# STUDIES ON STATISTICAL MODELS FOR POLYTOMOUSLY SCORED TEST ITEMS

ISBN: 90-36512425

Omslagontwerp: Esther Klaassen - van der Kruk

# STUDIES ON STATISTICAL MODELS
# FOR POLYTOMOUSLY SCORED TEST ITEMS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F. A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 17 december 1998 te 15.00 uur

door

Ludovica Maria Wilhelmina Akkermans

geboren op 20 december 1957
te Hasselt, België.

Promotor:
prof. dr. W. J. van der Linden

Assistent-promotor:
dr. C. A. W. Glas

Promotiecommissie:

prof. dr. G. J. Mellenbergh
prof. dr. W. Molenaar
prof. dr. J. C. M. M. Moonen
prof. dr. K. Sijtsma
prof. dr. N. D. Verhelst
dr. H. Kelderman

# Acknowledgements

In writing this dissertation I have experienced many people's support. Wim van der Linden en Kees Glas have patiently and thoroughly commented on many versions of the manuscript. Their constructive criticism has been very valuable to me.

I sincerely thank Wim van der Linden for stimulating me to spend a year abroad. Without the knowledge and confidence I gained in Great Britain this dissertation might not have been written at all. Kees Glas has been important to me too. He inspired me, and I wish to thank him for his confidence.

Eiji Muraki from ETS co-authors Chapter 1. I enjoyed working with him. Alun Thomas from the University of Bath has given stimulating comments on Chapter 5. Norman Verhelst, from OMD and CITO, has been coaching and advising me on several occasions as well. I express my gratitude to all these people for sharing their time and expertise with me.

My thanks also go to the members of the dissertation committee, for their willingness to read and judge the manuscript. Harrie Vorst, from the University of Amsterdam, is acknowledged for providing me with the data used in Chapter 4.

Finally, I thank my colleagues from the department of OMD for the co-operative and pleasant atmosphere that was always present. In particular, I wish to mention Sebie Oosterloo, who used to drop into my office to look at the map of the world hanging on my wall, and we would realize all the places we will probably never get to...

Wageningen, november 1998
Wies Akkermans

# Contents

# Introduction

This dissertation, which is structured as a collection of self-contained papers, will be concerned mainly with differences between item response models. The purpose of item response theory (IRT) is estimation of a hypothesized latent variable, such as, for example, intelligence or ability in geography. The latent variable will be denoted as $\theta$. As latent variables cannot be directly observed, inferences about them have to be made by indirect methods. To this end, the subject is presented with a test, consisting of a number of items. Usually, the subject's response on each test item is scored into an ordinal variable, called the item score, and the investigator has a model which describes the relation that is assumed between $\theta$ and the item scores. This model defines $\theta$. Obviously, the observed responses depend on item characteristics as well as on $\theta$, so these must be part of the model as well. Items may, for example, differ in their difficulty. The models considered in this dissertation will be parametric models.

If the item characteristics were known, one could combine the model and the observed scores to obtain an estimate of $\theta$. Very often, however, the item characteristics are not known and have to be derived from the same responses that are used to estimate $\theta$. The usual procedure is first to estimate the item characteristics, and possibly some parameters of the distribution of $\theta$. These estimates are then used in a second analysis in which an estimate for each subject's value on $\theta$ is obtained. For example, the estimate could be the expectation of the posterior distribution of $\theta$ given the score vector and the estimated item parameters. This two-stage estimation of $\theta$ increases the inaccuracy of the estimate. Relatively little effort is spent, however, in devising procedures to estimate $\theta$ directly from the data.

The present thesis does not focus on the estimation of $\theta$. Its main concern is with differences between item response models, and the possibility to distinguish between these models. However, because the primary purpose of IRT is the estimation of $\theta$, one chapter is included in which a method for direct estimation of $\theta$ is studied. Item parameters in exponential family item response models, such as the Rasch model (Rasch, 1960) or the partial credit model (Masters, 1982), can be estimated directly through conditioning on sufficient statistics for $\theta$. This method of estimation is known as conditional maximum likelihood estimation (CML). Modern commercially available software allows CML estimation of up to 1000 item parameters, but usually there is no explicit option for CML estimation of $\theta$. Verhelst, Glas and van der Sluis (1984) investigated the possibility of using this method for the estimation of $\theta$. They solved the numerical problems, but for large numbers of parameters the procedure remained slow. In 1992, Geyer and Thompson proposed a Markov Chain Monte Carlo method to approximate parameter estimates in exponential family models. In Chapter 5 of this thesis the application of this method to the CML estimation of $\theta$ is studied. The procedure produces fairly accurate results, but it appears to be just as time-consuming as exact calculation of these estimates.

