

From these properties of the model follows that the only factors affecting the precision of measurement at any given level of ability is the number of items in the test and the distribution of item parameters. But it also follows that the standard error of measurement varies as a function of ability, which can be illustrated with some examples.

For two tests, both with 40 items, the SEM(ξ) has been plotted against ξ in Figure 5.2. One of the tests (peaked) contains items which all have the same parameter (σ_i=0) and in the other test (spaced) the item parameters vary between -3 and 3 in equal steps. We see that the peaked test gives a higher SEM(ξ) for extreme person parameters while it gives a lower SEM(ξ) for the intermediate range of abilities.

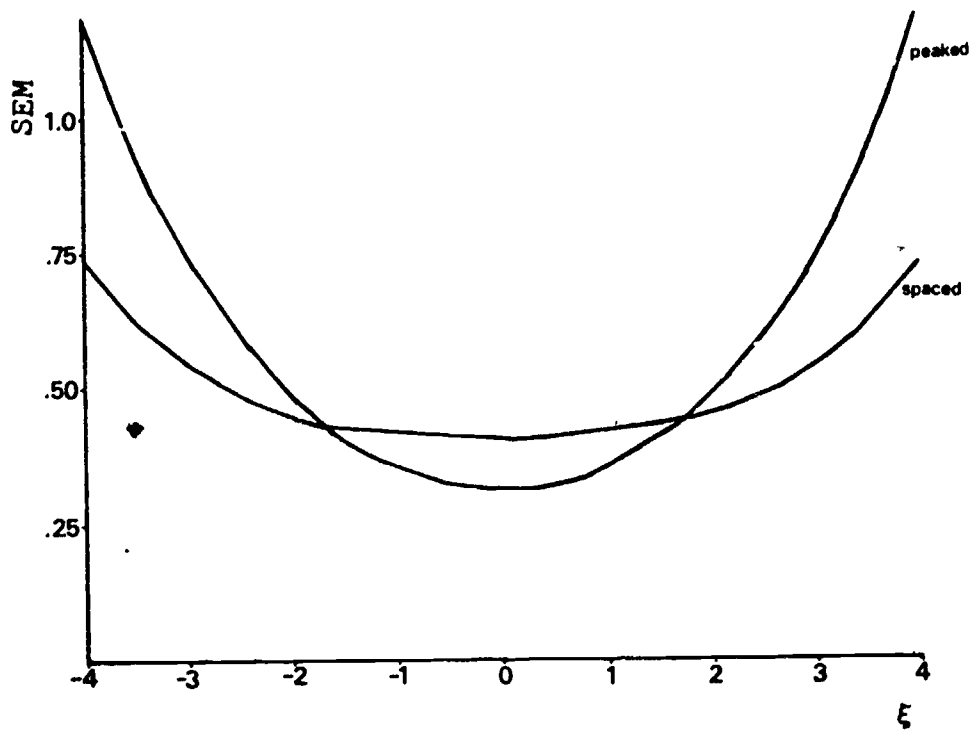


Figure 5.2. Standard errors of measurement of ability as a function of ability for two hypothetical tests.

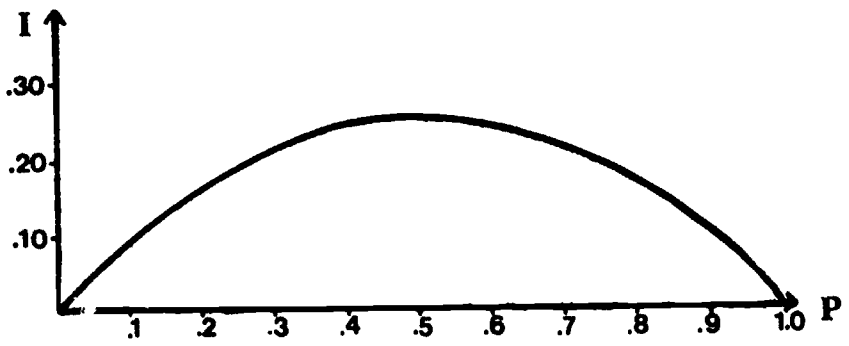


Figure 5.1 The information in an item as a function of probability of a correct answer.

From chapter 2.3 it is recalled that the information in a test with respect to a person parameter is the sum of the information contributed by each item:

$$(5.1.2) \qquad I_t(\xi_v) = \sum_{i=1}^{k} p_{vi}(1-p_{vi})$$

and it will also be recalled that the standard error of measurement SEM(ξ) is:

$$(5.1.3) \qquad SEM(\xi) = \frac{1}{\sqrt{I_t(\xi)}}$$

From these properties of the model follows that the only factors affecting the precision of measurement at any given level of ability is the number of items in the test and the distribution of item parameters. But it also follows that the standard error of measurement varies as a function of ability, which can be illustrated with some examples.

For two tests, both with 40 items, the SEM(ξ) has been plotted against ξ in Figure 5.2. One of the tests (peaked) contains items which all have the same parameter (σ_i=0) and in the other test (spaced) the item parameters vary between -3 and 3 in equal steps. We see that the peaked test gives a higher SEM(ξ) for extreme person parameters while it gives a lower SEM(ξ) for the intermediate range of abilities.

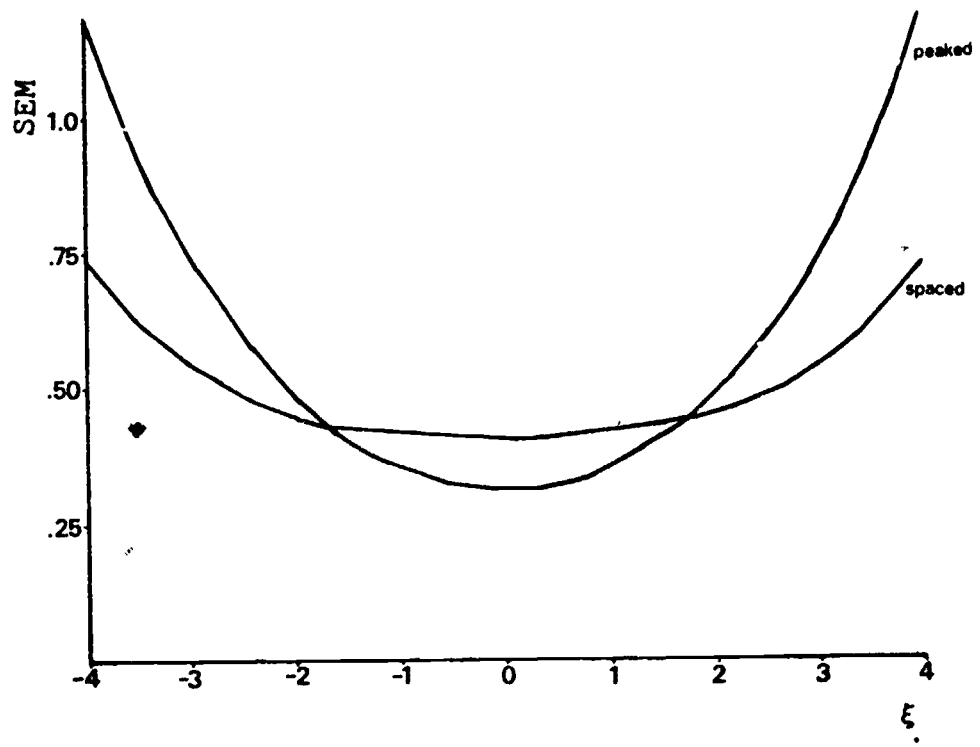


Figure 5.2. Standard errors of measurement of ability as a function of ability for two hypothetical tests.

The highest precision of measurement is of course always obtained for any given level of ability when at that level the probability of a correct answer is .50 for all the items. We can thus formulate the very simple rule that when the purpose is to measure just one level of ability, items should be selected which have the same parameter value as the ability to be measured. The number of items needed (k_w) to obtain any wanted precision (SEM_w) is of course easily determined:

$$k_w = \frac{1}{.25SEM_w^2} \qquad (5.1.4)$$

Mostly, however, a test is intended for use over a range of abilities and to reach any statement about how a test should be built up it is necessary to make assumptions about the distribution of person parameters. If we take a look at Figure 5.2 again in this light we find that any of the two tests can have the lowest mean of standard errors and thus have the best subject separation (or, equivalently, have the highest reliability). If, for example, the person parameters are distributed normally with zero mean and unit variance we would find that for more than 90% of the persons in the sample the peaked test has the lowest SEM(ξ) and would consequently yield the best subject separation. When the ISS's for the two tests were computed under these assumptions the values found were .88 and .85 for the peaked and spaced tests respectively (the corresponding values of KR_{20} were .89 and .85). If, however, we assume another distribution of person parameters such as a rectangular one or a normal distribution with a standard deviation which is considerably greater than unity it is easy to see that the peaked test will give an ISS lower than that for the spaced test.

The problem of how items with different parameters should be chosen so as to obtain maximum precision of measurement (in terms of the mean of the standard errors) has been studied in great detail by Douglas (1975) and Wright and Douglas (1975). They found that when the sample has a normal distribution of person parameters a peaked test centered on the

mean of the sample is optimal when the standard deviation (s) is not larger than 1.25-1.50 but that for samples with greater s uniformly spaced item difficulties should be used. For example, when s=1.75 an optimal difference of 6 between the highest and lowest item parameter was found (for this difference the term width, W, was used, so here W=6).

For rectangular distributions of person parameters lower values of s were found where a change from a peaked to a spaced test is motivated, the limit being around s=.75. Compared to the normal distribution a rectangular distribution of person parameters requires a greater spread of item parameters for the same s to obtain the best precision. For example when s=1.75 the optimum was found at W=10 for the rectangular distribution.

Wright and Douglas (1975) have summarized their studies in some simple rules for test construction: they do advise, for example that uniformly spaced item parameters with W=4s, where s is the best guess of the standard deviation in the sample, should be used. But it must of course be realized that use of such simple rules implies that some accuracy is sacrificed.

It is of some interest to compare the conclusions about optimal test design drawn here with those recommendations issued within the framework of the classical test theory. It has long been known that a test with uniform item difficulties (with a proportion of correct answers of .50, when no guessing is allowed) generally has a higher reliability than a test in which the item difficulties are spaced (e.g. Gulliksen, 1945; Lord, 1952). Another conclusion which has been drawn is that a better reliability is obtained if items with a high reliability, as measured for example with the biserial or point biserial correlation, are selected. But it has also been noted that for a peaked test there is an optimum item reliability beyond which the reliability of the test decreases; this is the so called attenuation paradox (e.g. Loevinger, 1954).

The explanation as to why the attenuation paradox occurs is quite simple if stated in general terms: If the reliability of all items is very high, the correlations between all items will approach unity (if the test is unidimensional) which means that a person who passes one item will pass all the others and that a person who fails one item will fail all the others. The distribution of scores will thus tend to be bimodal with a very good discrimination at one level of ability but with virtually no discrimination between examinees at other levels of ability. The attenuation paradox occurs only if the items all are of the same difficulty and the solution of the problem is, of course, to use items with spaced difficulties (e.g. Brogden, 1946; Cronbach & Warrington, 1952).

The conclusion was drawn above that when the variance of the person parameters in the sample is small a peaked test should be used, otherwise not. This conclusion is in fact identical to the solution of the problems caused by the attenuation paradox which follows from the fact that with a higher, for all items common, discrimination there is in the Rasch model a higher standard deviation of the person parameters.

In fact the real explanation of the attenuation paradox is of course that since the standard errors of measurement are larger for certain scores than for others, constructing the test so that for a sample it results in many scores which have a large standard error will have detrimental effects on the reliability. Thus what in classical theory is a paradox follows in the Rasch model (and all the other LT models) naturally from the fact that the standard errors of measurement vary as a function of ability.

5.2 Tailored testing

It is obvious that the strategy of giving the same set of test items to persons of all levels of ability will necessarily result in different precision of measurement at different levels of ability. The only possibility of obtaining standard

errors of measurement which are equal over a range of abilities is to give different items to different persons, i.e. tailored testing.

The LT models are of course extremely well-suited for tailored testing since it is possible to estimate on a common ability scale results obtained by different examinees on different items. The next section demonstrates how such a translation into a common metric can be effected with the Rasch model.

The basic principle is of course that all the persons should take items on which they have a probability of .50 of giving a correct answer. Ususally computer based administration of the items has to be used and there are different strategies by which items can be selected from a pool so as to keep as close to this requirement as possible (Lord, 1971, 1974b). There is of course additional use of the computer when, after the testing, the scores on the items are to be translated into the metric of the latent trait.

Wright and Douglas (1975) have, however, presented a system for self-tailored testing based on simple approximations in which a computer need not be involved either in selection of items or in computing person parameters:

> "The person to be measured can be handed a booklet of test items more or less equally spaced in increasing difficulty from easiest to hardest and invited to choose any starting place in the booklet with which he feels comfortable. From that self-chosen starting point the examinee can work at his own will and speed in either direction, forward into harder items or backward into easier ones, until he reaches his own performance limits or runs out of time. Whatever the level and length of the self-chosen segment, all that are needed to obtain an objective item-free person measure and its standard error are the serial numbers of the easiest and hardest items tried and the number of successes in between. These three observations are sufficient to look up in a

simple series of tables the person's estimated measure and the standard error of that estimate. (Wright & Douglas, 1975, p. 43-44).

As was mentioned above this system is based on certain approximations and it is easy to imagine practical problems in its application; it is, however, possible that it might work so well that it can be profitably exploited.

5.3 Test equating and linking

One area of great potential for applications of the Rasch model is equating and linking of tests, i.e. expressing on the same scale raw scores obtained on different tests (or sets of items).

Test equating

If it can be confidently assumed that two (or more) tests measure the same trait, equating of scores is a very simple task, which is illustrated below. However, since the assumption that the tests measure the same ability is critical we will first adress the problem of how to test this assumption.

Let us assume that the tests have been given to the same sample and that separate analyses of the tests have indicated a good fit to the model. If the tests measure the same ability then we must also find a good fit if all the items are analyzed together. This straightforward approach of testing the assumption is conceptually simple, but it may be impractical since when all the items are pooled, a very long test may be the result and it will be recalled that the overall numerical tests are cumbersome to compute when the number of items is large. Fortunately there exists a likelihood ratio test which directly tests the hypothesis that the two sets of items measure the same ability (Martin-Löf, 1973, p. 135-136). This test calls for some hand computations (or a short computer program) but requires otherwise only that the parameters are estimated for each test and for the pooled set of items.

Let us call the number of items in the two tests k_1 and k_2, with $k=k_1+k_2$. We define further $n_{r_1r_2}$ to be the number of persons with raw score r_1 on the first test and raw score r_2 on the second test. Let H_t be the maximum value of the logarithm of the likelihood function (3.2.1) and H_1 and H_2 the corresponding values for each test. Martin-Löf has then shown that the test statistic is:

(5.3.1)

$$\log\lambda = -\sum_{r_1=0}^{k_1}\sum_{r_2=0}^{k_2} n_{r_1r_2}\log\frac{n_{r_1r_2}}{n} + \sum_{r=0}^{k} n_r\log\frac{n_r}{n} + H_t - H_1 - H_2$$

and that $-2\log\lambda$ is approximately chi-square distributed with k_1k_2-1 degrees of freedom when $n\to\infty$.

The values of H_t, H_1 and H_2 are obtained on the computer printouts from the corresponding analyses. The values of the other terms appearing in (5.3.1) can be obtained either through hand calculations based on the bivariate and univariate distributions of test scores, or by writing a special program to perform these simple but sometimes tedious tasks.

The test presented above can be expected to be of use not only in testing the homogeneity of two distinct sets of items intended to be used as separate tests but also when very long tests are constructed. Since in such cases the overall numerical tests of goodness of fit are out of reach, at least if one is operating in an environment where computer time is of limited supply, a good strategy may be to develop out of the same pool of items two tests fitting the model and then investigate whether they can be put together into one long test.

Let's now turn to the problem of equating raw scores obtained on different tests. It will be recalled that in any estimation of the item parameters a constraint must be imposed, for

example that the sum of the item parameters expressed on the log scale is zero, which effects a fixation of the origin of the scale. The ability scales associated with two tests measuring the same ability are thus the same except for the arbitrary origin of the scales. But if the two tests are given to the same sample we are in the position to estimate the difference in origin of the scales since, of course, the same sample must have the same mean of ability whichever test is used.

There are two methods which can be employed to estimate the difference in origin of ability scales, both resulting in a simple additive constant to be used as a correction factor (see e.g. Kifer, 1976 ; Rentz & Bashaw, 1975). The first method, the so called "ability method" simply consists of calculating the difference in the means of ability estimated from the two tests and using the obtained difference as the correction factor. In the other method, the so called "difficulty method", the item parameters from both tests are estimated together and the difference between the means of the estimated item parameters is used as the correction factor.

These two methods give theoretically identical results but there is at least one thing that speaks in favour of the difficulty method: since the person parameters cannot be estimated for zero or perfect raw scores (such persons are excluded from the analysis) the ability method must not be used whenever different persons obtain such scores on the two tests.

The difficulty method will here be illustrated with some generated data. For a sample of 1 000 persons, distributed N(0,1), scores were generated for 40 items, of which 20 had the parameter -1 and 20 the parameter 1. It will be supposed that these two groups of items can be given as two forms, one simple and one difficult, and that we above all are interested in knowing which raw score on the simple form corresponds to which raw score on the difficult form.

The test of the homogeneity of the two forms gave $\chi^2=302.62$ with 399 degrees of freedom so it is obviously no problem to do the equating. Not surprisingly the difference between the means of the parameters of the simple and difficult items sets turned out to be -2 in the analysis. This value of -2 is the correction factor which of course means that we shall subtract 2 from (or rather add -2 to) the ability scale for the simple items to get the corresponding location of the ability scale for the difficult items. Table 5.1 presents the table of conversion from raw scores on the two forms into the ability scale of the difficult form.

Table 5.1 Person parameters expressed in the metric of the difficult test for raw scores obtained on the simple and difficult forms.

Raw score on the difficult form	Person parameter	Raw score on the simple form	Person parameter
1	-2.94	1	-4.94
2	-2.18	2	-4.18
3	-1.73	3	-3.73
4	-1.39	4	-3.39
5	-1.10	5	-3.10
6	-.85	6	-2.85
7	-.62	7	-2.62
8	-.41	8	-2.41
9	-.20	9	-2.20
10	.00	10	-2.00
11	.20	11	-1.80
12	.41	12	-1.59
13	.62	13	-1.38
14	.85	14	-1.15
15	1.10	15	-0.90
16	1.39	16	-.61
17	1.73	17	-.27
18	2.20	18	.20
19	2.94	19	.94

From the figures presented in the table it is easily understood that before the conversion was made the scale of ability for the simple form was numerically exactly the same as that for the difficult form. The reason for this is of course that separate analyses of the t' items sets produce exactly the same item parameters, (they are all equal to zero, more or less, as a consequence of the normation) so consequently there can be no difference between the numerical values of the person parameter corresponding to a certain raw score (but the distribution of person parameter is of course radically different).

Using linear interpolation methods a graph has been constructed (Figure 5.3) to show the relation between raw scores on the two forms, i.e. using the common scale of ability, raw scores on the simple form have been translated into raw scores on the difficult form. Obviously there is a curvilinear relationship between raw scores on the two forms.

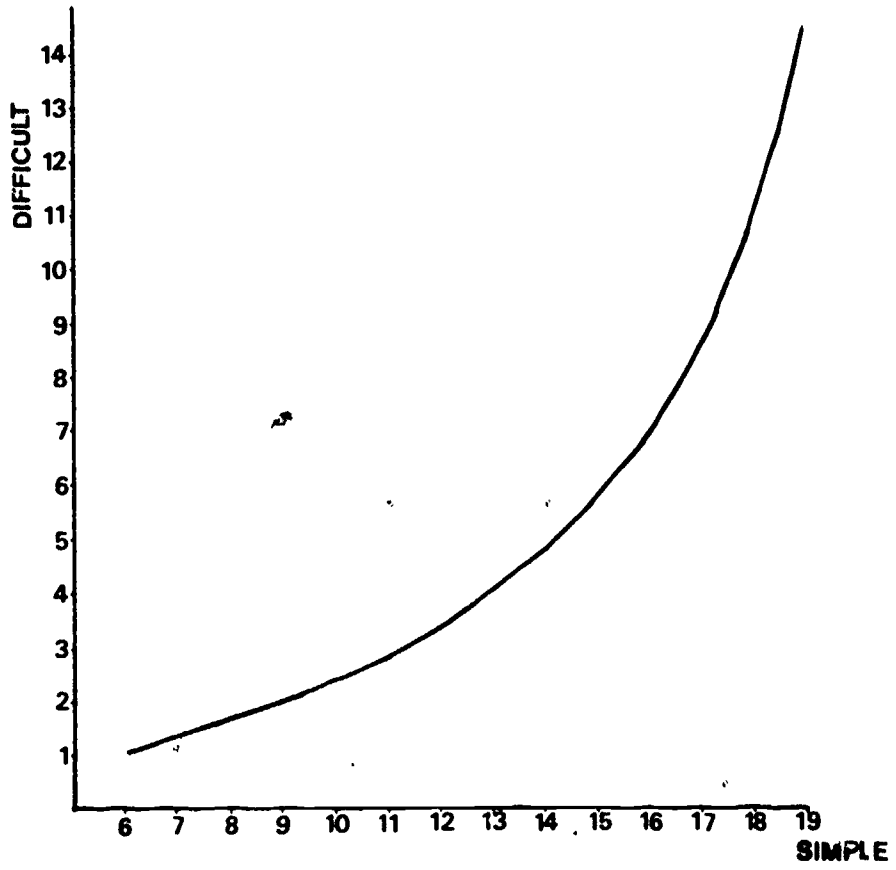


Figure 5.3. Raw scores on the simple and difficult forms corresponding to the same level of ability.

It should perhaps be pointed out that if we did test a sample with both forms we would not find this particular curvilinear relationship between raw scores on the two forms even for very large samples. The reason for this is that there is a regression towards the mean, i.e. those examinees with bad (good) luck on the simple form can on the whole not be expected to have an equally bad (good) luck on the difficult form and vice versa. The conversion should thus not be interpreted to mean that it gives the expected raw score on one form given the raw score on another form; rather it tells what raw score would have been found if the other form had been used instead, everything else being constant.

Test linking

Also in linking tests the purpose is to estimate on the same scale results obtained on different tests but in this case the tests are given to different samples; the linking is made possible through use of a subset of, say 10-20 items common to both (or all) tests.

A version of the difficulty method described above is used, in that the mean of the item parameters for the common subset is estimated in the context of each test. The difference between the means of the estimates of the parameters indicates of course the difference between the origins of the scales of the item parameters in the two tests and can be used as a correction or translation factor. Thereafter the ability scale associated with the "translated" item parameters must be computed (handy computer programs which perform this task can be found in Wright & Panchapakesan, 1969; Kifer, Mattsson & Carlid, 1975; Rentz & Bashaw, 1975) which makes it possible to translate into the ability scale of one of the tests, raw scores obtained on the other.

No example of how this can be done in practice is presented here; the reader is instead referred to Wright (1977) for further details and more elaborate linking designs and to Kifer, et al. (1975); see also Kifer, (1976), as they do present an easily followed example.

It need probably not be said that no linking should be attempted unless the tests measure the same ability. The fact that the tests do have a subset of items in common of course makes it possible to test this assumption: if all tests fit the model they must in fact measure the same ability.

5.4 Item banks

Virtually all the applications discussed in this chapter presuppose that there exists a pool of items measuring the same ability and for which items the difficulties have the same origin of scale. It is obvious than when such a pool of items is at hand a large range of measurement problems can be solved with great efficiency and simplicity; tests can be optimized for specific purposes and tailored testing becomes possible. Furthermore, all possible tests which can be constructed by selection of items from the pool are automatically equated (even though it is of course necessary to compute the associated ability scale for each selection of items so that the observed raw scores can be translated into the common metric).

The most effective way of developing item banks is of course to successively link new items into the bank, using the procedures of test linking described above. But it is important that an eye is kept on the fit of the items throughout: a bank consisting of heterogeneous items with a poor fit is probably worse than no bank at all; the strong claims which can be advanced in relation to the Rasch model are valid when the model holds true, otherwise not.

Chapter 6

GENERALIZATIONS OF THE RASCH MODEL

This report treats in detail only the simplest case, i.e. in which the model specifies only two parameters and there are only two categories of answer. (Even though the wording has been phrased in terms of measurement of ability there is of course nothing that says that the model cannot be used to measure personality, attitudes and so on). There are, however, developments of the basic model which can deal with more complex situations and the parameter structure of the model can be transformed in different ways. Some of these generalizations of the model will be briefly mentioned below.

6.1 The polychotomous case

It is possible to generalize the model to treat the case where there are more than two categories of answer, as is for example often the case in attitude questionnaires (Andersen, 1973; Fischer, 1974, p. 424 ff.; Allerup & Sorber, 1977).

Instead of observing whether a particular answer is correct or incorrect we observe which particular answer category (h, $h=1,\ldots,m$) a person v endorses on item i. We can represent the answer by using a selection vector $(A_{vi})=(A_{vi}^{(1)},\ldots,A_{vi}^{(h)},\ldots,A_{vi}^{(m)})$ which contains zeroes for all the alternatives not chosen and a one for the category endorsed. If there, for example, are three categories of answer and a person choses the last for a particular item this is represented with the selection vector (0,0,1).

Instead of one person parameter there is in the polychotomous case a vector of person parameters, the elements of which each indicate the tendency for each persons to chose each alternative: $(\theta_v)=(\theta_v^{(1)},\ldots,\theta_v^{(h)},\ldots,\theta_v^{(m)})$. In the same way there is for

each item a vector of parameters representing the tendency for each alternative to be chosen: $(\varepsilon_i)= (\varepsilon_i^{(1)},\ldots,\varepsilon_i^{(h)},\ldots,\varepsilon_i^{(m)})$.

We need, however, to impose a constraint on these vectors of parameters and we can use $\theta_v^{(m)}$ and $\varepsilon_i^{(m)}$ for unity normation, i.e. they are put equal to unity. We can then write the basic model in the following way:

$$(6.1.1)\qquad \begin{cases} P(A_{vi}^{(1)}=1|v,i)= \dfrac{\theta_v^{(1)}\varepsilon_i^{(1)}}{1+\sum_{h=1}^{m}\theta_v^{(h)}\varepsilon_i^{(h)}} \\ P(A_{vi}^{(h)}=1|v,i)= \dfrac{\theta_v^{(h)}\varepsilon_i^{(h)}}{1+\sum_{h=1}^{m}\theta_v^{(h)}\varepsilon_i^{(h)}} \\ P(A_{vi}^{(m)}=1|v,i)= \dfrac{1}{1+\sum_{h=1}^{m}\theta_v^{(h)}\varepsilon_i^{(h)}} \end{cases}$$

Thus, the ICC is for each answer category here multidimensional and there are m-1 dimensions. But of course the notion of unidimensionality is as important here as everywhere else so it may be asked whether the multidimensional model may, in fact, be reduced into a unidimensional one. This can be done if it is possible to find a unidimensional vector of item parameters (ε_i),i=1,...,k and a "scoring-vector" $(\phi^{(h)},h=1,\ldots,m)$ so that for all items $\log\varepsilon_i^{(h)}=\phi^{(h)}\varepsilon_i$.

There are great technical complexities in obtaining CML estimates of the parameters. Allerup and Sorber (1977) have, however, presented such a computer program, based on methods for computing the symmetric functions and solving the equations suggested by Andersen (1972). This program also tests the hypothesis that the multidimensional model can be reduced into a

unidimensional one and provides also the necessary information for performing goodness of fit tests. There do also exist approximations to the strictly conditional approach: Fischer (1974, p. 571) has presented such a program for the case where there are three categories of answer, and methods for obtaining unconditional estimates have also been developed (Andrich, 1977). Examples of applications of the polychotomous Rasch model have been presented by Fischer (1974, p. 478 ff.).

6.2 The linear logistic model

In the basic Rasch model there is one difficulty parameter for each item; it is, however, possible to construct models with another parameter structure. A very interesting model results when the item parameters are replaced with a smaller number of "basic parameters" ($\eta_j, j=1,\ldots,m$) representing, for example, hypothesized processes which appear with different frequency in different items. By specifying one parameter for each process and the frequency with which it has to be carried out, the difficulty parameters can be "explained". We thus want to investigate the hypothesis that $\log\varepsilon_i = \sum_{j=1}^{m} f_{ij}\eta_j$, which can be be made empirically when the matrix of frequencies $((f_{ij}))$ has the rank m, and when $m<k$.

The model has been presented in detail by Fischer (1974, p. 340 ff.; a computer program is also presented); Fischer (1974) and Lybeck (1974) discuss some very interesting possible applications in an educational context. It should be pointed out, however, that Kempf and Niehausen (1976) have criticized this approach on the basis of lack of interpretability of the "basic parameters". They suggest instead that error types should be analyzed with a polychotomous model.

Dynamic models in which "transfer effects" are represented with special parameters have also been proposed and used (Spada, 1976; Kempf, 1976; Kempf, Niehausen & Mach, 1976). Such models can be used to investigate learning effects from one item to another as a threat to the validity of the basic

model, but can of course also be used to investigate substantive problems of great interest.

6.3 Analyses of experimental data

The linear logistic models mentioned above can be used to analyze data from experimental studies (see e.g. Kempf et al., 1976). But as has been pointed out by Fischer (1974, p. 506) it is also possible to formulate linear logistic models resembling the analysis of variance model, i.e. with parameters representing treatment and interaction effects of different kinds. Such models would entail one single assumption (which is also empirically testable), namely that there is an additive or, equivalently, a multiplicative relationship between the parameters, and they would fill a deeply felt need for sound statistical models for the analysis of qualitative data.

Chapter 7

THE PML PROGRAM

The computer program is written in FORTRAN IV and was developed on the IBM machines (360/65 and 370/148) at GUC (Gothenburg Universities Computing Center). The program should, however, only to a small degree be machine dependent (one version of the program at least) so it can probably relatively easy be implemented on other machines.

7.1 The two versions of PML

There are two versions of the program: one OSIRIS version calling routines in the OSIRIS III (1973) subroutine library and a non-OSIRIS version (or rather a simplified OSIRIS version) in which all routines called are included in the source code. The OSIRIS version can of course only be used at computer installations where the OSIRIS system is implemented.

The OSIRIS system has four important advantages:

- A self descriptive data structure is used, i.e. for each data file there is an associated dictionary file containing descriptions of the data file such as variable numbers, variable locations and names of the variables. This implies that the variables (items) can be referred to with a variable number which remains constant from analysis to analysis and that the variables are easily identified on the printout since they have a unique name.
- Specification of the control parameters for each run is easy since keywords are specified in a completely free format.
- Selection of any subset of cases is easily effected through the special filtering feature.
- Since the input routines are coded in Assembler they are very fast.

In the non-OSIRIS version of the program some of these advantages are lacking: no filtering is possible and fixed format specification of a few of the control parameters is necessary. However, to maximize the similiarity between the versions, and to gain some of the advantages of OSIRIS, a simplified OSIRIS structure has been created (this work has been done by Jan-Gunnar Tingsell at the Department of educational research, University of Göteborg) in which a simplified dictionary file is used along with the data file (see below).

Thw two versions of the program thus differ with respect to the input routines used; in the analysis parts of the programs there are no differences.

7.2 Obtaining a copy of the program

The scource code punched on cards (or written on a tape sent to me) may be acquired from the Institute of Education, University of Göteborg by writing to the present author. A fee is charged corresponding to the price of the cards and the costs involved in handling and shipping. Please indicate whether the OSIRIS or the non-OSIRIS version of PML is desired.

7.3 Using PML

Since the control information needed for the two versions of the program is somewhat different and is specified in different ways, the instructions for use will be specified separately. Some advice about choice of options is also given below, but only in connection with the OSIRIS version.

How to use the OSIRIS version

The control cards for the OSIRIS version are specified according to the standard OSIRIS III (1973) syntax and there is no need to describe the details here. Three or four statements (mostly corresponding to the same number of cards) are

necessary as input:

1. Filter statement (optional)
2. Title card (80 characters of information to label the output)
3. Global parameters (selected from the 15 parameters described below)
4. Variable list.

The global parameters are selected from those described below (defaults are underlined)

PRINT=DICT/<u>NODI</u>

DICT: Print the dictionary
NODI: Do not print the dictionary

DESC/<u>NODE</u>

DESC: Only descriptive information (e.g. proportion of correct answers, point-biserial correlations, and the item by score group frequency matrix of correct answers) is supplied without any estimation of item and person parameters. This keyword can be specified to make sure in an economical way that there are no items with a very high proportion of correct answers (which causes a slow convergence). Another usage is to have a look at the $((n_{ir}))$ matrix in order to specify a suitable minimum group size for the Andersen test (see page 49 above).
NODE: A full analysis, according to the other options chosen is performed.

MAXI=N

The maximum number of iterations in the estimation of the item parameters. The default is N=250. If convergence has not been obtained within the specified number of iterations PML will assume that this has occured when MAXI is reached and will continue with the other tasks set up. The maximum number of iterations in estimating the person parameters is taken as 4N.

ERROR=N

The accuracy required in the estimation of item and person parameters in terms of number of decimal places. The default is N=3. For some purposes a lower accuracy can be demanded but certainly not when the overall numerical tests are to be computed. The variance- covariance matrices which are inverted in the computation of the Martin-Löf test (see page 51 ff.) may for example not be positive definite when accuracy is too low.

ALGO=DIFF/SUM

DIFF: The symmetric functions are computed with the Difference algorithm (see page 31). Since this algorithm is sensitive to roundoff errors it should not be used when the number of items is large and/or there is a great range of item parameters. This algorithm seldom works when k>40 and it seldom fails when k<20. It should be pointed out, however, that even though this algorithm may work well in estimating the item parameters for the whole sample it may break down when the Andersen test is computed. When this test is requested this algorithm should thus be avoided unless k<20. There is no risk, however, of getting wrong results as a consequence of roundoff errors since the program is stopped when computational accuracy gets too low.

SUM: The symmetric functions are computed with the Summation algorithm (see page 32). This algorithm works in those cases in which the Difference algorithm fails but it is somewhat slower.

PREC=SING/DOUB

This keyword is effective only when ALGO=SUM is chosen.

SING: The symmetric functions are computed with single precision arithmetic.

This keyword should be chosen only when it is essential to keep the amount of computer time to a minimum. Observe that there is no test of computational accuracy in the SUM algorithm.
DOUB: The symmetric functions are computed with double precision arithmetic.

START=APPR/UNIT

APPR: The approximation suggested by Martin-Löf (1973, see page 34 above) is used to compute start values for the iterations. This keyword can be chosen regularly.
UNIT: Unities are used as start values for the iterations.

NORM=N

N= the variable number of the item chosen for unity normation. The default is the item of medium difficulty.

EXTR/NOEX

EXTR: The Aitken extrapolation (see page 35) is used to speed up convergence of the iterations. This default value can be used regularly but if the iterations should diverge the extrapolation may be the explanation.
NOEX: No extrapolation is done.