The other chapters all deal with topics that come, in a sense, prior to any estimation. They deal with the question which model to choose (Chapters 2 - 4), or, given a model, which items to choose (Chapter 1), to obtain the best description of the item response process, and the most reliable estimate of $\theta$. With the exception of the second part of Chapter 4, the item response models are studied on a mathematical level, that is, they are considered as a set of known functions of known parameters.

The item information function $\mathrm{I}(\theta)$, which is defined in Chapter 1, can be used to construct an asymptotic confidence interval around $\hat{\theta}$, where $\hat{\theta}$ is an estimate of $\theta$. For accurate estimation of $\theta$ it is desirable to have items with high values on $\mathrm{I}(\theta)$ in the region of $\theta$ which the test is intended to measure. Item information is related to item difficulty: for example, persons of higher ability should get more difficult items than persons of lower ability. A good test has the difficulty of its items in the range of the candidates' abilities. Furthermore, if an item is considered suitable for persons of a certain ability level, it should also discriminate between subjects of this ability who do, and who do not have the knowledge required to pass the test. The item information and discrimination functions can help in the decision to include an item in the test or not.

With binary items, where the response is scored as 'correct' or 'incorrect', the concepts of item difficulty and of item discrimination, although arbitrary, are clearly defined, and the maximum value of these functions is usually easily found. With polytomous items, where the response is scored as $0, 1, \ldots \mathrm{M}$, the concept of item difficulty is not clearly defined, nor is that of item discrimination. Moreover, the information and discrimination functions may have several local maxima. In Chapter 1 these functions are explored for trinary items under the partial credit model (Masters, 1982). The conditions are obtained under which they are uni- or bimodal, and the location and value of the maxima are derived.

The remainder of the dissertation, that is, Chapters 2 through 4, are devoted to an investigation of features of several models for polytomous item responses. In the literature, polytomous items are sometimes distinguished by certain item features, and it is then suggested to describe these items by models that reflect their features. Mellenbergh (1995) argues that it is plausible that the type of model used should be determined by features of the item and by the cognitive processes involved in answering the item. Van Engelenburg (1997) assumes that the process of solving a polytomous item consists of taking a number of dichotomous steps, and he distinguishes task features, which determine the way in which the dichotomous steps are linked up. Task features include the step process, which can be either simultaneous or sequential; the continuation rule (try-all or stop-if-fail), and the ordering mechanism (fixed or not fixed). In this view item response models should reflect the task features. Van Engelenburg's task features are formal features, they are not related to item contents nor to cognitive processes.

In the present thesis this reasoning is carried one step further. It is argued (a) that the interest in IRT is more in scores than in items, (b) that polytomous items should be distinguished by the scoring rule that is applied to the responses, and (c) that models should reflect the scoring rule.

Three different scoring rules are distinguished; there is a close resemblance to Engelenburg's task features, but the emphasis is nevertheless different. These three scoring rules are labelled graded, parallel, and sequential scoring. With graded scoring, the response given by the examinee is evaluated in one overall judgement. With parallel scoring the judge refers to a collection of features, and credit is given for each of these features that is displayed by the response. As an example: in parallel scoring of an essay credit could possibly be obtained for each of the following features: correspondence of the title with the contents, logical flow of the argument, correct use of spelling, grammar and punctuation, the presence of a clear conclusion. If the question were to name three European capitals, then with parallel scoring credit would be given for each correctly mentioned capital. With sequential scoring too, credit is given for each feature in a collection of features that the response displays, but here the search for features is made in a fixed order, and as soon as a feature is not displayed, a score is given and further features are not considered. Sequential scoring may occur in the testing of psycho-motor skills, where an action it tried until the first success, or repeated until the first failure.

The distinction between tasks or items on the one hand and scoring rules on the other hand may at present seem a little contrived, but examples will be encountered of different scoring rules that can be applied to the same item, leading to possibly different scores.

Several families of item response models can be used to describe the distribution of the score vector obtained on a test. In this paper the following families of models are considered: the family of partial credit models (Masters, 1982; Andrich, 1978; Andersen, 1977), the family of graded response models (Samejima, 1969), and the family of sequential models (Tutz 1990, 1997; Verhelst, Glas and de Vries, 1997). The question is whether a particular data set calls for a particular item response model. This question can be approached from a theoretical, a practical, and a statistical point of view:

**A** From the theoretical point of view one can investigate the mathematical differences that exist between the three families. Do the families have features that make them especially useful for certain kinds of data? Is it possible, for example, to connect models to scoring rules? What properties should scoring rules have in order for a certain model to describe the scores?

**B** If it appears that certain models and scoring rules go together well, one could study the question which model to choose from a practical point of view. Practical questions are: how large is the influence of using a 'wrong' model on the ability estimate? Is it possible to distinguish data sets that were generated under different families of models? If the influence of a wrong model on the ability estimate were small, or if it were very difficult to distinguish data sets that were generated under different families, then, from a practical perspective, the question of the choice of a suitable family of models might not be so important after all.

**C** Finally, if certain models and scoring rules go together well, and if the mathematical differences between the families do have practical consequences, one could also approach the problem from a statistical point of view and ask whether it is possible to derive statistical criteria for deciding which family of models can best be used with a given empirical data set.

Chapters 2 and 3 of this thesis address Question A; the material presented in Chapter 4 has to do with Questions B and C.

Concerning Question A, that is, the investigation of model properties, Samejima (1972) already pointed out the possibility of models for sequential and graded processes (see also Samejima, 1997). Molenaar (1983) examined the mathematical relation between on the one hand the Rasch (1960) model for binary items, and on the other hand the partial credit model (PCM), the graded response model (GRM), and the sequential model (SM) for polytomous items. Mellenbergh (1995) took Bock's (1972) nominal response model as a starting point and then distinguished three different order preserving mechanisms leading to three different types of models for ordinal polytomous data. Van Engelenburg (1997, Chapter 2) investigated subtask features to determine the kind of model that suits a particular kind of items. The contributions mentioned above investigate several models or several response features. More specific comparisons, focusing on only one family of item response models, or on only one response feature, have also been made. Jansen and Roskam (1986) concluded that only the GRM is a natural model for rating scales. The SM quite naturally follows from the assumption of a 1 parameter logistic curve for the conditional probability of taking the 'next step' in a sequence of binary tasks (Van Engelenburg, 1997; Mellenbergh, 1995; Molenaar, 1983). Furthermore, Huynh (1994) showed that, under some conditions on the item parameters, the score probabilities under the PCM have the same distribution as the total score on a set of independent binary Rasch items. All these findings seem to point to the conclusion that the three families of models considered in this dissertation could well be used to describe different kinds of data.

In Chapter 2 it will be demonstrated that under *every* model for polytomously scored items, the distribution of the score is equal to the distribution of the total score on a set of binary items that are maximally dependent given their marginals. The response functions of these binary items can be derived from the model for the polytomous item score as well. This does not imply, however, that every model for polytomous items is suitable for modeling binary variables in a sequential design: in Chapter 3 it will be shown that a specification error is made if the PCM or the GRM were applied to binary variables in a sequential design. This holds for the application of the PCM or the GRM to sequentially scored polytomous items as well. Using the SM, no such error is made.

The suitability of the SM for sequentially scored polytomous items should come as no surprise, as the SM was explicitly derived to describe sequential scoring. What is new is that it is demonstrated that the GRM and the PCM are mathematically *un*suited for dealing with sequentially scored items. Combining the two approaches, it can be concluded that of the PCM, the GRM and the SM, only the SM (or a model of identical structure, differing from the SM only in the form assumed for the binary response functions) is suitable for use with sequential scoring.

Regarding Question B, there is literature suggesting that, notwithstanding the theoretical differences between the models, it may be difficult to distinguish them in practice. It is sometimes found that several models fit the same empirical data set reasonably well. Verhelst, Glas and de Vries (1997) found a comparable fit when they applied the SM or the PCM to their data. Furthermore, using an algorithm (De Vries, 1988) to find a set of PCM curves for a set of SM curves, such that the area between the two sets of curves is minimized, Verhelst, Glas and de Vries (1997) concluded that the response curves under these two models can be very close, but that these models are not a reparameterization of each other. Finally, there are simulation studies that seem to indicate that data generated according to the specifications of one of the three models are not always easily recognized as such (Maydeu-Olivares, Drasgow and Mead, 1994; Van Engelenburg, 1997, Chapter 1).