PERS/NOPE

PERS: The person parameters are estimated.
NOPE: The person parameters are not estimated. In a process of item selection and goodness of fit testing it may be a waste to estimate the person parameters in each analysis. But it is of course not possible to obtain estimates of the standard errors of the estimated item parameters if the person parameters are not computed.

PLOT/<u>NOPL</u>	PLOT: For each item a printerplot is made of the observed proportion of correct answers against the proportion predicted for each score group (see chapter 3.1). Observe that these plots produce a large amount of lines as output. NOPL: No plots are made.
BINO/<u>NOBI</u>	BINO: For each item and for each score group a binomial test is carried out to test the difference between observed and predicted frequencies of correct answer (see pages 46-47). The power of these tests is lower than the "power" of the printerplots but may at times be useful. They also present the numerical information on which the printerplots are based. NOBI: No binomial tests are carried out.
NOBS=N	N+1 is the smallest size allowed for a score group if it is to be considered in the printerplots or in the binomial tests. The default is N=5.
TEST=CHIS/LIKE/BOTH/ <u>NONE</u>	CHIS: The Martin-Löf chi-square goodness of fit test is computed (see chapter 3.2). LIKE: The Andersen conditional likelihood ratio test is computed (see chapter 3.2). BOTH: Both the overall numerical tests are computed. This keyword should be chosen only rarely, especially if k is large, for economical reasons. NONE: No overall test is computed.
NIND=N	N is the minimum number of persons allowed within each range of scores when the Andersen test is computed. The default is N=100.

A fairly typical example of the setup, including the JCL, required for executing the OSIRIS version of PML on an IBM machine under OS is shown below:

```
//UPEJEG JOB ...
/*JOBPARM RTIME=3,LINES=6K
// EXEC ... (referring to the library where PML is to be found)
//DICTIN DD ... (description of the dictionary file)
//DATAIN DD ... (description of the data file)
//FT12F001 DD UNIT=SYSSQ,DISP=(,PASS),SPACE=(TRK,(50,20)),
//       DCB=(RECFM=VBS,BLKSIZE=6000)      (description of the
         scratch file used in the computation of the Martin-
         Löf test)
//FT01F001 DD *      (observe that the instream is defined
                     as unit 1)
INCLUDE V3=1*       (Filter card)
BOYS IN GRADE 6     (Title card)
ALGO=SUM PLOT*      (Parameter card)
V121-V140,V145*     (Variable list)
/*
```

How to use the non-OSIRIS version

In OSIRIS the dictionary file is created with a special program. Also in the non-OSIRIS version of PML a dictionary file is used; here, however, the dictionary is simply punched on cards (but of course the card images can be stored on a disc or a tape). The non-OSIRIS dictionary must be prepared in the following way:

1st card

pos 1-3 Logical record length (LRECL) for each record in the data file. (If the data are on cards LRECL is of course 80; if there are more items than can be contained on one card it is necessary first to create a file with a greater LRECL).

pos 4-6 The variable number of the first item described in the dictionary (need not be 1).

pos 7-9 The variable number of the last item described in the dictionary.

2nd and following cards:

pos 1-3 Variable number
pos 4-27 Variable name
pos 28-30 Column location in the data file

The variables must be continously numbered between the first and the last variable number, but there is no restriction as to where in the record the different variables are located. It must be observed, however, that the information for each item must be punched in only one column (i.e. using I1 format), and that the responses of course must be coded 0 and 1. At most 200 items can be described in the dictionary.

An example is given below:

```
181 28 43
028 VOK A 1                    062
029 VOK A 2                    063
030 VOK A 3                    064
031 VOK A 4                    065
032 VOK A 5                    066
033 VOK A 6                    067
034 VOK A 7                    068
035 VOK A 8                    069
036 VOK A 9                    070
037 VOK A 10                   071
038 VOK A 11                   072
039 VOK A 12                   073
040 VOK A 13                   074
041 VOK A 14                   075
042 VOK A 15                   076
043 VOK A 16                   077
```

This dictionary describes 16 items in a data file with LRECL=181. The variable number for the first item has been taken to be 28; if, for example there is another subtest preceeding this one which in a later step is to be analyzed together with these items, the same variable numbers can be used.

In executing the non-OSIRIS version of PML there are 4 control statements (usually the same number of cards) which must be supplied:

1. Title card (80 characters of information to label the output)
2. Keyword parameter card (keywords are selected from those described below)
3. Fixed format parameter card (is prepared according to the instructions given below)
4. Variable list (see below)

The keyword parameter card should contain a selection from the keywords described below:

NODI/DICT	NODI: The dictionary is not printed. DICT: The descriptions in the dictionary for the variables selected in the variable list are printed.
DIFF/SUMM	DIFF: The symmetric functions are computed with the Difference algorithm. SUMM: The symmetric functions are computed with the Summation algorithm.
DOUB/SING	This keyword is effective only when SUMM is chosen. DOUB: Double precision arithmetic is used. SING: Single precision arithmetic is used for computing the symmetric functions.
APPR/UNIT	APPR: Start values for the iterations are computed according to an approximation (see page 34). UNIT: Unities are used as start values in solving the equations for the item parameters.
EXTR/NOEX	EXTR: The Aitken extrapolation (see page 35) is used to speed up convergence of the iterations. NOEX: No extrapolation is used.

PERS/NOPE	PERS: The person parameters are estimated. NOPE: The person parameters are not estimated.
NODE/DESC	NODE: A full analysis is performed. DESC: Only descriptive information is presented, without any estimation of item and person parameters.
PLOT/**NOPL**	PLOT: For each item a printerplot is made as a graphic test. NOPL: No printerplot is made.
BINO/**NOBI**	BINO: For each item and for each score group a binomial test is carried out to test the difference between observed and predicted frequencies of correct answers. NOBI: No binomial test is carried out.
NONE/CHIS/LIKE/BOTH	NONE: No overall numerical test of goodness of fit is computed. CHIS: The Martin-Löf chi-square goodness of fit test is computed. LIKE: The Andersen conditional likelihood ratio test is computed. BOTH: Both the overall numerical tests are computed.

The keywords selected to override the defaults are written on the keyword parameter card, beginning in the first position. The keywords are specified in any order and are separated with comma or blank. The list of keywords must be ended with an asterisk.

An example is given below:
DICT SUMM PLOT LIKE*

The fixed format parameter card is prepared in the following way:

Pos	
1-4	Maximum number of iterations in estimating the item parameters (MAXI). If left blank MAXI is assumed to be 250. The maximum number of iterations in estimating the person parameters is taken to be 4 times MAXI.
5-8	The accuracy required in the estimation of the item and person parameters in terms of number of decimal places (ERRO). If left blank ERRO is assumed to be 3.
9-12	The minimum number of persons allowed within each range of scores when the Andersen test is computed (NIND). If left blank NIND is assumed to be 100.
13-16	The smallest size allowed for a score group if it is to be considered in the printerplots or the binomial tests (NOBS). If left blank NOBS is assumed to be 5.
17-20	The variable number, according to the dictionary, of the item chosen for unity normation(NORM). If left blank the item of medium difficulty is used for unity normation.

Even if there is nothing punched on the fixed format parameter card it must be physically in place, after the keyword parameter card. An example is given below:

```
100   4 150  10
```

The variable list must contain a list of the variable numbers for those items to be included in the analysis. Each variable number must be specified with three digits (e.g. 006) and the numbers should be separated with comma or hyphen, where the hyphen indicates that a range of items are selected. The variable list must be started in position 1 and as many cards as are necessary may be used. Each card must be filled, however, and the comma is the only sign which is allowed in column 80, if continuation to a new card is to be made. The variable list must be ended with an asterisk. An example of a variable list could be:

005,008-017,034*

In executing the non-OSIRIS version of PML the control cards are read from unit 1, the dictionary from unit 13 and the data from unit 14. A fairly typical example of the setup, including the JCL, for executing this version of PML on an IBM machine is shown below:

```
//UPEJEG  JOB
//  EXEC ...  (referring to the library where PML is to be found)
//FT12F001  DD  UNIT=SYSSQ,DISP=(,PASS),SPACE=(TRK,(50,20)),
//      DCB=(RECFM=VBS,BLKSIZE=6000) (description of the scratch
                                      file used in the computa-
                                      tion of the Martin-Löf test)
//FT14F001  DD ...  (description of the data file)
//FT01F001  DD *
GRADE 6              (Title card)
SUMM PLOT*           (Keyword parameter card
 150                 (Fixed format parameter card)
121-140,145*         (Variable list)
//FT13F001  DD *
256 78192            (The dictionary)
078GRAMMARTEST 1, ITEM 1   112
.
.
.
192GRAMMARTEST 1, ITEM 115 226
/
```

7.4 The most important subroutines

READ — reads the data, forms the $((n_{ir}))$ matrix and computes the proportions of corrects answers, the point-biserial correlations (with the item included in the test) and the KR_{20}.

PAREST — administeres the iterative solution of the equations for the item parameters.

GAMMA — is used to compute the symmetric functions with the Difference algorithm. This routine has been written by Fischer (1974).

GAMMA2 supervises the computation of the symmetric functions with the Summation algorithm and calls repearedly the

GAM routine,which is a slightly changed version of a routine presented by Fischer (1974), or the

GAME routine, which is a single precisicn version of GAM.

AITKEN computes the Aitken extrapolation, if requested. It is called by PAREST.

PERS estimates the person parameters iteratively using the Newton-Raphson method. This subroutine has been taken from Fischer (1974) but code for computing start values has been added. The present version also computes the standard errors of the person parameters and the routine calls

ITINFO which computes the standard errors of the item parameters.

ITTEST administers the analysis of the items and calls

PLOTT which produces the printerplots and

DPIBIN which computes the cumulative binomial distribution. The latter routine has been taken from Allerup and Sorber (1977).

PMLCHI administers the computation of the Martin-Löf chi-square test but most of the computational work is carried out in

STORVA and in the two SSP routines

DMFSD and

DSINV which invert the variance- covariance matrices.

EBACHI groups the score groups and computes the Andersen likelihood ratio test by calling PAREST as many times as groups found.

7.5 Dimensioning of the program

The version which is delivered is dimensioned for $k_{max}=60$. Dummy dimensions are, however, used almost throughout so it is easy to dimension the program for both smaller and larger

problems. The following arrays must be changed in MAIN with K as the maximum number of items:

```
INTEGER  V(K),VMD1(K),VMD2(K),NIS(K,K),NR(K),AOI(K)
INTEGER*2 LIST(K),KDIFF(K)
REAL*4 WK2(2,K),W(K)
REAL*8 EPS(K),EPSI(K),G(K),GI(K,K),WK3(3,K),THETA(K),SAVE(K),
       VARKOV(K*(K+1)/2)
```

In GAM there are two arrays the dimensions of which must be changed:

```
REAL*8 X(K),Y(K)
```

and in GAME there are three:

```
REAL*4 E(K),X(K),Y(K)
```

Since the program tests that no attempts are made to analyze greater sets of items than it is dimensioned for an IF statement must be changed too. This test is made in MAIN immediately after the variable list has been read.

Furthermore, in any implementation of PML there is one more array the size of which must be considered. As was mentioned above on page 53 an array is used to store as many matrices of second derivatives of the symmetric functions as possible. This array (STOR) should be dimensioned to be as large as the available core allows. It is also necessary that the size of STOR is represented as the integer constant in the statement immediately preceeding the call to PMLCHI in MAIN.

7.6 A sample printout

On the following pages a sample printout from a run with the non-OSIRIS version of PML is shown.

Page 1

NUMBER SERIES

FOLLOWING PARAMETERS OVERRIDES THE DEFAULTS:

DICT
PLOT
BINO
ROTH

012-020*

NUMBER OF ITEMS..... 9
MAXIMUM NUMBER OF ITERATIONS.. 250
CRITERION FOR CONVERGENCE..... 0.0010
THE SYMMETRIC FUNCTIONS WILL BE COMPUTED WITH THE DIFFERENCE METHOD.
THE AITKEN EXTRAPOLATION WILL BE USED TO SPEED UP CONVERGENCE.
SCORE GROUPS WITH 5 OR FEWER PERSONS ARE NOT CONSIDERED IN THE PLOTS OR THE BINOMIAL TESTS.

Page 2

THE FOLLOWING VARIABLES ARE INCLUDED:

VAR	NAME	TLOC
12	NUMBER SERIES ITEM 12	106
13	NUMBER SERIES ITEM 13	107
14	NUMBER SERIES ITEM 14	108
15	NUMBER SERIES ITEM 15	109
16	NUMBER SERIES ITEM 16	110
17	NUMBER SERIES ITEM 17	111
18	NUMBER SERIES ITEM 18	112
19	NUMBER SERIES ITEM 19	113
20	NUMBER SERIES ITEM 20	114

NUMBER OF CASES READ..... 566
NUMBER OF CASES WITH A ZERO SCORE..... 53
NUMBER OF CASES WITH A FULL SCORE..... 44
NUMBER OF CASES REMAINING FOR ANALYSIS. 469

VAR.NO	VARIABLE NAME	PROPORTION CORRECT	POINT BISERIAL CORRELATION
12	NUMBER SERIES ITEM 12	0.578	0.530
13	NUMBER SERIES ITEM 13	0.716	0.440
14	NUMBER SERIES ITEM 14	0.597	0.494
15	NUMBER SERIES ITEM 15	0.678	0.457
16	NUMBER SERIES ITEM 16	0.552	0.503
17	NUMBER SERIES ITEM 17	0.512	0.468
18	NUMBER SERIES ITEM 18	0.516	0.556
19	NUMBER SERIES ITEM 19	0.456	0.518
20	NUMBER SERIES ITEM 20	0.501	0.510

NORMATION ON VARIABLE 16

THE RELIABILITY (KR-20) IS 0.64

THE ITEM BY SCOREGROUP FREQUENCY MATRIX OF CORRECT ANSWERS

P.NO	1	2	3	4	5	6	7	8	
12	0	7	18	35	34	37	63	77	271
13	8	16	30	35	52	51	65	79	336
14	4	8	22	29	37	39	66	75	280
15	9	13	30	28	46	44	68	80	318
16	0	6	11	31	35	41	55	80	259
17	5	9	12	20	30	38	55	71	240
18	5	2	7	19	37	40	61	71	242
19	2	3	15	13	26	31	51	73	214
20	5	6	14	14	28	34	55	74	235
	38	35	53	56	65	60	77	85	

NUMBER OF ITERATIONS FOR CONVERGENCE: 7

LOGLIKE = -0.1690032H91D+04

This is the maximum of the logarithm of the likelihood function.

Page 5

ITEM PARAMETERS

ITEM.NO	UNITY NORMATION	PRODUCT NORMATION	PRODUCT NORMATION(LOG)	STANDARD ERROR	CONFIDENCE INTERVAL(95 %)	
12	1.13370	1.04068	-0.03987	0.10858	-0.25268	0.17294
13	2.35754	2.16410	-0.77200	0.11618	-0.99971	-0.54430
14	1.24718	1.14484	-0.13527	0.10919	-0.34928	0.07874
15	1.90206	1.74599	-0.55732	0.11323	-0.77926	-0.33538
16	1.00000	0.91795	0.08562	0.10794	-0.12595	0.29718
17	0.82239	0.75491	0.28116	0.10732	0.07081	0.49151
18	0.83937	0.77049	0.26072	0.10736	0.05029	0.47116
19	0.63178	0.57994	0.54483	0.10719	0.33474	0.75492
20	0.78152	0.71739	0.33213	0.10723	0.12196	0.54230

Page 6

ABILITY PARAMETERS

SCORE	PRODUCT NORMATION	PRODUCT NORMATION(LOG)	STANDARD ERROR	CONFIDENCE INTERVAL(95 %)	
1	0.11698	-2.14575	1.07098	-4.24487	-0.04663
2	0.27317	-1.29765	0.81468	-2.89442	0.29911
3	0.48783	-0.71779	0.72088	-2.13071	0.69513
4	0.79557	-0.22869	0.68449	-1.57029	1.11290
5	1.26563	0.23557	0.68370	-1.10448	1.57562
6	2.05961	0.72252	0.71864	-0.68603	2.13106
7	3.66238	1.29811	0.81142	-0.29226	2.88849
8	8.49118	2.13903	1.06722	0.04728	4.23077

NUMBER OF ITERATIONS FOR CONVERGENCE OF THE ABILITIES: 36.