In the first part of Chapter 4 the possibility of distinguishing between data generated under two different families of item response models is further investigated. The research reported on is a modified replication of the simulation study performed by Maydeu-Olivares et al. (1994; see also Levine et al., 1992). These authors defined an ideal observer as an observer making statistically optimal decisions. In their studies, an ideal observer was confronted with two completely specified item response models, and with two score vectors. Each score vector had been simulated under one of the two models. The two models were either members from the same family, or from different families. Careful attention was given to the selection of the models to be actually compared. The observer, whose part was played by a computer, had to match each score vector to the model that generated it. The combination decided on was the one whose likelihood was the larger of the two. This decision rule can be reformulated as a likelihood ratio test. The generation and classification of the two response patterns was repeated a large number of times, for the same model specification, and the percentage of correct classifications was interpreted as an index for the difference between the two models. In Chapter 4 the rate of correct classification in this decision experiment is increased by classifying an entire data matrix instead of only a single score vector. For those models that were selected for the actual investigation, the results indicate that when two members of the same family are compared, the sample size needed for a 95 percent rate of correct classification has to be almost twice as large as when two members of different families are compared. This leads to the conclusion that the consequences of the model differences are large enough to be of practical relevance.

Hence one arrives at question C: the development of statistical criteria for the choice of a family of models. Several indices have been proposed in the literature to aid in the choice of non-nested models. Among those are Akaikes AIC (Akaike, 1973, 1974) and Bozdogan's CAIC (Bozdogan, 1987). In a Bayesian context, methods have been proposed by Schwarz (1978), and recently by Gelfand and Ghosh (1998). See Haughton (1996) for an overview. Gelfand and Ghosh (1998) show that most of these criteria are a combination of a goodness-of-fit measure and a penalty for the number of parameters in the model. When the objective is to choose between families of item response models, however, no such penalty term is required, because the competing models will usually have an equal number of parameters.

The procedure proposed in the second part of Chapter 4 is as follows. For each family of models under consideration, that member is identified that fits the data best. Then, using a zero/one loss function and a maximum likelihood decision rule, the model having the largest likelihood is decided on. The power under the alternatives and the size of the type I error are ascertained by means of simulation. The results of applying the procedure to some simulated data sets seem to indicate that it may be a useful tool in deciding upon an item response model. When the procedure is subsequently applied to an empirical data set, consisting of rating scale data, the GRM is decided on.

The dissertation ends with an epilogue, in which some consequences are discussed of emphasizing, in the question of model choice, scoring rules rather than item types.

# Chapter 1

# Item information and discrimination functions for trinary PCM items[1]

## Abstract

For trinary partial credit items the shape of the item information and the item discrimination function is examined in relation to the item parameters. In particular, it is shown that these functions are unimodal if $\delta_2 - \delta_1 < 4 \ln 2$ and bimodal otherwise. The locations and values of the maxima are derived. Furthermore, it is demonstrated that the value of the maximum is decreasing in $\delta_2 - \delta_1$. Consequently, the maximum of a unimodal item information function is always larger than the maximum of a bimodal one, and similarly for the item discrimination function.

Key words: partial credit model, trinary items, item information function, item discrimination function, maximum item information.

## 1.1 Introduction and Definitions

Let a graded item admit a score $X$ in $0, 1 \ldots$ M. A higher score indicates a better performance. Examinee ability will be denoted by $\theta$. The category response function (CRF) gives the probability of obtaining a score $k$, as a function of $\theta$.

[1] This paper has been written in cooperation with dr. E. Muraki from ETS, Princeton. It has been published as: Akkermans, W. and E. Muraki (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62*, 569-578.

CRFs will be denoted by the symbol $\mathrm{P}_k(\theta)$:

$$\mathrm{P}_k(\theta) = \Pr(X = k; \theta), \qquad \text{for } k = 0, 1 \ldots \mathrm{M}.$$

In the partial credit model (PCM; Andrich, 1978; Masters, 1982) the CRFs are given by

$$\mathrm{P}_k(\theta) = \frac{\exp \sum_{p=0}^{k} (\theta \Leftrightarrow \delta_p)}{\sum_{r=0}^{M} \left[\exp \sum_{p=0}^{r} (\theta \Leftrightarrow \delta_p)\right]}, \tag{1.1}$$

with $\sum_{p=0}^{0} (\theta \Leftrightarrow \delta_p) \equiv 0$. The parameters $\delta_p$ are the scale values at which two consecutive CRFs intersect. The model for binary items formulated by Rasch (1960) can be seen as a special case of the PCM: the binary Rasch item is a PCM item with $\mathrm{M} = 1$.

The expected response function (ERF) is defined as the (normalized) expected score as a function of $\theta$:

$$\mathrm{ERF}(\theta) = \frac{1}{\mathrm{M}} \sum_{k=0}^{M} k \mathrm{P}_k(\theta).$$

The derivative of the ERF is known as the item discrimination function; in this paper it will be denoted by the symbol $\mathrm{G}(\theta)$:

$$\mathrm{G}(\theta) = \tfrac{\partial}{\partial \theta} \mathrm{ERF}(\theta).$$

Using the fact that in the PCM the derivatives of the CRFs are given by

$$\tfrac{\partial}{\partial \theta} \mathrm{P}_k(\theta) = \mathrm{P}_k(\theta)[k \Leftrightarrow \textstyle\sum_{r=0}^{M} r \mathrm{P}_r(\theta)], \tag{1.2}$$

in this model the derivative of the ERF is equal to

$$\mathrm{G}(\theta) = \frac{1}{\mathrm{M}} \left\{ \sum_{k=0}^{M} k^2 \mathrm{P}_k(\theta) \Leftrightarrow \left[ \sum_{k=0}^{M} k \mathrm{P}_k(\theta) \right]^2 \right\}.$$

Let $\mathrm{L}(\theta|X)$ be the likelihood function of $\theta$ given the observed response $X$. Again using (1.2), the item information function $\mathrm{I}(\theta) = \mathrm{E}\left[\Leftrightarrow \frac{\partial^2}{\partial \theta^2} \ln \mathrm{L}(\theta|X); \theta\right]$ follows as

$$\begin{aligned} \mathrm{I}(\theta) &= \sum_{k=0}^{M} \left[\tfrac{\partial}{\partial \theta} \mathrm{P}_k(\theta)\right]^2 / \mathrm{P}_k(\theta) \\ &= \sum_{k=0}^{M} k^2 \mathrm{P}_k(\theta) \Leftrightarrow \left[ \sum_{k=0}^{M} k \mathrm{P}_k(\theta) \right]^2 . \end{aligned} \tag{1.3}$$

Note that in the PCM the functions $\mathrm{I}(\theta)$ and $\mathrm{G}(\theta)$ are proportional. Let a trinary item be a graded item with maximum score 2. Huynh (1994) shows that $\delta_2 \Leftrightarrow \delta_1 \geq 2\ln 2$ is a necessary and sufficient condition for the likelihood of the score on a trinary partial credit item to be fully equivalent to the likelihood of the total score on a set of 2 independent binary Rasch items. Subsequently, Huynh introduces the term "indecomposable" (Huynh, 1996) to refer to trinary PCM items with $\delta_2 \Leftrightarrow \delta_1 < 2\ln 2$; and he proves that the likelihood of the score on any PCM item is equivalent to the likelihood of the total score on a set of independent binary and indecomposable trinary PCM items. Huynh also uses the term trinary Rasch item for the trinary PCM item.

Binary Rasch items have been thoroughly investigated, their features are well known. For example, the information function of a binary Rasch item is unimodal, its maximum occurs at $\theta = \delta$, and the value of the maximum is equal to 1/4. It is also well known that in the Rasch model the value of the derivative of the ERF, evaluated at $\theta = \delta$, equals 1/4. Much less appears to be known about the characteristics of trinary Rasch (PCM) items. Thus, as the PCM appears to be built up of both binary and trinary Rasch items, there seems to be a need for investigating the trinary Rasch item.

This paper will concentrate on the item information function $\mathrm{I}(\theta)$ and the item discrimination function $\mathrm{G}(\theta)$ of the trinary PCM item. The conditions will be derived under which these functions have either one or two modes, and the locations and values of the maxima will be determined. Because of the proportionality of $\mathrm{I}(\theta)$ and $\mathrm{G}(\theta)$ in this model, the calculations will be carried out for $\mathrm{I}(\theta)$ only.