ON THE LOG SCALE THE MEAN OF THE SAMPLE IS 0.35 WITH THE VARIANCE 1.63

THE INDEX OF SUBJECT SEPARATION IS 0.57

Page 7

PROBABILITY OF OBTAINING A CERTAIN RAW SCORE GIVEN THE PERSON PARAMETER

SCORE	PARAMETER	RAW SCORE 0	1	2	3	4	5	6	7	8	9
1	0.11694	0.34	0.39	0.20	0.06	0.01	0.00	0.00	0.00	0.00	0.00
2	0.27317	0.10	0.27	0.31	0.21	0.09	0.02	0.00	0.00	0.00	0.00
3	0.48783	0.02	0.11	0.24	0.28	0.21	0.10	0.03	0.01	0.00	0.00
4	0.79557	0.00	0.03	0.11	0.22	0.27	0.21	0.11	0.04	0.01	0.00
5	1.26563	0.00	0.01	0.04	0.11	0.21	0.27	0.22	0.11	0.03	0.00
6	2.05961	0.00	0.00	0.01	0.03	0.10	0.21	0.28	0.24	0.11	0.02
7	3.26238	0.00	0.00	0.00	0.00	0.02	0.09	0.21	0.31	0.27	0.10
8	4.49119	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.20	0.39	0.34

Page 8

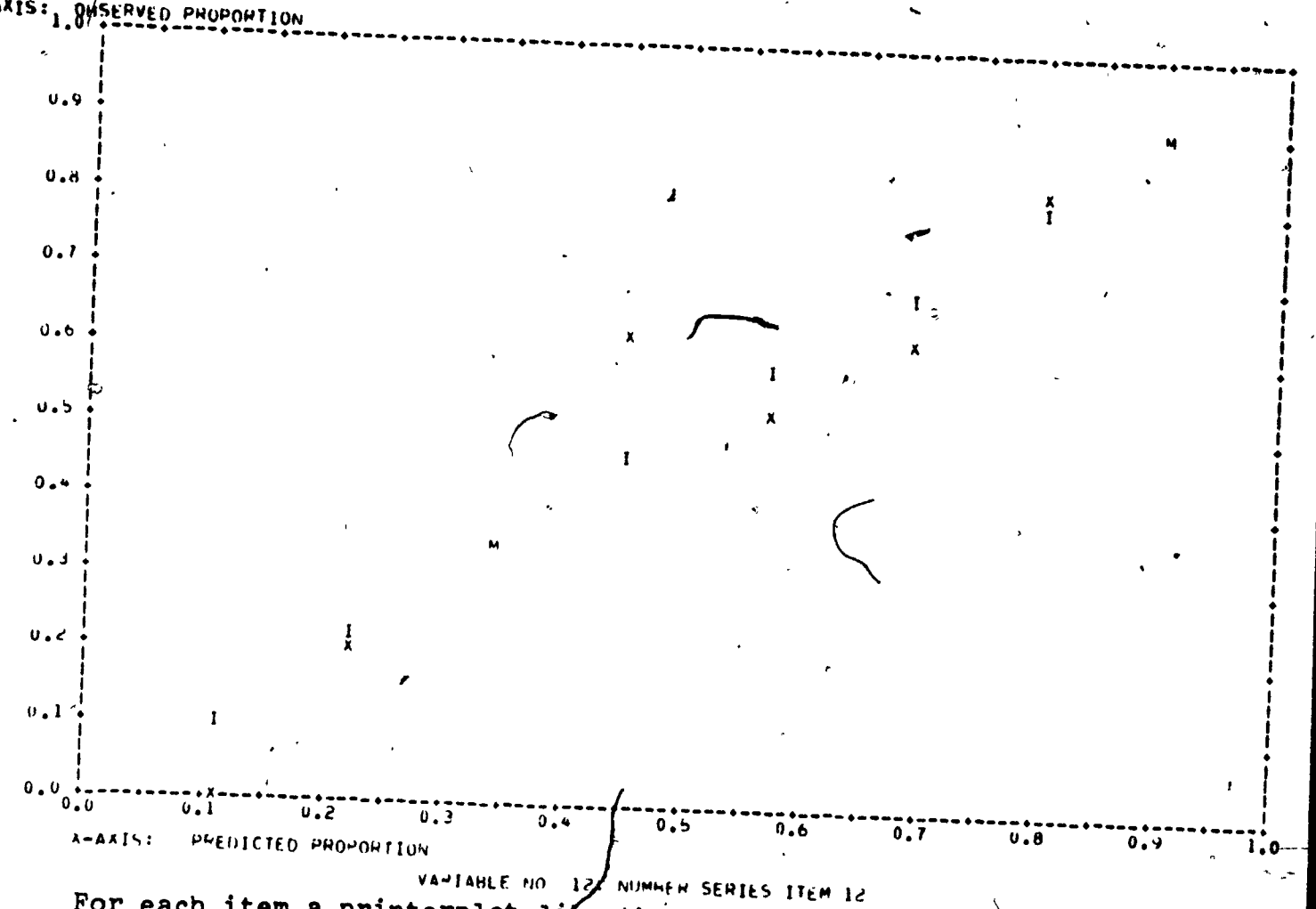
THE VALUES OF THE SYMMETRIC FUNCTIONS

ORDER	
1	0.9836284273254234D+01
2	0.4184849364742866360+02
3	0.1015340421594559D+03
4	0.1547101774616976D+03
5	0.1534318588065385D+03
6	0.1001617491097121D+03
7	0.4115407261513027D+02
8	0.9699397611944833D+01
9	0.1000000000000001D+01

VARIABLE NO 12: NUMBER SERIES ITEM 12

TIME GROUP	FREQUENCY	CORRECT ANSWERS	OBSERVED PROPORTION	PREDICTED PROPORTION	P-VALUE	
1	38	0	0.0	0.106	0.0144	TOO LOW OBSERVED PROP
2	35	7	0.200	0.219	0.4914	
3	53	18	0.340	0.335	0.5261	
4	56	35	0.625	0.454	0.0074	TOO HIGH OBSERVED PROP
5	65	34	0.523	0.571	0.2534	
6	60	37	0.617	0.686	0.1554	
7	77	63	0.818	0.796	0.3756	
8	85	77	0.906	0.901	0.5310	

The table shows the results from the binomial tests. There is of course one table for each item.



For each item a printerplot like the one above is produced. The symbol X is used to mark along the Y-axis the observed proportion, and the symbol I is used to indicte the predicted proportion, i.e. the I's are placed on the diagoral. When the X and the I coincide an M is printed.

THE SCORE GROUPS CONTRIBUTE TO THE CHI-SQUARE SUM AS FOLLOWS:

SCORE GROUP	NUMBER OF OBSERVATIONS	CONTRIBUTION
1	38	13.054
2	35	6.161
3	53	10.573
4	56	15.750
5	65	3.478
6	60	5.493
7	77	5.815
8	85	5.705

THE MARTIN-LOF CHI-SQUARE GOODNESS OF FIT TEST GIVES CHI-SQUARE= 66.032 WITH 56 DEGREES OF FREEDOM. P=0.16886.
THE REDUNDANCY IS: 0.0195358

THE MINIMUM NUMBER OF OBSERVATIONS WITHIN EACH GROUP ALLOWED WHEN COMPUTING THE LIKELIHOOD RATIO TEST IS 100
THE FOLLOWING GROUPING HAS BEEN USED

RANGE	NUMBER OF OBSERVATIONS	
1 - 3	126	NUMBER OF ITERATIONS FOR CONVERGENCE: 7
4 - 6	181	NUMBER OF ITERATIONS FOR CONVERGENCE: 8
7 - 8	162	NUMBER OF ITERATIONS FOR CONVERGENCE: 10

THE LIKELIHOOD RATIO GOODNESS OF FIT TEST GIVES CHI-SQUARE= 24.016 WITH 15 DEGREES OF FREEDOM. P=0.06415.
THE REDUNDANCY IS: 0.0071053

7.7 The source code of the non-OSIRIS version of PML

```
C*****************************************************************
C
C       PPPPPP      M      M    L
C       P     P     MM    MM    L      -- A PROGRAM WRITTEN BY JAN-ERIC
C       P     P     M M  M M    L        GUSTAFSSON, INSTITUTE OF
C       PPPPPP      M  M  M    L        EDUCATION, UNIVERSITY OF
C       P           M     M    L        GOTEBORG, DECEMBER 14, 1977.
C       P           M     M    L
C       P           M     M    LLLLLL
C
C*****************************************************************
C THE PROGRAM COMPUTES CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATES OF
C PARAMETERS IN THE DICHOTOMOUS RASCH MODEL AND PERFORMS GOODNESS OF
C FIT TESTS.
C
C REFERENCES:
C ALLERUP,P., & SORBER,G.(1977) THE RASCH MODEL FOR QUESTIONNAIRES. WITH
C    A COMPUTER PROGRAM(2ND ED.). DANISH INSTITUTE FOR EDUCATIONAL
C    RESEARCH, COPENHAGEN, 1977.+
C FISCHER,G.H. (1974) EINFUHRUNG IN DIE THEORIE PSYKOLOGISCHEN TESTS.
C    BERN: VERLAG HANS HUBER.
C MARTIN-LOF,P.(1973) STATISTISKA MODELLER (STATISTICAL MODELS).
C    MIMEOGRAPHED. INSTITUTET FOR FORSAKRINGSMATEMATIK OCH MATEMATISK
C    STATISTIK, STOCKHOLMS UNIVERSITET.
C
C*****************************************************************
C
C  THIS IS THE NON-OSIRIS VERSION OF PML.
C
C*****************************************************************
C
C  THE CONTROL CARDS ARE READ FROM UNIT 1
C
C  PRINTED OUTPUT ON UNIT 6
C   DICTIONARY ON UNIT 13
C   DATA ON UNIT 14
C
C
C
C  WHEN THE PARAMETER TEST=CHIS IS CHOSEN A SCRATCH FILE
C  (UNIT 12) FOR PRELIMINARY READING AND WRITING IS USED FOR LARGER PROBLEMS.
C
C*****************************************************************
C  THE FOLLOWING DIMENSION STATEMENT SHOULD DEFINE AN ARRAY THAT IS
C  AS LARGE AS THE AVAILABLE CORE ALLOWS. THE SIZE OF THE ARRAY
C  SHOULD ALSO BE REPRESENTED AS THE INTEGER CONSTANT IN THE STATEMENT
C  IMMEDIATELY PRECEDING THE CALL TO PMLCHI IN THE MAIN PROGRAM.
      DIMENSION STOR(1:3000)
C
      INTEGER LABEL(20),  ILOC(15),IVAL(15),ISTR(1),V(60),
     1VMO1(60),VMO2(60),NIS(60,60),NR(60),MO1(60),SW
      INTEGER*2 LIST(60),FLIFF(60)
      REAL*4 WK2(2,50),W(60)
      REAL*8 EPS(50),EPS1(60),G(50),G1(60,60),FLINF,WK3(3,60),
     1THETA(50),ERROR,SAVE(60),VARCOV(1:30)
      NKEY=15
      IPRINT=0
      INP=1
      IOUT=6
      KMAX=60
C
C  READ THE FILTER AND TITLE-CARDS
C
      SW=0
      READ(INP,1000) LABEL
1000  FORMAT(20A4)
      WRITE(6,1001) LABEL
1001  FORMAT('1',20X,20A4)
C
C  DEFINE AND INTERPRET THE PARAMETERS
C
      CALL GETKEY(IVAL,INP,IOUT,NKEY,ISTR)
      LISDIC=IVAL(1)
      MAXI=IVAL(2)
      ERROR=1.0/10**IVAL(3)
      METHOD=IVAL(4)
      IPREC=IVAL(5)
      ITEST=IVAL(6)
      IPUN=IVAL(7)
      IPLOT=IVAL(8)
      ISTART=IVAL(9)
      IFK=IVAL(10)
      MINST=IVAL(11)
      NORM=IVAL(12)
      JAHIL=IVAL(13)
      NORM=IVAL(14)
      IDESCR=IVAL(15)
C
C  READ THE VARIABLE LIST
      CALL ALIST(LIST,INP,IOUT,NV)
      K=NV
      IF(NV.GT.60.OR.NV.LE.2)GO TO 40
C
C  PRINT OUT THE CONTROL INFORMATION
```

```
      WRITE(6,1002)NV,MAXI,ERROR
 1002 FORMAT('0NUMBER OF ITEMS',I5,'.'),I5//' MAXIMUM NUMBER OF ITERATIO
     1NS.',I5//' CRITERION FOR CONVERGENCE.....',E7.)
      IF(METHOD.EQ.0)WRITE(6,1003)
 1003 FORMAT('0THE SYMMETRIC FUNCTIONS WILL BE COMPUTED WITH THE DIFFERE
     1NCE METHOD.')
      IF(METHOD.EQ.1)WRITE(6,1004)
 1004 FORMAT('0THE SYMMETRIC FUNCTIONS WILL BE COMPUTED WITH THE SUMMATI
     1ON METHOD.')
      IF(METHOD.EQ.1.AND.IPREC.EQ.0)WRITE(6,1005)
 1005 FORMAT(' DOUBLE PRECISION ARITHMETIC WILL BE USED.')
      IF(METHOD.EQ.1.AND.IPREC.EQ.1)WRITE(6,1006)
 1006 FORMAT(' SINGLE PRECISION ARITHMETIC WILL BE USED.')
      IF(ISTART.EQ.1)WRITE(6,1007)
 1007 FORMAT('0UNITIES WILL BE USED AS STARTVALUES FOR THE ITERATIONS.')
      IF(IEX.EQ.1)WRITE(6,1008)
 1008 FORMAT('0THE AITKEN EXTRAPOLATION WILL BE USED TO SPEED UP CONVERG
     1ENCE.')
      IF(IMIN.EQ.0.OR.IPLOT.EQ.0)WRITE(6,1012)NOBS
 1012 FORMAT('0SCORE GROUPS WITH',I3,' OR FEWER PERSONS ARE',
     1' NOT CONSIDERED IN THE PLOTS OR THE BINOMIAL TESTS.')
C
CC    READ THE DATA
C
      CALL READ(NV,LIST,V,VMD1,VMD2,LISDIC,KMAX,WK2,AOI,NR,NIS,NORM,
     1  NIND,INP,IOUT)
      IF(IDESCH.EQ.0)STOP
C
CC    COMPUTE START-VALUES ACCORDING TO THE VALUE OF ISTART
C
      CALL START(EPS,K,NR,AOI,ISTART,NORM,LIST)
C
CC    ESTIMATE THE ITEM PARAMETERS
C
      CALL PAREST(METHOD,IPREC,IEX,EPS,EPSI,G,GI,MAXI,ERROR,
     1AOI,NP,NIS,K,FLIKE,ITER,KITER,WK3,NORM)
      WRITE(6,1010)FLIKE
 1010 FORMAT('1-',30('*'),'1X,',*',2HX,'*',1X,'*',28X,'*',1X,'* LOGLIKE=',
     1D18.10,' *'/1X,'*',2HX,'*',1X,'*',28X,'*',1X,30('*'))
      DO 10 I=1,K
   10 SAVE(I)=EPS(I)
C
CC    ESTIMATE THE ABILITIES AND PRINT OUT THE PARAMETERS
C
      MAXP=4*MAXI
      CALL PERS(EPS,EPSI,THETA,K,MAXP,ERROR,NR,WK3,LIST,JAKIL,G)
C
CC    IF ASKED FOR, INVESTIGATE EACH ITEM
C
      IF(IMIN.EQ.0.OR.IPLOT.EQ.0)CALL ITTEST(IMIN,IPLOT,SAVE,G,GI,K,
     1NP,NIS,WK2,W,LIST,NOBS,IOUT)
C
CC    IF ASKED FOR, COMPUTE THE MARTIN-LOF CHI-SQUARE TEST
C
      KK1=K*(K-1)/2
C
C     THE INTEGER CONSTANT IN THE FOLLOWING STATEMENT SHOULD BE THE SIZE
C     OF THE ARRAY 'ISTOR' AS DIMENSIONED IN THE FIRST STATEMENT AND IT
C     SHOULD BE AS LARGE AS POSSIBLE.
C
      NTRAD=43000/KK1
      IF(ITEST.EQ.1.OR.ITEST.EQ.3)CALL PMLCHI(SAVE,D,GI,
     1EPSI,K,NIS,NP,VARKOV,THETA,METHOD,IPREC,LIST,STOR,NTRAD,
     2KK1,NIND,W,FLIKE)
C
CC    IF ASKED FOR, COMPUTE THE LIKELIHOOD RATIO TEST
C
      ERROR=1.0/10**IVAL(3)
      IF(ITEST.EQ.2.OR.ITEST.EQ.3) CALL FRACHI(METHOD,IPREC,IEX,
     1EPS,EPSI,G,GI,MAXI,ERROR,V,VMD1,NIS,K,FLIKE,WK3,NORM,NP,
     2VMD2,LISTA,MINST)
      GO TO 999
  900 WRITE(6,901)
  901 FORMAT('0TOO FEW OR TOO MANY ITEMS SPECIFIED')
      STOP 900
  910 WRITE(6,911)
  911 FORMAT('0ERROR ON THE PARAMETER CARD')
      STOP 910
  999 STOP
      END


      SUBROUTINE READ(K,LIST,V,VMD1,VMD2,LISDIC,KMAX,WK,AOI,NP,NIS,
     1  NIND,INP,IOUT)
C
C     READS THE DATA AND FORMS THE ITEM BY SCORE-GROUP FREQUENCY MATRIX.
C     POINT BISERIAL CORRELATIONS ARE ALSO COMPUTED.
C
      INTEGER GCLOSE
      DIMENSION WK(2,K)
      INTEGER*2 LIST(1)
      INTEGER SVN,TYPE,EVN,V(1),VMD1(1),VMD2(1),NAME(6),AOI(1),NR(1),
     1NIS(K,K)
      KM1=K-1
      MODE=0
      IPRCO=0
      NV=K
      SUM=0.0
      SUM2=0.0
      NIND=0
      NNULL=0
      NFULL=0
      DO 5 J=1,K
      DO 4 L=1,K
    4 NIS(J,L)=0
      NR(J)=0
      AOI(J)=0
      DO 5 I=1,2
    5 WK(I,J)=0.0
```

```
C
C READ THE DICTIONARY AND THE CASES
C
    1       CALL GETDIC(LISDIC,NV,INP,IOUT,LIST)
            CALL CASE(V,KL)
            IF(KL.NE.0)GO TO 11
            IR=0
            DO 10 I=1,K
   10       IR=IR+V(I)
            IF(IR.EQ.0.OR.IR.EQ.K)GO TO 2
            SUM=SUM+IR
            SUM2=SUM2+IR**2
C
C  INCREMENT THE NIS-MATRIX
C
            DO 20 I=1,K
   20       NIS(IR,I)=NIS(IR,I)+V(I)
            WK(I,V(I),I)=WK(I,V(I),I)+IR
            NIND=NIND+1
            GO TO 1
    2       IF(IR.EQ.0)NNULL=NNULL+1
            IF(IR.EQ.K)NFULL=NFULL+1
            GO TO 1
   11       CONTINUE
C
            NTOT=NIND+NNULL+NFULL
            WRITE(6,1002)NTOT,NNULL,NFULL,NIND
 1002       FORMAT(1H-,'NUMBER OF CASES READ',19('.'),I6//' NUMBER OF CASES WITH
           1A ZERO SCORE',6('.'),I6//' NUMBER OF CASES WITH A FULL SCORE',
           26('.'),I6//' NUMBER OF CASES REMAINING FOR ANALYSIS.',.,I6)
C
C  COMPUTE THE ROW AND COLUMN VECTORS
            DO 30 I=1,K
            DO 30 J=1,K
            AOI(I)=AOI(I)+NIS(J,I)
   30       NR(I)=NR(I)+NIS(I,J)
            DO 40 I=1,KM1
   40       NR(I)=NR(I)/I
C
C  SELECT ITEM FOR NORMATION
C
            IF(NORM.GT.0)GO TO 65
            DO 50 I=1,K
            V(I)=AOI(I)
   50       VMOI(I)=I
            DO 60 I=1,KM1
            L=K-I
            DO 60 J=1,L
            IJ=I+J
            IF(V(I)-V(IJ))60,60,55
   55       ISP1=V(I)
            ISP2=VMOI(I)
            V(I)=V(IJ)
            VMOI(I)=VMOI(IJ)
            V(IJ)=ISP1
            VMOI(IJ)=ISP2
   60       CONTINUE
            NORM=(K+1)/2
            NORM=VMOI(NORM)
            GO TO 69
   65       CONTINUE
            DO 67 I=1,K
            IF(NORM.EQ.LIST(I))GO TO 68
   67       CONTINUE
            WRITE(6,1013)
 1013       FORMAT('0THE ITEM SELECTED FOR UNITY NORMATION IS NOT IN
           1' THE VARIABLE LIST')
            STOP 916
   68       NORM=I
   69       CONTINUE
C
C  COMPUTE AND PRINT THE KR-20, THE POINT-BISERIALS AND PROPORTIONS OF
C  CORRECT ANSWERS.
C
            SD=SQRT((SUM2-(SUM**2/NIND))/(NIND-1))
            WRITE(6,1003)
 1003       FORMAT('1VAR.NO.   VARIABLE NAME',13X,'PROPORTION CORRECT',4X,
           1'POINT BISERIAL CORRELATION')
            RXX=0.
            DO 70 I=1,K
            PROP=AOI(I)/1.0/NIND
            RXX=RXX+PROP*(1.-PROP)
            RPBIS=((WK(I,I)/AOI(I)-WK(1,I)/(NIND-AOI(I)))/SD)*
           1SQRT(PROP*(1.0-PROP))
            CALL GNAME(I,NAME)
   70       WRITE(6,1004)LIST(I),NAME,PROP,RPBIS
 1004       FORMAT('0',I5,4X,5A4,F15.3,F15.3)
            WRITE(6,1005) LIST(NORM)
            RXX=K*(SD**2-RXX)/((K-1)*SD**2)
            WRITE(6,1011)RXX
 1011       FORMAT(////' THE RELIABILITY (KR-20) IS',F6.2)
 1005       FORMAT('0NORMATION ON VARIABLE',I5)
C
C  PRINT THE NIS-MATRIX
C
            WRITE(6,1006)
 1006       FORMAT('1',20X,'THE ITEM BY SCOREGROUP FREQUENCY MATRIX OF CORRECT
           1 ANSWERS')
            NPRINT=(KM1-1)/18+1
            IPRINT=1
    7       IF(IPRINT.EQ.NPRINT)GO TO 8
            L=IPRINT*18-17
            M=IPRINT*18
            WRITE(6,1007) (IP,IP=L,M)
 1007       FORMAT(////6X,18I6)
            WRITE(6,1010)
 1010       FORMAT(' VAR.NO')
```

```
      DO 80 J=1,K
80    WRITE(6,1008) LIST(J),(NIS(IR,J),IR=L,M)
      WRITE(6,1009) (NR(IR),IR=L,M)
1008  FORMAT(1H0,I5,1X,1916)
1009  FORMAT(1H0,5X,1816)
      IPRINT=IPRINT+1
      GO TO 7
8     L=IPRINT*18-17
      WRITE(6,1007) (IR,IR=L,KM1)
      WRITE(6,1010)
      DO 90 I=1,K
90    WRITE(6,1008) LIST(I),(NIS(IR,I),IR=L,KM1),AOI(I)
      DO 100 I=1,K
      IF(AOI(I)-NINO)101,99,99
101   IF(AOI(I))99,99,100
100   CONTINUE
      RETURN
99    WRITE(6,1012)
1012  FORMAT(1H0,'THERE ARE ZERO AND/OR PERFECT ITEM SCORES.')
      STOP 920
      END


      SUBROUTINE START(EPS,K,NR,AOI,NEJSTA,NORM,LIST)
C
C  COMPUTES START VALUES FOR THE ITERATIONS
C
      REAL*8 EPS(1),RSUM,DEL
      INTEGER NR(1),AOI(1)
      INTEGER*2 LIST(1)
      WRITE(6,1001)
1001  FORMAT(//////)
      IF(NEJSTA.EQ.1)GO TO 1
      RSUM=0.000D+00
      DEL=0.000D+00
      DO 10 I=1,K
      RSUM=RSUM+AOI(I)*1.00D+00
10    DEL=DEL+1.0D+00*NR(I)*(I*(K-I))/(K*(K-1))
      RSUM=RSUM/K
      DO 20 I=1,K
20    EPS(I)=DEXP((AOI(I)*1.00D+00-RSUM)/DEL)
      DEL=EPS(NORM)
      DO 21 I=1,K
      EPS(I)=EPS(I)/DEL
21    CONTINUE
      RETURN
1     DO 30 I=1,K
30    EPS(I)=1.00D+00
      RETURN
      END


      SUBROUTINE PAREST(METHOD,IPREC,IEX,EPS,EPSI,G,GI,MAXI,ERROR,
     1AOI,NR,NIS,K,FLIKE,ITER,KDIFF,WK3,NORM)
C
C  ESTIMATES THE ITEM PARAMETERS
C
      REAL*8 EPS(1),EPSI(1),G(1),GI(K,K),TEST,FLIKE,WK3(3,K),PR
      INTEGER AOI(1),NR(1),NIS(K,K)
      INTEGER*2 KDIFF(1)
      LOOP=0
      FLIKE=0.000
      DO 1 ITER=1,MAXI
      IF(METHOD.EQ.0)CALL GAMMA(EPS,G,GI,K)
      IF(METHOD.EQ.1)CALL GAMMA2(EPS,G,GI,K,IPREC)
      IF(IEX.EQ.0)CALL IMPRV(EPS,EPSI,G,GI,K,IPREC,K,AOI,NR,ERROR,
     1NORM)
      IF(IEX.EQ.1)CALL AITKEN(WK3,EPS,EPSI,G,GI,K,NR,AOI,ERROR,
     1KDIFF,LOOP,ITER,NORM)
      NDIFF=0
      DO 30 I=1,K
30    NDIFF=NDIFF+KDIFF(I)
      IF(NDIFF.EQ.0)GO TO 20
10    CONTINUE
20    WRITE(6,1001) ITER
1001  FORMAT(1H0,'NUMBER OF ITERATIONS FOR CONVERGENCE:',I5)
      PR=1.000
      DO 40 I=1,K
40    PR=PR*EPSI(I)
      PR=DEXP(DLOG(PR)/K)
C
C  CARRY OUT THE PRODUCTNORMING
C
      DO 50 I=1,K
      EPS(I)=EPSI(I)/PR
      WK3(1,I)=EPSI(I)
      WK3(2,I)=EPS(I)
50    WK3(3,I)=-DLOG(EPS(I))
C
C  COMPUTE THE SYMMETRIC FUNCTIONS FOR THE PRODUCT NORMED PARAMETERS
C
      IF(METHOD.EQ.0)CALL GAMMA(EPS,G,GI,K)
      IF(METHOD.EQ.1)CALL GAMMA2(EPS,G,GI,K,IPREC)
C
C  COMPUTE THE LOG LIKELIHOOD
      DO 60 I=1,K
60    FLIKE=FLIKE+AOI(I)*DLOG(EPS(I))-NR(I)*DLOG(G(I))
      RETURN
      END
```

```
      SUBROUTINE IMPROV(EPS,EPSI,G,GI,KDIFF,K,AOI,NR,ERROR,NORM)
C
C  COMPUTES NEW VALUES OF THE ITEM PARAMETERS IN THE ITERATION
C
      REAL*8 EPS(I),EPSI(I),G(I),GI(K,K),D,ERROR
      INTEGER NR(I),AOI(I)
      INTEGER*2 KDIFF(I)
      DO 10 I=1,K
      EPSI(I)=0.000
      DO 20 J=1,K
20    EPSI(I)=EPSI(I)+NR(J)*GI(I,J)/G(J)
10    EPSI(I)=AOI(I)/EPSI(I)
      D=EPSI(NORM)
      DO 30 I=1,K
      KDIFF(I)=0
      EPSI(I)=EPSI(I)/D
      IF(DABS(EPSI(I)-EPS(I)).GT.ERROR)KDIFF(I)=1
30    EPS(I)=EPSI(I)
      RETURN
      END



      SUBROUTINE GAMMA(EPS,G,GI,N)
C
C  COMPUTES THE SYMMETRIC FUNCTIONS WITH A RECURSIVE FORMULA THAT USES
C  SUBTRACTIONS. THE ALGORITHM IS SENSITIVE TO ROUND-OFF ERRORS WHEN THE
C  NUMBER OF ITEMS IS LARGE.
C  THE ROUTINE IS TAKEN FROM FISCHER (1974).
C
      REAL*8 G(I),GI(N,N)
      REAL*8 EPS(I),TEST
C     GI(I,J) I=1,....,N J=0,...,N-1 GRUNDFUNKTION J-TER ORDNUNG OHNE I
C
      NORM=(N+1)/2
      NORM1=NORM+1
      G(1)=0.
      DO 200 I=1,N
      GI(I,1)=1.
200   G(1)=G(1)+EPS(I)
      DO 230 J=2,NORM1
      DO 210 I=1,N
210   GI(I,J)=G(J-1)-GI(I,J-1)*EPS(I)
      G(J)=0.
      DO 220 I=1,N
220   G(J)=G(J)+GI(I,J)*EPS(I)
230   G(J)=G(J)/DFLOAT(J)
      TEST=G(NORM1)
      G(N)=1.
      DO 250 I=1,N
      GI(I,N)=1.
      DO 240 J=1,N
      IF(I.EQ.J)GO TO 240
      GI(I,N)=GI(I,N)*EPS(J)
240   CONTINUE
250   G(N)=G(N)*EPS(I)
      J1=N-NORM-1
      DO 270 J=1,J1
      G(N-J)=0.
      DO 260 I=1,N
260   G(N-J)=G(N-J)+GI(I,N-J+1)
      G(N-J)=G(N-J)/DFLOAT(J)
      DO 270 I=1,N
270   GI(I,N-J)=(G(N-J)-GI(I,N-J+1))/EPS(I)
      TEST=1.-TEST/G(NORM1)
      IF(DABS(TEST)-1.D-4)310,310,280
280   PRINT 500,TEST
      STOP 920
310   RETURN
500   FORMAT('0COMPUTATIONAL ACCURACY TOO LOW IN GAMMA',D16.7)
      END



      SUBROUTINE GAMMA2(EPS,G,GI,K,IPREC)
C
C  THE ROUTINE SUPERVISES THE COMPUTATION OF THE SYMMETRIC FUNCTIONS
C  AND THEIR DERIVATIVES WITH THE SUMMATION METHOD.
C
      REAL*8 EPS(I),G(I),GI(K,K),STORE
      KM1=K-1
      DO 10 I=1,K
      STORE=EPS(I)
      EPS(I)=0.000