## 1.2 Condition for Unimodality

Let $\mathrm{Var}(X;\theta)$ be the variance of $X$ as a function of $\theta$. Because $\mathrm{Var}(X;\theta) = \sum_{k=0}^{M} k^2 \mathrm{P}_k(\theta) \Leftrightarrow \left[\sum_{k=0}^{M} k\mathrm{P}_k(\theta)\right]^2$, it appears that for the PCM $\mathrm{Var}(X;\theta)$ is equal to $\mathrm{I}(\theta)$, as given in (1.3). As it is well known that $\mathrm{Var}(X;\theta)$ approaches 0 for

$\theta \rightarrow \pm\infty$, it follows that also

$$\lim_{\theta\rightarrow\infty} \mathrm{I}(\theta) = \lim_{\theta\rightarrow-\infty} \mathrm{I}(\theta) = 0. \tag{1.4}$$

Defining $\overline{\delta} = (\delta_1 + \delta_2)/2$, it is easy to show that for a trinary PCM item,

$$\mathrm{I}(\theta) \text{ is symmetric around } \theta = \overline{\delta}.$$

Starting from (1.3) and once more using (1.2), the derivative of the item information function can be found; it is equal to

$$\tfrac{\partial}{\partial\theta}\mathrm{I}(\theta) = \sum_{k=0}^{M} k^3\mathrm{P}_k(\theta) \Leftrightarrow 3\left[\sum_{k=0}^{M} k\mathrm{P}_k(\theta)\right]\sum_{k=0}^{M} k^2\mathrm{P}_k(\theta) + 2\left[\sum_{k=0}^{M} k\mathrm{P}_k(\theta)\right]^3.$$

If we let $\xi = \exp(\theta)$, $\varepsilon_1 = \exp(\Leftrightarrow\delta_1)$, $\varepsilon_2 = \exp(\Leftrightarrow\delta_2)$ and substitute these into the expressions for $\mathrm{P}_k(\theta)$, then after some algebra and rearranging this derivative can be expressed as

$$\tfrac{\partial}{\partial\theta}\mathrm{I}(\theta) = \xi\varepsilon_1\frac{(1 \Leftrightarrow \xi^2\varepsilon_1\varepsilon_2)[\xi^2\varepsilon_1\varepsilon_2 \Leftrightarrow \xi(\varepsilon_1 \Leftrightarrow 8\varepsilon_2) + 1]}{(1 + \xi\varepsilon_1 + \xi^2\varepsilon_1\varepsilon_2)^3}.$$

This function becomes or approaches zero in each of the following cases:

$$\text{if} \quad \xi \rightarrow \infty, \quad \text{i.e. if} \quad \theta \rightarrow \infty, \tag{1.5}$$

$$\text{if} \quad \xi \rightarrow 0, \quad \text{i.e. if} \quad \theta \rightarrow \Leftrightarrow\infty, \tag{1.6}$$

$$\text{if} \quad \xi^2 = 1/(\varepsilon_1\varepsilon_2), \quad \text{i.e. if} \quad \theta = \overline{\delta}, \tag{1.7}$$

$$\text{if} \quad \xi^2\varepsilon_1\varepsilon_2 \Leftrightarrow \xi(\varepsilon_1 \Leftrightarrow 8\varepsilon_2) + 1 = 0. \tag{1.8}$$

To start with, therefore, there always exist one finite and two asymptotic solutions to $\frac{\partial}{\partial\theta}\mathrm{I}(\theta) = 0$. The number of solutions to $\frac{\partial}{\partial\theta}\mathrm{I}(\theta) = 0$ furthermore depends upon the discriminant $D = (\varepsilon_1 \Leftrightarrow 8\varepsilon_2)^2 \Leftrightarrow 4\varepsilon_1\varepsilon_2$ of the quadratic in (1.8). This quadratic has two real solutions if its discriminant is positive, that is, if

$$\frac{\varepsilon_1}{\varepsilon_2} \leq 4 \quad \text{or} \quad \frac{\varepsilon_1}{\varepsilon_2} \geq 16. \tag{1.9}$$

Let the real solutions to this quadratic, if they exist, be denoted by $\xi_1$ and $\xi_2$. In order for these solutions to be valid they both have to be positive, as