      IF(IPREC.EQ.0)CALL GAM(EPS,K,G)
      IF(IPREC.EQ.1)CALL GAME(EPS,K,G)
      GI(I,1)=1.000
      DO 20 J=1,KM1
20    GI(I,J+1)=G(J)
      EPS(I)=STORE
10    CONTINUE
      CALL GAM(EPS,K,G)
      RETURN
      END
```

```
      SUBROUTINE GAM(EPS,K,F)
C
C  THE ROUTINE COMPUTES THE SYMMETRIC FUNCTIONS WITH A RECURSIVE FORMULA
C  THAT ONLY USES MULTIPLICATION AND ADDITION OF POSITIVE NUMBERS.
C  DOUBLE PRECISION ARITHMETIC IS USED.
C  THE ROUTINE IS TAKEN FROM FISCHER (1974).
C
      REAL*8 EPS(1),F(1),X(60),Y(60)
      DO 10 I=1,K
10    E(I)=EPS(I)
      X(1)=E(1)
      DO 30 I=2,K
      X(I)=0.0D0
      Y(1)=X(1)*E(1)
      DO 20 J=2,I
20    Y(J)=X(J)*X(J-1)*E(I)
      DO 30 J=1,I
30    X(J)=Y(J)
      DO 40 J=1,K
40    F(J)=X(J)
      RETURN
      END


      SUBROUTINE GAME(EPS,K,G)
C
C  THIS IS A SINGLE PRECISION VERSION OF THE GAM ROUTINE.
C
      REAL*8 EPS(1),G(1)
      REAL*4 E(60),X(60),Y(60)
      DO 10 I=1,K
10    E(I)=SNGL(EPS(I))
      X(1)=E(1)
      DO 30 I=2,K
      X(I)=0.0
      Y(1)=X(1)*E(I)
      DO 20 J=2,I
20    Y(J)=X(J)*X(J-1)*E(I)
      DO 30 J=1,I
30    X(J)=Y(J)
      DO 40 J=1,K
40    G(J)=DBLE(X(J))
      RETURN
      END


      SUBROUTINE AITKEN(A,EPS,EPSI,G,GI,KK,NN,AOI,ERROR,KDIFF,LOOP,
     1ITER,NORM)
C
C  THE ROUTINE COMPUTES THE AITKEN EXTRAPOLATION TO SPEED UP CONVERGENCE
C
      REAL*8 A(3,K),EPS(1),EPSI(1),G(1),GI(K,2),H,HD
      INTEGER NN(1),AOI(1)
      INTEGER*2 KDIFF(1)
      IF(ITER.LT.4)GO TO 2
      NDIFF=0
      LOOP=LOOP+1
      CALL IMPROV(EPS,EPSI,G,GI,KDIFF,K,AOI,NN,ERROR,NORM)
      DO 10 I=1,K
      NDIFF=NDIFF+KDIFF(I)
10    A(LOOP,I)=EPS(I)
      IF(NDIFF.EQ.0)RETURN
      IF(LOOP.LT.3)RETURN
      DO 20 I=1,K
      IF(KDIFF(I).EQ.0)GO TO 20
      HD=A(3,I)-A(2,I)
      H=ABS(HD/(A(1,I)+A(3,I)-2*A(2,I)))
      IF(H.GT.1.0D0)H=1.0D0
      EPS(I)=A(3,I)+HD*H
20    CONTINUE
      DO 30 I=1,K
30    A(1,I)=EPS(I)
      LOOP=1
      RETURN
    2 CALL IMPROV(EPS,EPSI,G,GI,KDIFF,K,AOI,NN,ERROR,NORM)
      RETURN
      END


      SUBROUTINE PERS(E,D,THETA,K,MAX,FEHLER,NP,WB,LIST,JAHIL,G)
C
C  ESTIMATES THE PERSON PARAMETERS ITERATIVELY USING THE NEWTON-RAPHSON
C  METHOD. THE ROUTINE IS AN IMPROVED VERSION OF THE ONE PRESENTED BY
C  FISCHER (1974).
C
      REAL*8 E(1),D(1),THETA(1),FEHLER,SUMM,AK1(3,K),G(1),
     1ERR,Z1,Z2,LOG
      INTEGER NN(1)
      INTEGER*2 LIST(1)
      KK=K-1
      IF(JAHIL.EQ.1)GO TO 201
      IF(MAX.EQ.1)GO TO 20
C
C  DETERMINE THE RANGE OF VARIATION OF THE ITEM PARAMETERS
C
      Z1=E(1)
      Z2=Z1
      DO 4 I=2,K
      IF(E(I)-Z1)1,2,2
1     Z1=E(I)
2     IF(E(I)-Z2)4,4,3
3     Z2=E(I)
4     CONTINUE
      Z2=DLOG(Z2)-DLOG(Z1)
```

```
      IF(Z2.LT.2.0D0)GO TO 8
C
C  COMPUTE START VALUES UNDER THE ASSUMPTION OF EQUALLY SPACED
C  ITEM PARAMETERS
C
      DO 5 I=1,KK
      Z1=1.0*I/K
      D(I)=DEXP(Z2*(Z1-.5001))
5     D(I)=D(I)*(1.000-DEXP(-Z1*Z2))/(1.000-DEXP(-Z2*(1.0D0-Z1)))
      GO TO 7
C
C  COMPUTE START VALUES UNDER THE ASSUMPTION OF EQUAL ITEM PARAMETERS
C
8     CONTINUE
      DO 6 I=1,KK
      Z1=1.0*I/K
6     D(I)=Z1/(1.0D0-Z1)
7     CONTINUE
   20 DO 60 ITER=1,MAX
      NDIF=0
      DO 40 J=1,KK
      SU=0.
      DO 30 I=1,K
   30 SU=SU+E(I)/(1.+F(I)*D(J))
      DD=J
   40 THETA(J)=DD/SU
      DO 50 J=1,KK
      IF(DABS(THETA(J)-D(J)).GT.FEHLER)NDIF=NDIF+1
   50 D(J)=THETA(J)
      IF(NDIF.EQ.0)GO TO 70
   60 CONTINUE
   70 IF(MAX.EQ.1)RETURN
C
C  COMPUTE THE STANDARD ERRORS OF THE ITEM PARAMETERS AND PRINT OUT THE
C  INFORMATION.
C
      CALL ITINFO(WK3,D,NR,K,LIST)
      WRITE(6,1001)
1001  FORMAT('1',30X,'ABILITY PARAMETERS'//6X,'PRODUCT NORMATION   PRODU
     1CT NORMATION(LOG)   STANDARD ERROR   CONFIDENCE INTERVAL(95 %)'/
     2' SCORE')
      FEHLER=0.DO
      SU=0.DO
      DD=0.DO
      NIND=0
      DO 110 I=1,KK
      INF=0.0D+00
      DO 111 J=1,K
  111 INF=INF+(E(J)*D(I))/(1.0+E(J)*D(I))**2
      LOG=DLOG(D(I))
      FEHLER=FEHLER+NR(I)/INF
      SU=SU+NR(I)*LOG
      DD=DD+NR(I)*LOG**2
      NIND=NIND+NR(I)
      INF=1./DSQRT(INF)
      Z1=LOG-1.96*INF
      Z2=LOG+1.96*INF
  110 WRITE(6,1002)I,D(I),LOG,INF,Z1,Z2
 1002 FORMAT('0',I3,F15.5,F19.5,F24.5,F19.5,F12.5)
      WRITE(6,1003) ITER
 1003 FORMAT('0NUMBER OF ITERATIONS FOR CONVERGENCE OF THE ABILITIES:'
     1I5,'.')
      DD=100-(SU**2/NIND))/(NIND-1)
      FEHLER=FEHLER/NIND
      SU=SU/NIND
      FEHLER=(DD-FEHLER)/DD
      WRITE(6,1005)SU,DD
      WRITE(6,1006)FEHLER
 1005 FORMAT('0ON THE LOG SCALE THE MEAN OF THE SAMPLE IS',F6.2,' WITH T
     1HE VARIANCE',F7.2)
      IF(K.GT.30)GO TO 125
      WRITE(6,1012)
      NULL=0
      WRITE(6,1013) NULL,(I,I=1,K)
      WRITE(6,1014)
      DO 120 I=1,KK
      DD=1.0D0
      DO 130 J=1,K
      Z1=1.000+E(J)*D(I)
      DD=DD*Z1
130   CONTINUE
      DO 140 J=1,K
140   THETA(J)=E(J)*D(I)**J/DD
      Z1=1.0E0/DD
      WRITE(6,1015)I,(Z1,(THETA(J),J=1,K)
120   CONTINUE
125   CONTINUE
      WRITE(6,1010)
      DO 150 I=1,K
150   WRITE(6,1011) I,G(I)
      RETURN
1006  FORMAT('0THE INDEX OF SUBJECT SEPARATION IS',F5.2)
201   WRITE(6,1007)
1007  FORMAT('1',30X,'ITEM PARAMETERS'//5X,'UNITY NORMATION   PRODUCT NO
     1RMATION   PRODUCT NORMATION LOG')
      DO 210 I=1,K
210   WRITE(6,1008) LIST(I),(WK3(J,I),J=1,3)
1008  FORMAT('0',I5,F14.5,F18.5,F19.5)
1010  FORMAT('1          THE VALUES OF THE SYMMETRIC FUNCTIONS'//' ORDER'
     1)
1011  FORMAT('0',I3,G32.16)
1012  FORMAT('1',15X,'PROBABILITY OF OBTAINING A CERTAIN RAW SCORE GIVEN
     1 THE PERSON PARAMETER'//20X,'RAW SCORE')
1013  FORMAT('0',10X,20I5.5/(11X,20I5))
1014  FORMAT(' SCORE  PARAMETER')
1015  FORMAT('0',I3,F11.5,3X,20E5.2,5/(14X,20E5.2))
      END
```

```
      SUBROUTINE ITINFO(WK3,PERS,NR,K,LIST)
C
C  COMPUTES THE STANDARD ERRORS OF THE ITEM PARAMETERS AND PRINTS OUT
C  THE INFORMATION ABOUT THE ITEM PARAMETERS.
C
      REAL*8 WK3(3,K),PERS(1),INF,Z1,Z2
      INTEGER NR(1)
      INTEGER*2 LIST(1)
      KM1=K-1
      WRITE(6,1001)
      DO 10 I=1,K
      INF=0.0D+00
      DO 20 J=1,KM1
   20 INF=INF+NR(J)*(WK3(2,I)*PERS(J))/(1.0+WK3(2,I)*PERS(J))**2
      INF=1./DSQRT(INF)
      Z1=WK3(3,I)-1.96*INF
      Z2=WK3(3,I)+1.96*INF
10    WRITE(6,1002) LIST(I),(WK3(J,I),J=1,3),INF,Z1,Z2
      RETURN
1001  FORMAT('1',30X,'ITEM PARAMETERS'//9X,'UNITY NORMATION   PRODUCT NO
     1RMATION   PRODUCT NORMATION(LOG)   STANDARD ERROR   CONFIDENCE INT
     2ERVAL(95 %)'/' VAR.NO')
1002  FORMAT('0',I5,F14.5,F18.5,F19.5,F24.5,F19.5,F12.5)
      END



      SUBROUTINE ITTEST(IRIN,IPLOT,EPS,G,GI,K,NR,NIS,WK2,W,LIST,NOBS,
     1IOUT)
C
C  PLOTS FOR EACH ITEM OBSERVED AGAINST PREDICTED PROPORTIONS OF CORRECT
C  ANSWERS WITHIN SCORE GROUPS. OPTIONALLY A BINOMIAL TEST ON THE
C  FREQUENCIES IS CARRIED OUT.
C
      REAL*8 EPS(1),G(1),GI(K,K)
      INTEGER NR(1),NIS(K,K),NAME(6)
      INTEGER*2 LIST(1)
      DIMENSION A(47),WK2(K,2),W(1)
      DATA A/20*'   ','PRED','ICTE','D PR','OPOR','TION',5*'   ',
     1'OPSE','RVED',' PRO','PORT','ION ',5*'   ',0.0,1.0,0.0,1.0,
     23*'XXX'/
      M=2
      INC=1
      KM1=K-1
C
C  START THE LOOP OVER THE ITEMS
C
      DO 10 I=1,K
      NR=0
      IF(IRIN.EQ.1)GO TO 2
      CALL GNAME(I,NAME)
      WRITE(6,1001) LIST(I),NAME
1001  FORMAT('1',40X,'VARIABLE NO',I4,': ',6A4//'SCORE GROUP   FREQUENCY
     1  CORRECT ANSWERS   OBSERVED PROPORTION   PREDICTED PROPORTION   P
     2-VALUE'/)
C
C  LOOP OVER SCORE GROUPS
C
2     DO 20 J=1,KM1
      IF(NR(J)-NOBS)11,11,12
11    NR=NR+1
      GO TO 20
12    CONTINUE
      W(J-NR)=EPS(I)*GI(I,J)/G(J)
      WK2(J-NR+1)=W(J-NR)
      WK2(J-NR,2)=NIS(J,I)*1.0/NR(J)
      IF(IRIN.EQ.1)GO TO 20
      NOBS=NIS(J,I)
      PP=W(J-NR)
C
C  IF THE OBSERVED FREQUENCY IS HIGHER THAN THE PREDICTED, INVESTIGATE THE
C  UPPER PART OF THE DISTRIBUTION
C
      IF(WK2(J-NR+1)-WK2(J-NR,2))3,4,4
3     NOBS=NR(J)-NOBS
      PP=1.0-PP
4     P=OBBIN(NOBS,NR(J),PP)
      WRITE(6,1002) J,NR(J),NIS(J,I),WK2(J-NR,2),W(J-NR),P
      IF(WK2(J-NR+1).GT.WK2(J-NR,2).AND.P.LT..025)WRITE(6,1004)
      IF(WK2(J-NR+1).LT.WK2(J-NR,2).AND.P.LT..025)WRITE(6,1005)
1004  FORMAT('+',96X,'TOO LOW  OBSERVED PROP.')
1005  FORMAT('+',96X,'TOO HIGH OBSERVED PROP.')
1002  FORMAT(I7,2I14,F20.3,F23.3,F17.4)
20    CONTINUE
      IF(IPLOT.NE.0)GO TO 10
      KMR=KM1-NR
      CALL PLOTT(W,WK2,KMR,M,INC,K,A,IOUT)
      CALL GNAME(I,NAME)
      WRITE(6,1003) LIST(I),NAME
1003  FORMAT('0',40X,'VARIABLE NO',I4,': ',6A4)
10    CONTINUE
      RETURN
      END
```

```
      FUNCTION UPIBIN(M,N,P)
C     UPIBIN COMPUTES THE CUMULATIVE BINOMIAL PROBABILITY DISTRIBUTION
C     BY A METHOD SUGGESTED BY SOREN SPELLING.
C
C     PARAMETERS:
C
C       M  INTEGER VARIABLE WHOSE CUMULATIVE PROBABILITY IS TO BE COMPUTED
C
C       N  INTEGER NUMBER OF TRIALS IN THE DISTRIBUTION
C
C       P  REAL PROBABILITY OF SUCCESS IN ONE TRIAL
C
C     CODED BY GORDON SORHER, DPI, (DANISH INSTITUTE FOR EDUCATIONAL
C     RESEARCH), COPENHAGEN, 26 AUGUST, 1976.
      FN=FLOAT(N)
      FN1=FN+1.0
      J0=IFIX(FN*P+0.5001)
      JI=J0+1
      P1=(1.0-P)/P
      S=0.0
      S1=0.0
      S2=0.0
      T=0.0
      IF(J0.EQ.0)GO TO 16
      D=1.0
C
      DO 14 I=1,J0
      J=J0-I+1
      D=D*J*P1/(FN1-J)
      S=S+D
      IF(S1-S)10,16,16
10    CONTINUE
      S1=S
      IF(J-M-1)12,12,14
12    CONTINUE
      T=T+D
14    CONTINUE
16    CONTINUE
C
      IF(J0.EQ.N)GO TO 26
      S=0.0
      D=1.0
      P1=1.0/P1
C
      DO 24 J=J1,N
      D=D*(FN1-J)*P1/J
      S=S+D
      IF(S2-S)20,26,26
20    CONTINUE
      S2=S
      IF(J-M)22,22,24
22    CONTINUE
      T=T+D
24    CONTINUE
26    CONTINUE
      IF(M.GE.J0)T=T+1.0
C
      UPIBIN=T/(1.0+S1+S2)
C
      RETURN
      END



      SUBROUTINE PMLCHI(EPS,G,GI,G2,K,NIS,NR,VARKOV,VEK,
     1METHOD,IPREC,LIST,STOR,NTRAD,KK1,NIMD,W,FLIKE)
C
C  THE ROUTINE ADMINISTERS THE COMPUTATION OF THE CHI-SQUARE GOODNESS
C  OF FIT TEST SUGGESTED BY MARTIN-LOF (1973).
C
      INTEGER*2 LIST(1)
      INTEGER NIS(K,K),NR(1)
      REAL*8 EPS(1),G(1),GI(K,K),VARKOV     VEK(1),G2(1),
     1CHI,M,THETA,ALF,FLIKE
      REAL*4 STOR(NTRAD,KK1),W(1)
      IER2=0
      KS=K-1
      KM1=K-1
      CHI=0.000
      WRITE(6,1007)
1007  FORMAT('1THE SCORE GROUPS CONTRIBUTE TO THE CHI-SQUARE SUM AS FOLL
     1OWS:'//' SCORE GROUP     NUMBER OF OBSERVATIONS     CONTRIBUTION')
C
C  DETERMINE HOW MANY TIMES IT IS NECESSARY TO FILL THE MATRIX WHERE
C  THE SECOND DERIVATIVES OF THE SYMMETRIC FUNCTIONS ARE STORED.
C
111   NLOOP=KM1/NTRAD
      IORD=0
      IF(NLOOP.EQ.0)GO TO 141
      NRAD=NTRAD
C
C  CALL THE ROUTINE WHERE THE COMPUTATIONS ARE CARRIED OUT.
C
      DO 140 LOOP=1,NLOOP
      CALL STORVA(EPS,G,GI,K,KK1,VEK,G2,VARKOV,NTRAD,CHI,NRAD,LOOP,
     1KS,STOR,IORD,NR,NIS,IPREC,W,IER2)
      IF(IER2.NE.0)RETURN
140   CONTINUE
C
C  IF THERE ARE ANY SCORE-GROUPS LEFT, TAKE CARE OF THEM.
C
141   IF(IORD-KM1)142,130,130
142   LOOP=KM1/NTRAD+1
      NRAD=KM1-IORD
      CALL STORVA(EPS,G,GI,K,KK1,VEK,G2,VARKOV,NTRAD,CHI,NRAD,LOOP,
     1KS,STOR,IORD,NR,NIS,IPREC,W,IER2)
```

```
      IF(IER2.NE.0)RETURN
  130 KS=KS-1
      NDF=KS*(K-1)
      CALL MDCDF1(CHI,NDF,ALF,IER)
      WRITE(6,1003) CHI,NDF,ALF
 1003 FORMAT(////' THE MARTIN-LOF CHI-SQUARE GOODNESS OF FIT TEST GIVES
     1CHI-SQUARE=',F14.3,' WITH',I6,' DEGREES OF FREEDOM, P=',F7.5,'.')
C
C  COMPUTE THE REDUNDANCY
C
      H=-CHI/(2.0*FLIKE)
      WRITE(6,100H) H
 100H FORMAT('0THE REDUNDANCY IS:',F20.7)
      RETURN
      END



      SUBROUTINE DSINV(A,N,EPS,IER)
C
C  THIS SSP-ROUTINE INVERTS A SYMMETRIC POSITIVE DEFINITE MATRIX
C
      REAL*8 A(1),DIN,WORK
      CALL DMFSD(A,N,EPS,IER)
      IF(IER)9,1,1
    1 IPIV=N*(N+1)/2
      IND=IPIV
      DO 6 I=1,N
      DIN=1.D0/A(IPIV)
      A(IPIV)=DIN
      MIN=N
      KEND=I-I
      LANF=N-KEND
      IF(KEND)5,5,2
    2 J=IND
      DO 4 K=1,KEND
      WORK=0.D0
      MIN=MIN-1
      LHOR=IPIV
      LVER=J
      DO 3 L=LANF,MIN
      LVER=LVER+1
      LHOR=LHOR+L
    3 WORK=WORK+A(LVER)*A(LHOR)
      A(J)=-WORK*DIN
    4 J=J-MIN
    5 IPIV=IPIV-MIN
    6 IND=IND-1
      DO 8 I=1,N
      IPIV=IPIV+I
      J=IPIV
      DO 8 K=I,N
      WORK=0.D0
      LHOR=J
      DO 7 L=K,N
      LVER=LHOR+K-I
      WORK=WORK+A(LHOR)*A(LVER)
    7 LHOR=LHOR+L
      A(J)=WORK
    8 J=J+K
    9 RETURN
      END



      SUBROUTINE DMFSD(A,N,EPS,IER)

C  THIS SSP-ROUTINE FACTORS A GIVEN SYMMETRIC POSITIVE DEFINITE MATRIX.
C  IT IS CALLED BY DSINV.
C
      REAL*8 A(1),DPIV,DSUM
      IF(N-1)12,1,1
    1 IER=0
      KPIV=0
      DO 11 K=1,N
      KPIV=KPIV+K
      IND=KPIV
      LEND=K-1
      TOL=ABS(EPS*SNGL(A(KPIV)))
      DO 11 I=K,N
      DSUM=0.D0
      IF(LEND)2,4,2
    2 DO 3 L=1,LEND
      LANF=KPIV-L
      LIND=IND-L
    3 DSUM=DSUM+A(LANF)*A(LIND)
    4 DSUM=A(IND)-DSUM
      IF(I-K)10,5,10
    5 IF(SNGL(DSUM)-TOL)6,6,9
    6 IF(DSUM)12,12,7
    7 IF(IER)8,8,9
    8 IER=K-1
    9 DPIV=DSQRT(DSUM)
      A(KPIV)=DPIV
      DPIV=1.D0/DPIV
      GO TO 11
   10 A(IND)=DSUM*DPIV
   11 IND=IND+I
      RETURN
   12 IER=-1
      RETURN
      END
```

```
      SUBROUTINE STORVA(EPS,G,G1,K,KK1,VEK,G2,VARKOV,NTRAD,CHI,
     1NRAD,LOOP,KS,STOR,IORD,NR,NIS,IPREC,W,IER2)
C
C  THIS ROUTINE IS CALLED BY PMLCHI. THE SECOND DERIVATIVES OF THE
C  SYMMETRIC FUNCTIONS ARE STORED AS SINGLE PRECISION NUMBERS IN THE STOR
C  ARRAY. WHEN THIS ARRAY IS TOO SMALL USE IS MADE OF A SCRATCH FILE
C  (UNIT 12) FOR BINARY WRITING AND READING.
C
      REAL*8 EPS(1),G(1),G2(1),VEK(1),G1(K,K),STOI,STOJ,CHI,CHI2,
     1VARKOV(1)
      REAL*4 STOR(NTRAD,KK1),W(1)
      INTEGER NR(1),NIS(K,K)
      TOL=1.E-5
      KM1=K-1
      KM2=K-2
      NLOOP=KM1/NTRAD
      JL=0
      IF(LOOP.GT.1)GO TO 11
C
C  COMPUTE THE SECOND DERIVATIVES OF THE SYMMETRIC FUNCTIONS THROUGH
C  REPEATED CALLS TO THE GAM (OR GAME) ROUTINE
C
      DO 10 I=2,K
      L=I-1
      DO 10 J=1,L
      JL=JL+1
      STOI=EPS(I)
      STOJ=EPS(J)
      EPS(I)=0.D0
      EPS(J)=0.D0
      IF(IPREC)17,17,18
17    CALL GAM(EPS,K,G2)
      GO TO 19
18    CALL GAME(EPS,K,G2)
19    W(1)=0.
      W(2)=1.0
      DO 20 LORD=1,KM2
20    W(LORD+2)=G2(LORD)
      EPS(I)=STOI
      EPS(J)=STOJ
C
C  COPY THE INFORMATION FOR THE SCORE GROUPS BEING TREATED IN THIS LOOP
C  INTO THE STOR ARRAY
C
      IORD=NTRAD*(LOOP-1)
      DO 30 LR=1,NRAD
      IORD=IORD+1
30    STOR(LR,JL)=W(IORD)
C
C  IF NECESSARY WRITE THE INFORMATION ON THE SCRATCH FILE
C
      IF(NLOOP.GT.0) WRITE(12) (W(KLO),KLO=1,KM1)
10    CONTINUE
      GO TO 12
C
C  READ THE SCRATCH FILE AND COPY THE INFORMATION FOR THE SCORE GROUPS
C  BEING TREATED IN THIS LOOP INTO THE STOR ARRAY
C
11    REWIND 12
      DO 110 I=2,K
      L=I-1
      DO 110 J=1,L
      JL=JL+1
      READ(12)  (W(KLO),KLO=1,KM1)
      IORD=NTRAD*(LOOP-1)
      DO 130 LR=1,NRAD
      IORD=IORD+1
130   STOR(LR,JL)=W(IORD)
110   CONTINUE
12    IORD=NTRAD*(LOOP-1)
C
C  LOOP OVER THE SCORE GROUPS BEING TREATED IN THIS LOOP
C
      DO 40 LR=1,NRAD
      IORD=IORD+1
      IF(NR(IORD))39,39,41
39    KS=KS-1
      GO TO 40
C
C  COMPUTE THE VARIANCE- COVARIANCE MATRICES
C
41    JL=0
      JK=0
      DO 50 J=1,K
      DO 50 L=1,J
      JL=JL+1
      IF(L.EQ.J)GO TO 141
      JK=JK+1
      VARKOV(JL)=NR(IORD)*EPS(J)*EPS(L)*STOR(LR,JK)/G(IORD)
      GO TO 50
141   VARKOV(JL)=EPS(J)*G1(J,IORD)*NR(IORD)/G(IORD)
C
C  COMPUTE THE DIFFERENCES BETWEEN OBSERVED AND PREDICTED FREQUENCIES
C
      VEK(J)=NIS(IORD,J)-VARKOV(JL)
50    CONTINUE
C
C  INVERT THE VARIANCE- COVARIANCE MATRICES
C
      CALL DSINV(VARKOV,K,TOL,IER)
      IF(IER.LT.0)GO TO 960
      IF(IER.GT.0)GO TO 970
```

```
C
C  PRE- AND POSTMULTIPLY THE INVERTED MATRICES WITH THE VECTOR OF
C  DIFFERENCES
C
      CHI2=0.00
      IJ=0
      DO 60 I=1,K
      DO 60 IO=1,I
      IJ=IJ+1
      IF(I.EQ.IO)GO TO 61
      CHI2=CHI2+2.*VEK(I)*VEK(IO)*VARKOV(IJ)
      GO TO 60
   61 CHI2=CHI2+VARKOV(IJ)*VEK(I)**2
   60 CONTINUE
      CHI=CHI+CHI2
      WRITE(6,1003)IORD,NR(IORD),CHI2
   40 CONTINUE
      RETURN
  960 WRITE(6,1001) IORD
      IER2=1
      RETURN
  970 WRITE(6,1002)IORD
      IER2=2
      RETURN
 1001 FORMAT(*0THE MARTIN-LOF TEST CANNOT BE COMPUTED SINCE THE VARIANCE
     1-COVARIANCE MATRIX FOR SCOREGROUP*,I3,* IS NOT POSITIVE DEFINITE*)
 1002 FORMAT(*0THE MARTIN-LOF TEST CANNOT BE COMPUTED SINCE TOO LOW COMP
     1UTATIONAL ACCURACY WAS OBTAINED*/* IN THE INVERSION OF THE VARIANC
     2E-COVARIANCE MATRIX FOR SCORE GROUP*,I3)
 1003 FORMAT(*0*,I6,I20,F25.3)
      END



      SUBROUTINE EMACHI(METHOD,IPREC,IEX,EPS,EPSI,G,GI,MAXI,
     1ERROR,AOG,NAG,NIS,K,FLIKE,WK3,NORM,NR,KODI,KODL,MINST)
C
C  THE ROUTINE COMPUTES THE LIKELIHOOD RATIO GOODNESS OF FIT TEST
C  AS SUGGESTED BY E.H. ANDERSEN (1973)
C
      REAL*8 EPS(1),EPSI(1),G(1),GI(K,K),WK3(3,K),FLIKE,FLOG,FLOT
      INTEGER NAG(1),AOG(1),NIS(K,K),NR(1),KODI(1),KODL(1)
      INTEGER*2 KDIFF(60)
      FLOG=FLIKE
      FLOT=FLIKE
      IORD=0
      LORD=K
      NGRI=0
      NGRL=0
C
C  SEEK FIRST A SUITABLE GROUPING OF THE SCORE GROUPS
C  GO UPWARDS FIRST
C
      DO 10 I=1,K
   10 AOG(I)=0
      NRT=0
    2 IORD=IORD+1
C
C  TEST IF THERE ARE ANY SCORE GROUPS LEFT
C
      IF(IORD-(LORD-1))3,3,199
C
C  ADD A SCORE GROUP
C
    3 DO 20 I=1,K
   20 AOG(I)=AOG(I)+NIS(IORD,I)
      NRT=NRT+NR(IORD)
C
C  TEST FOR ZERO AND FULL ITEM PROPORTIONS
C
      DO 30 I=1,K
      IF(AOG(I))2,2,24
   24 IF(AOG(I)-NRT)30,2,2
   30 CONTINUE
C  TEST IF THE NUMBER OF SUBJECTS IS SUFFICIENT
C
      IF(NRT-MINST)2,23,23
C  A GROUPING IS ACCEPTED
C
   23 NGRI=NGRI+1
      KODI(NGRI)=IORD
C
C  REPEAT THE PROCESS GOING DOWNWARDS
      DO 40 I=1,K
   40 AOG(I)=0
      NRT=0
  102 LORD=LORD-1
      IF((IORD+1)-LORD)103,103,199
  103 DO 50 I=1,K
   50 AOG(I)=AOG(I)+NIS(LORD,I)
      NRT=NRT+NR(LORD)
C  TEST FOR ZERO AND FULL
      DO 60 I=1,K
      IF(AOG(I))102,102,129
  129 IF(AOG(I)-NRT)60,102,102
   60 CONTINUE
      IF(NRT-MINST)102,123,123
  123 NGRL=NGRL+1
      KODL(NGRL)=LORD
      IF((IORD+1)-LORD)1,204,204
C
C  THE FOLLOWING STATEMENT IS REACHED IF THE NUMBER OF SCORE GROUPS IS
C  EXHAUSTED DURING A GROUPING
  199 CONTINUE
```

```
C       IF(NGRI.LT.1.OR.NGRL.LT.1)GO TO 999
C
C  ADD THE REMAINING SCORE GROUPS TO ADJOINING GROUPS FOUND
C
200     IORD=KOUI(NGRI)
        LORD=KOUL(NGRL)
        IF(IORD+1.EQ.LORD)GO TO 204
201     IORD=IORD+1
        IF((IORD+1)-LORD)202,203,203
202     LORD=LORD-1
        IF((IORD+1)-LORD)201,203,203
203     KOUI(NGRI)=IORD
        KOUL(NGRL)=LORD
204     CONTINUE
        LIGH=0
        NGRT=NGRI+NGRL
        WRITE(6,1006)MINST
        WRITE(6,1001)
C
C  ESTIMATE THE ITEM PARAMETERS WITHIN EACH OF THE GROUPS
C
        DO 300 I=1,NGRI
        LOW=LIGH+1
        LIGH=KOUI(I)
309     CONTINUE
        DO 310 J=1,K
        AOG(J)=0
310     NRG(J)=0
        NRT=0
        DO 320 J=LOW,LIGH
        NRG(J)=NR(J)
        NRT=NRT+NR(J)
        DO 320 L=1,K
320     AOG(L)=AOG(L)+NIS(J,L)
        WRITE(6,1002)LOW,LIGH,NRT
        CALL PAREST(METHOD,IPREC,IFX,EPS,EPSI,G,GI,MAXI,ERROR,AOG,NRG,
       1NIS,K,FLIKE,ITER,KDIFF,WK3,NORM)
        FLOG=FLOG-FLIKE
300     CONTINUE
        DO 400 I=1,NGRL
        LOW=KOUL(NGRL-I+1)
        IF(I-NGRL)402,401,401
  402   LIGH=KOUL(NGRL-I)-1
        GO TO 403
401     LIGH=K-1
403     DO 410 J=1,K
        AOG(J)=0
410     NRG(J)=0
        NRT=0
        DO 420 J=LOW,LIGH
        NRG(J)=NR(J)
        NRT=NRT+NR(J)
        DO 420 L=1,K
420     AOG(L)=AOG(L)+NIS(J,L)
        WRITE(6,1002)LOW,LIGH,NRT
        CALL PAREST(METHOD,IPREC,IFX,EPS,EPSI,G,GI,MAXI,ERROR,AOG,NRG,
       1NIS,K,FLIKE,ITER,KDIFF,WK3,NORM)
        FLOG=FLOG-FLIKE
400     CONTINUE
        FLOT=FLOG/FLOT
        FLOG=-2*FLOG
        NDF=(NGRT-1)*(K-1)
        CALL MODIFI(FLOG,NDF,FLIKE,IER)
        WRITE(6,1003)FLOG,NDF,FLIKE
        WRITE(6,1007)FLOT
        RETURN
999     WRITE(6,1005)
        RETURN
1001    FORMAT(*0THE FOLLWING GROUPING HAS BEEN USED
       1      *///*     RANGE     NUMBER OF OBSERVATIONS*)
1002    FORMAT(*0*,I3,* - *,I3,I14)
1003    FORMAT(////* THE LIKELIHOOD RATIO GOODNESS OF FIT TEST GIVES CHI-S
       1QUARE =*,F14.3,* WITH*,I6,* DEGREES OF FREEDOM. P=*,F7.5,*.*)
1005    FORMAT(*0 THE LIKELIHOOD RATIO TEST CANNOT BE COMPUTED*)
1006    FORMAT(/////* THE MINIMUM NUMBER OF OBSERVATIONS WITHIN EACH*,
       1* GROUP ALLOWED WHEN COMPUTING THE LIKELIHOOD RATIO TEST IS:*,I4)
1007    FORMAT(*0THE REDUNDANCY IS:*,F20.7)
        END



        SUBROUTINE XLIST(LIST,INP,IOUT,J)
C
C       WRITTEN BY JAN-GUNNAR LINGSELL, DEPARTMENT FOR EDUCATIONAL
C       RESEARCH, UNIVERSITY OF GOTEBORG.
C
C       ----------------
C
C       READS THE VARIABLE LIST AND RETURNS THE VARIABLE NUMBERS IN LIST
C
C       EX:
C           001,005,007-015,020*
C
C       ALWAYS 3 DIGITS IN THE NUMBER, BEGIN IN POS 1.
C       FIRST CARD MUST BE COMPLETE BEFORE STARTING ON NEXT CARD
C       SEPARATE WITH COMMA OR HYPHEN
C       AFTER LAST VARIABLE PUNCH AN ASTERISK (*).
C-----------------------------------------------------------------------
C
        INTEGER*2 LIST(1)
        DIMENSION IV(20),II(20),IECK(4),CARD(20)
        DATA IECK/'-',',','*',' '/
C
        J=1
        I=1
```

```
1     READ(INP,100,END=99) CARD,IV,T
100   FORMAT(20A4,T1,20(I3,1X),T1,20(3X,A1))
101   FORMAT('0',20A4)
      WRITE(IOUT,101) CARD
      DO 15 I=1,20
      IF(T(I).EQ.TECK(1)) GOTO 5
      LIST(J)=IV(I)
      J=J+1
      IF(T(I).EQ.TECK(3)) GOTO 20
      IF(T(I).EQ.TECK(4)) GOTO 99
      GOTO 10
5     CONTINUE
      IP1=I+1
      I1=IV(I)
      I2=IV(IP1)
      DO 6 IX=I1,I2
      LIST(J)=IX
6     J=J+1
      Y=Y+1
      IF(T(I).EQ.TECK(3)) GOTO 20
      IF(T(I).EQ.TECK(4)) GOTO 99
10    I=I+1
15    CONTINUE
      GOTO 1
20    J=J-1
      RETURN
99    WRITE(IOUT,999)
999   FORMAT('0ERROR IN VARIABLE LIST.**********')
      STOP 77
      END



      SUBROUTINE GETKEY(IVAL,INP,IOUT,NKEY,*)
C
C     WRITTEN BY JAN-GUNNAR TINGSELL, DEPARTMENT FOR EDUCATIONAL
C     RESEARCH, UNIVERSITY OF GOTEBORG.
C
C     -----------------
C
C
C     PARAMETERS:
C     1:ST PARAMETER CARD INCLUDES KEYWORDS.
C     THEY ARE WRITTEN WITH 4 CHARACTERS SEPARATED BY COMMA OR BLANK
C     BEGINNING IN FIRST POSITION.
C     ONLY PARAMETERS OVERRIDING THE DEFAULTS ARE REQUIRED.
C     IF ALL DEFAULTS ARE CHOOSEN, THE CARD MUST STILL BE THERE
C
C     2:ND PARAMETER CARD INCLUDES THE PARAMETERS MAXI, ERRO,
C     NIND, NOBS AND NORM IN THIS ORDER WITH FIXED FORMAT 4 POS.
C     THIS CARD IS ALSO ALWAYS REQUIRED.
C
C     ***************************************************************
C
      DIMENSION IVAL(1)
      DIMENSION LISTA(30),LISTB(30),LISTC(30),LISTC(15)
      DATA LISTA/'NOCL','DICT','DIFF','SUMM','DOUM','SING','NOSP',
     1  'CHIS','LINE','HOTH','NOMI','MINO','NOPL','PLOT','APPR',
     2  'UNIT','FATR','NOEX','PERS','NOPE','NODE','DESC'/
      DATA LISTB/1,0,0,1,0,1,0,1,2,3,1,0,1,0,0,1,1,0,0,1,1,0/
      DATA LISTC/1,1,4,4,5,5,6,6,6,6,7,7,8,8,9,9,10,10,
     1   13,13,15,15/
      N1=22
C     N1 = NUMBER OF ELEMENTS IN LISTA
C
C
C     DEFAULT-VALUES
      IVAL(1)=1
      IVAL(2)=250
      IVAL(3)=3
      IVAL(4)=0
      IVAL(5)=0
      IVAL(6)=0
      IVAL(7)=1
      IVAL(8)=1
      IVAL(9)=0
      IVAL(10)=1
      IVAL(11)=100
      IVAL(12)=5
      IVAL(13)=0
      IVAL(14)=0
      IVAL(15)=1
C
C
      READ(INP,101) LISTC
      WRITE(IOUT,103)
      DO 11 I=1,15
      K=0
      DO 10 J=1,N1
      IF(LISTC(I).EQ.LISTA(J)) K=J
10    CONTINUE
      IF(K.EQ.0) GOTO 11
      M=LISTB(K)
      L=LISTB(K)
      IVAL(M)=L
      WRITE(IOUT,104) LISTA(K)
11    CONTINUE
      READ(INP,102) LISTC
      IF(LISTC(1).NE.0) IVAL(2)=LISTC(1)
      IF(LISTC(2).NE.0) IVAL(3)=LISTC(2)
      IF(LISTC(3).NE.0) IVAL(11)=LISTC(3)
      IF(LISTC(4).NE.0) IVAL(12)=LISTC(4)
      IF(LISTC(5).NE.0) IVAL(14)=LISTC(5)
C
```

```
      IF(LISTC(1).NE.0) WRITE(IOUT,201) LISTC(1)
      IF(LISTC(2).NE.0) WRITE(IOUT,202) LISTC(2)
      IF(LISTC(3).NE.0) WRITE(IOUT,203) LISTC(3)
      IF(LISTC(4).NE.0) WRITE(IOUT,204) LISTC(4)
      IF(LISTC(5).NE.0) WRITE(IOUT,205) LISTC(5)
      RETURN
101   FORMAT(15(A4,1X))
102   FORMAT(I5I4)
103   FORMAT('0FOLLOWING PARAMETERS OVERRIDES THE DEFAULTS:'/)
104   FORMAT(5X,A4)
201   FORMAT('0',4X,'MAXI=',I4)
202   FORMAT(' ',4X,'ERRO=',I4)
203   FORMAT(' ',4X,'MIND=',I4)
204   FORMAT(' ',4X,'NOBS=',I4)
205   FORMAT(' ',4X,'NOMS=',I4)
      END



      SUBROUTINE GETDIC(LISDIC,NV,INP,IOUT,LIST)
C     WRITTEN BY JAN-GUNNAR TINGSELL, DEPARTMENT FOR EDUCATIONAL
C     RESEARCH, UNIVERSITY OF GOTEBORG.
C ******************************************
C     READS DICTIONARY AND DATA FILES.
C
C     THE ITEMS MUST BE DESCRIBED IN A DICTIONARY FILE ON UNIT 13
C     DATA FILE UNIT 14.
C     DICTIONARY DESCRIPTIONS:
C        POS 1-3     VARIABLE NUMBER
C        POS 4-27    VARIABLE NAME
C        POS 28-30   COLUMN LOCATION IN DATAFILE
C
C     EACH CASE MUST BE INCLUDED IN ONE RECORD
C
C     FIRST RECORD IN DICTIONARY FILE CONTAINS:
C        LRECL  FOR DATA FILE
C        ISTART  STARTING VARIABLE NUMBER IN DICTIONARY
C        IEND    ENDING VARIABLE NUMBER
C        FORMAT(3I3)
C
C     MAX 200 VARIABLES CAN BE DEFINED IN DICTIONARY
C     THE VARIABLES MUST BE CONTINOUSLY NUMBERED BETWEEN ISTART
C     AND IEND.
C ******************************************
C
      INTEGER*2 LIST(1),LOC(200),NR(200)
      INTEGER NAME(200,6),NAMN(6),V(1),X(1000),DATFIL
      INFIL=13
      DATFIL=14
C
C     READ DICTIONARY FILE
C
      READ(INFIL,100) LRECL,ISTART,IEND
      I=1
10    READ(INFIL,101,END=15) NR(I),(NAME(I,J),J=1,6),LOC(I)
      I=I+1
      IF(I.GT.200) STOP 15
      GOTO 10
15    CONTINUE
C
C     CHECK VARIABLE LIST AND PRINT DICTIONRY
C
      IF(LISDIC.EQ.0) WRITE(IOUT,102)
      DO 20 I=1,NV
      J=LIST(I)
      IF(J.LT.ISTART.OR.J.GT.IEND) GOTO 90
      II=LIST(I)-ISTART+1
      DO 20 I2=ISTART,IEND
      IF(I2.EQ.J.AND.LISDIC.EQ.0) WRITE(IOUT,103) NR(II),
     *    (NAME(II,I3),I3=1,6),LOC(II)
20    CONTINUE
      RETURN
90    WRITE(IOUT,104) J
      STOP 16
      ENTRY CASE(V,KL)
      KL=0
C     READ ONE CASE AND RETURN THE VARIABLES IN V.
C
21    READ(DATFIL,105,END=30) (X(I),I=1,LRECL)
      DO 25 I=1,NV
      JS=LIST(I)-ISTART+1
      LOC=LOC(JS)
25    V(I)=X(LOC)
      RETURN
30    KL=1
C     END OF FILE
      RETURN
      ENTRY GNAME(I,NAMN)
C
C     GET THE VARIABLE NAMES
C
      II=LIST(I)-ISTART+1
      DO 50 J=1,6
50    NAMN(J)=NAME(II,J)
      RETURN
100   FORMAT(10I3)
101   FORMAT(I3,6A4,I3)
102   FORMAT('1THE FOLLOWING VARIABLES ARE INCLUDED:'/'0',
     1    'VAR   NAME',22X,'LOC'/)
103   FORMAT(' ',I3,3X,6A4,2X,I3)
104   FORMAT('0VARIABLE:',I4,' NOT IN DICTIONARY')
105   FORMAT(250I1,250I1,250I1,250I1)
      END
```

```
      SUBROUTINE PLOTT(X,Y,N,M,INC,IA,A,IOUT)
C
C     WRITTEN BY JAN-GUNNAR TINGSELL, DEPARTMENT FOR EDUCATIONAL
C     RESEARCH, UNIVERSITY OF GOTEBORG.
C
      DIMENSION X(1),Y(IA,1),A(50),T(4),IY(2)
      LOGICAL*1 MTR(101,51),MX(5151),BLANK(4),TECK(12)
      EQUIVALENCE (MTR(1,1),MX(1)),(BLANK(1),BL),
     1    (T(1),TECK(1))
      DATA BL/'  '/,T/'IX-','1678','90 M'/
C
C
      IF(M.NE.2) GOTO 98
C
CCC   BLANK OUT THE PLOT
C
      DO 10 I=1,5151
10    MX(I)=BLANK(1)
CCC
CCC   X- AND Y-VALUES ALWAYS >=0.0 AND <=1.0
CCC   JUST 2 FUNCTIONS CAN BE PLOTTED
CCCCC
CCC   CONSTRUCT THE FRAMES
C
      DO 11 I=1,101
      MTR(I,1)=TECK(3)
      IF(((I-1)/5)*5.EQ.(I-1))  MTR(I,1)=TECK(4)
11    MTR(I,51)=MTR(I,1)
C
      DO 12 J=2,50
      MTR(1,J)=TECK(5)
      IF(((J-1)/5)*5.EQ.(J-1))  MTR(1,J)=TECK(4)
12    MTR(101,J)=MTR(1,J)
CC
C
      DO 25 I=1,N
      IF(X(I).LT.0.0.OR.X(I).GT.1.0) GOTO 99
      IX=100.0*A(I)+1.5
      DO 21 IM=1,M
      IF(Y(I,IM).LT.0.0.OR.Y(I,IM).GT.1.0) GOTO 99
      IY(IM)=100.0*(Y(I,IM)/2)+1.5
21    CONTINUE
      IF(IY(1).EQ.IY(2)) GOTO 23
      DO 22 IM=1,M
      IIY=IY(IM)
22    MTR(IX,IIY)=TECK(IM)
      GOTO 24
23    IIY=IY(1)
      MTR(IX,IIY)=TECK(12)
24    CONTINUE
25    CONTINUE
CCC
CCC   PRINTOUTS
C
      WRITE(IOUT,106)  (A(I),I=31,40)
      DO 30 JJ=1,50
      J=51-JJ
      JP1=J+1
      XJ=2.0*J/100.0
      IF((J/5)*5.NE.J) WRITE(IOUT,101) XJ,(MTR(I,JP1),I=1,101)
      IF((J/5)*5.EQ.J) WRITE(IOUT,102) (MTR(I,JP1),I=1,101)
30    CONTINUE
      WRITE(IOUT,103)  XJ,(MTR(I,1),I=1,101)
      WRITE(IOUT,104) (I,I=1,9)
      WRITE(IOUT,105) (A(I),I=21,30)
      RETURN
98    WRITE(IOUT,201)
      RETURN
99    WRITE(IOUT,202)
      RETURN
100   FORMAT('1',//' Y-AXIS:  ',10A4)
101   FORMAT(' ',F10.1,1X,101A1)
102   FORMAT(' ',11X,101A1)
104   FORMAT(' ',10X,'0.0',9(7X,'0.',I1),7X,'1.0')
105   FORMAT('0',11X,'X-AXIS:    ',10A4)
201   FORMAT('0***  ERROR IN PLOTT ***     NUMBER OF FUNCTIONS'
     *  ' NOT EQUAL 2.')
202   FORMAT('0***  ERROR IN PLOTT ***     X- OR Y-VALUES NOT IN '
     *  'RANGE  0 TO 1.')
      END
```