$\xi = \exp(\theta) > 0$. A necessary and sufficient condition for two real numbers to be both positive is that both their sum and their product be positive. The sum of the two roots to the quadratic $px^2 + qx + r$ equals $\Leftrightarrow q/p$, and their product is $r/p$; therefore we need $(\varepsilon_1 \Leftrightarrow 8\varepsilon_2)/\varepsilon_1\varepsilon_2 > 0$ and $1/\varepsilon_1\varepsilon_2 > 0$. The latter condition poses no problem; the former is fulfilled if $\varepsilon_1/\varepsilon_2 > 8$. Hence, if there are two real roots $\xi_1, \xi_2$ to the quadratic in (1.8), these can only both be positive if

$$\frac{\varepsilon_1}{\varepsilon_2} > 8. \tag{1.10}$$

Because of (1.10) only the second possibility in (1.9) is useful. Remembering that $\varepsilon_1/\varepsilon_2 = \exp(\delta_2 \Leftrightarrow \delta_1)$ it may be concluded that the quadratic in (1.8) has

$$\text{no solutions} \quad \text{if} \quad \delta_2 \Leftrightarrow \delta_1 < 4 \ln 2,$$

$$\text{two solutions} \quad \text{if} \quad \delta_2 \Leftrightarrow \delta_1 \geq 4 \ln 2.$$

In the first case, the information function will only have the one finite and two asymptotic extremes derived in (1.5) - (1.7) above, that is, at $\theta = \overline{\delta}$ and for $\theta \rightarrow \pm\infty$; in the second case there are two more extremes in the information function. These two cases will be examined separately below.

## 1.3 Unimodal Item Information Function

Using (1.5) - (1.7), the symmetry of $\mathrm{I}(\theta)$, Equation (1.4), and the fact that the information function is always positive, it may be concluded that if there is only one finite extreme in the information function this function has to be unimodal with a maximum occurring at $\theta = \overline{\delta}$. In order to find the value of this single maximum $\mathrm{I}(\overline{\delta})$ note that $\overline{\delta} \Leftrightarrow \delta_1 = (\delta_2 \Leftrightarrow \delta_1)/2$ and that $\exp(2\overline{\delta} \Leftrightarrow \delta_1 \Leftrightarrow \delta_2) = 1$, so that, using (1.3) and (1.1), it can be verified that

$$\mathrm{I}(\overline{\delta}) = \frac{2}{2 + \exp[(\delta_2 \Leftrightarrow \delta_1)/2]}. \tag{1.11}$$

From (1.11) it follows that

$$\mathrm{I}(\overline{\delta}) \quad \text{is decreasing in} \quad \delta_2 \Leftrightarrow \delta_1. \tag{1.12}$$

The maximum of the information at $\theta = \overline{\delta}$ therefore occurs if $\delta_2 \Leftrightarrow \delta_1 \to \Leftrightarrow \infty$, that is, if $\delta_2 \ll \delta_1$; and the minimum of the information at $\theta = \overline{\delta}$ is reached if $\delta_2 \Leftrightarrow \delta_1 \to \infty$, that is, when $\delta_2 \gg \delta_1$. However, for $\delta_2 \Leftrightarrow \delta_1 \geq 4 \ln 2$ the information function is no longer single peaked: hence the minimum value of the maximum at $\theta = \overline{\delta}$ for a single peaked information function occurs for $\delta_2 \Leftrightarrow \delta_1 = 4 \ln 2$. Using (1.11), values for the maximum at $\theta = \overline{\delta}$ may be easily obtained; some values for the maximum of a unimodal information function are:

$$\mathrm{I}(\overline{\delta}) \approx 1 \quad \text{for} \quad \delta_2 \ll \delta_1, \tag{1.13}$$

$$\mathrm{I}(\overline{\delta}) = \frac{2}{3} \quad \text{for} \quad \delta_1 = \delta_2, \tag{1.14}$$

$$\mathrm{I}(\overline{\delta}) = \frac{1}{2} \quad \text{for} \quad \delta_2 \Leftrightarrow \delta_1 = 2 \ln 2, \tag{1.15}$$

$$\mathrm{I}(\overline{\delta}) = \frac{1}{3} \quad \text{for} \quad \delta_2 \Leftrightarrow \delta_1 = 4 \ln 2. \tag{1.16}$$

From (1.12) and the fact that for $\delta_2 \Leftrightarrow \delta_1 \geq 4 \ln 2$ the item information function is no longer unimodal, it follows that the value of the single maximum at $\theta = \overline{\delta}$ is bounded by 1 and 1/3.