REFERENCES

Allerup, P., Mylov, P. & Spelling, S. (1977) Developmental curves through item analysis. Copenhagen, The Danish Institute for Educational Research, 1977.41.

Allerup, P. & Sorber, G. (1977) The Rasch model for questionnaires. With a computer program (2nd ed). Copenhagen, The Danish Institute for Educational Research, 1977.4.

Andersen, E.B. (1972) The solution of a set of conditional estimation equations. Journal of Royal Statistical Society. 34, 42-54.

Andersen, E.B. (1973a) Conditional inference and models for measuring. Copenhagen: Mentalhygienisk Forlag.

Andersen, E.B. (1973b) A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Andrich, D. (1977) A general psychometric model for ordered categories scored with successive integers. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977.

Andrich, D. & Douglas, G.A. (977) Reliability: Distinctions between item consistency and subject separation with the simple logistic model. Paper presented at the Annual Meeting of the American Educational Association, New York, April 1977.

Baker, F.B. (1977) Advances in item analysis. Review of Educational Research, 47, 151-178

Bock, R.D. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock, R.D., Lieberman, M. (1970) Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Brogden, H.E. (1946) Variation in test validity with variation in the distribution of item difficulties, number of items and degree of their intercorrelation. Psychometrika, 11, 197-214.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972) The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L.J., & Warrington, W.G. (1952) Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 17, 129-147.

Dahlquist, G., & Björck, Å. (1974) Numerical methods. Englewood Cliffs: Prentice-hall.

Douglas, G.A. (1975) Test design strategies for the Rasch psychometric model. Unpublished doctoral dissertation, University of Chicago.

Ferguson, G.A. (1941) The factorial interpretation of test difficulty. Psychometrika, 6, 323-329.

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

Fischer, G.H., & Allerup, P. (1968) Rechentechnische Fragen zu Raschs eindimensionalen Modell. In G.H. Fischer (Ed) Psychologische Testtheorie. Bern: Huber.

Fisher, C.H., & Scheiblechner, H. (1970) Algorithmen und Programme für das probabilistische Testmodell von Rasch. Psychologische Beitrage, 12, 23-51.

Gulliksen, H.O. (1945) The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 10, 79-91.

Gulliksen, H.O. (1950) Theory of mental tests. New York: Wiley.

Gustafsson, J.-E. (1976) Verbal and figural aptitudes in relation to instructional methods. Studies in aptitude-treatment interactions. Göteborg: Acta Universitatis Gothoburgensis.

Hambleton, R.K. et al. (1977) Developments in latent trait theory: A review of models, technical issues, and applications. Paper presented at a joint meeting of NCME and AERA in New York, April 1977.

Kempf, W.F. (1976) Notwendige und hinreichende Bedingungen für ein allgemeines dynamisches Testmodell. In Kempf, Niehausen & Mach, (1976), pp 52-62.

Kempf, W.F., & Niehausen, B. (1976) Algorithmen und Programme für ein logistiches Testmodell mit additiven Nebenbedingungen bezüglich der Itemparameter. In Kempf, Niehausen & Mach, (1976), pp 16-51.

Kempf, W.F., Niehausen, B., & Mach, G. (Eds.) (1976) Logistiche Testmodelle mit additiven Nebenbedingungen. Institut für die Pädagogik der Naturwissenschaften an der Christian-Albrechts-Universität Kiel, No 22.

Kifer, E. (1976) Estimating scores on a common metric using the Rasch model: An IEA application. Paper presented at the Annual Meeting of the American Educational Research Association, San Fransisco, 1976.

Kifer, E.W., Mattsson, I., & Carlid, M. (1975) Item analysis using the Rasch model. An application on IEA data. Reports from the Institute for the Study of International Problems in Education, University of Stockholm, no 12.

Kilborn, W., & Johansson, B. (1976) Elevernas räknefärdigheter (The pupils' arithmetic skills). Pedagogiska institutionen, Göteborgs universitet, PUMP-projektet, 11.

Lawley, D.N. (1943) On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61-A, 273-287.

Lazarsfeld, P.F. (1950) The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer et al. Measurement and Prediction. Princeton: Princeton University Press.

Lazarsfeld, P.F., & Henry, N.W. (1968) Latent structure analysis. New York: Houghton Mifflin.

Lewis, T.G., & Payne, W.H. (1973) Generalized feedback shift register pseudo random number algorithm. Journal of the Association for Computing Machinery, 20, 456-468.

Loevinger, J. (1954) The attenuation paradox in test theory. Psychological Bulletin, 51, 493-504.

Lord, F.M. (1952) A theory of test scores. Psychometric Monographs, No. 7.

Lord, F.M. (1953) The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13, 517-548.

Lord, F.M. (1971) Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 31, 3-31.

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

Lord, F.M. (1974b) Individualized testing and item characteristic curve theory. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (eds). Contemporary developments in mathematical psychology, Vol II. San Fransisco: Freeman.

Lord, F.M., & Novick, M.R. (1968) Statistical theories of mental test scores. Reading: Addison-Wesley.

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

Lumsden, J (1976) Test theory. In M.R. Rosenzweig & L.W. Porter (Eds) Annual Review of Psychology, Volume 27. Palo Alto: Annual Reviews Inc.

Lybeck, L. (1974) En mätteori för naturvetenskaplig undervisning (A theory of measurement for science instruction). Rapporter från Pedagogiska institutionen, Göteborgs universitet, nr. 110.

Martin-Löf, P. (1973) Statistiska modeller. Anteckningar från seminarier läsåret 1969-70 utarbetade av Rolf Sundberg. 2:a uppl. (Statistical models. Notes from seminars 1969-70 by Rolf Sundberg. 2nd ed.) Institutet för försäkringsmatematik och matematisk statistik vid Stockholms universitet.

Martin-Löf, P. (1974a) The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. Scandinavian Journal of Statistics, 1, 3-18.

Martin-Löf, P. (1974b) Exact tests, confidence, regions and estimates. Proceedings of conference on foundamental questions in statistical inference. Memoirs, no. 1, Department of Theoretical Statistics, Institute of Mathematics, University of Aaarhus.

McNemar, Q. (1946) Opinion-Attitude methodology. Psychological Bulletin, 43, 289-374.

Mead, R. (1976a) Assessing the fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Fransisco, 1976.

Mead, R. (1976b) Assessment of fit of data to the Rasch model through analysis of residuals. Unpublished doctoral dissertation, University of Chicago.

Neyman, J., & Scott, E.L. (1948) Consistent estimates based on partially consistent observations. Econometrika, 16, 1-5.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute for Educational Research.

Rasch, G. (1961) On general laws and the meaning of measurement in psychology. Berkeley symposium on mathematical statistics and probability. Berkeley: University of California Press.

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

Rentz, R.R., & Bashaw, W.L. (1975) Equating reading tests with the Rasch model. Athens, Georgia: Educational Resources Laboratory.

Spada, H. (1976) Modelle des Denkens und Lernens. Bern: Huber.

Svensson, A. (1964) Sociala och regionala faktorers samband med över- och underprestation i skolarbetet (The relation of social and regional factors to over- and underachievement) Rapporter från Pedagogiska institutionen, Göteborgs universitet, nr 13.

Svensson, A. (1971) Relative achievement. School performance in relation to intelligence, sex and home environment. Stockholm: Almqvist & Wiksell.

Wernersson, I (1977) Könsdifferentiering i grundskolan. (Sex differentiation in school) Göteborg: Acta Universitatits Gothoburgensis.

Willmott, A.S., & Fowles, D.E. (1974) The objective interpretation of test performance. The Rasch model applied. NFER Publishing Company Ltd.

Wright, B.D. (1977) Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, No 2.

Wright, B.D., & Douglas, G.A. (1975) Best test design and self-tailored testing. Research Memorandum No 19, Statistical Laboratory, Department of Education, University of Chicago.

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, University of Chicago.

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

46. Gustafsson,Jan-Eric: Inconsistencies in aptitude-treatment interactions as a function of procedures in the studies and methods of analysis. Project MID 16, March 1976.

47. Gustafsson,Jan-Eric: Differential effects of imagery instructions on pupils with different abilities. Project MID 17, April 1976.

48. Ekholm, Mats: Social development in school. Summary and excerpts. Project SOS 23, May 1976.

49. Stangvik, Gunnar: Approaches to the analysis of learner-task interactions and some implications for the study of pedagogical processes. Project YP 7, January 1976.

50. Andrae, Annika: Non-graded instruction in small rural lower secondary schools. A presentation of the PANG-project. Paper read at the INTERSKOLA conference, July 1976, Project PANG 20, July 1976.

51. Patriksson, Göran: Attitudes toward olympic games of Swedish adolescents. Paper presented at the international congress of physical activity sciences in Quebec City 11-16 July 1976. September 1976.

52. Gustafsson, Jan-Eric: A note on the importance of studying class effects in aptitude-treatment interactions. Project MID 19, September 1976.

53. Gustafsson, Jan-Eric: Spatial ability and the suppression of visualization by reading. Project MID 20, September 1976.

54. Dahlgren,Lars Owe & Marton, Ference: Investigation into the learning and reaching of basic concepts in economics - a research project on higher education. Paper presented at the Congress of the European Association for Research and Development in Higher Education. August 30 - September 3, 1976, Laouvain la Neuve, Belgium, September 1976.

55. Marton, Ference: Skill as an aspect of knowledge. Some implications from research on students conceptions of central phenomena in their subjects. Paper presented at the Second International Conference on Improving University Teaching, July 13-16, 1976. Heidelberg, F.R. Germany, September 1976.

56. Andrae. Annika (Ed.): Non-graded instruction. Research organization and design. Administration and daily teaching experiences in small rural lower secondary schools. Experiences from the PANG-project. Paper read at the INTERSKOLA conference in Sveg, July 1976. Project PANG 23, August 1976.

57. Sandgren, Björn & Åsberg, Rodney: On cognition and social change. A report from a pilot study regarding the effect of schooling on cognitive growth and attitudes towards social change in Pakistan, October 1976.

58. Sandgren, Björn: Relation between cognition and social development. January 1977.

59. Entwistle, Noel: Changing approaches to research into personality and learning. February 1977.

60. Härnqvist, Kjell: Primary Mental Abilities at Collective and Individual Level. October 1977.

61. Härnqvist, Kjell: Enduring Effects of Schooling - A Neglected Area in Educational Research, October 1977.

62. Härnqvist, Kjell: A Note on the Correlations between Increments, Cumulated Attainment and a Predictor. October 1977.

63. Gustafsson, Jan-Eric: The Rasch Model for Dichotomous Items: Theory, Applications and a Computer Program. December 1977.