## 1.4 Bimodal Item Information Function

If the quadratic in (1.8) has two real roots, then because of the symmetry of $\mathrm{I}(\theta)$ there must be maxima at these roots and a minimum in between, that is, at $\theta = \overline{\delta}$. Solving the quadratic it follows that the two maxima will occur at $\xi_{1,2} = \left\{\varepsilon_1 \Leftrightarrow 8\varepsilon_2 \pm [(\varepsilon_1 \Leftrightarrow 8\varepsilon_2)^2 \Leftrightarrow 4\varepsilon_1\varepsilon_2]^{1/2}\right\}/2\varepsilon_1\varepsilon_2$. Remembering that $\xi = \exp(\theta)$, $\varepsilon_1 = \exp(\Leftrightarrow \delta_1)$ and $\varepsilon_2 = \exp(\Leftrightarrow \delta_2)$, this can be rewritten as

$$\theta_{1,2} = \overline{\delta} \pm \ln \left\{ \frac{\varepsilon_1 \Leftrightarrow 8\varepsilon_2 + [(\varepsilon_1 \Leftrightarrow 8\varepsilon_2)^2 \Leftrightarrow 4\varepsilon_1\varepsilon_2]^{1/2}}{2(\varepsilon_1\varepsilon_2)^{1/2}} \right\}. \tag{1.17}$$

In the appendix it is shown that the value of these maxima is equal to

$$\mathrm{I}(\theta_1) = \mathrm{I}(\theta_2) = \frac{1}{4\,\{1 \Leftrightarrow 4\exp[\Leftrightarrow(\delta_2 \Leftrightarrow \delta_1)]\}}. \tag{1.18}$$

Again,

$$\mathrm{I}(\theta_1) \quad \text{and} \quad \mathrm{I}(\theta_2) \quad \text{are decreasing in} \quad \delta_2 \Leftrightarrow \delta_1. \tag{1.19}$$

For $\delta_2 \Leftrightarrow \delta_1 = 4 \ln 2$, both maxima still are located at $\theta = \overline{\delta}$, and their value is 1/3. For $\delta_2 \gg \delta_1$, the maxima will be located near $\delta_2$ and $\delta_1$, respectively, as can be seen upon taking the limit of (1.17) for $\delta_2 \Leftrightarrow \delta_1 \to \infty$, that is, for $\varepsilon_2/\varepsilon_1 \to 0$. In this case the exponential $\exp[\Leftrightarrow(\delta_2 \Leftrightarrow \delta_1)]$ will approach 0 and hence

$$\mathrm{I}(\theta_1) \approx \mathrm{I}(\delta_1) \approx \frac{1}{4} \quad \text{for} \quad \delta_2 \gg \delta_1, \tag{1.20}$$

and similarly for the maximum located near $\delta_2$. It may be concluded that the values of the maxima of a bimodal information function are bounded by 1/3 and 1/4.

As an example of a trinary item, let $(\delta_1, \delta_2) = (\Leftrightarrow 2, 2)$. For this item the maxima will be located at $\theta = \pm 1.8154$, and their value is .270. An item with $(\delta_1, \delta_2) = (\Leftrightarrow 3, 3)$ has its maxima at $\theta = \pm 2.9974$, and their value is .252. Note that indeed this value is nearly equal to 1/4.

## 1.5 Discrimination Function

With the appropriate modifications, all the above holds for the item discrimination function $\mathrm{G}(\theta)$ as well. In particular, $\mathrm{G}(\theta)$ will be bi- or unimodal under the same conditions, and the maxima will have the same location. Their values are obtained upon dividing (1.11) and (1.18) by M, which in this case is equal to 2.

If $\mathrm{G}(\theta)$ has one maximum, the second derivative of the expected response function changes sign exactly once: the ERF then has one point of inflection. A function with only one point of inflection is smooth. Hence the ERF of a PCM item with $\delta_2 \Leftrightarrow \delta_1 < 4 \ln 2$ will be smooth. In fact, its smoothness is comparable to the smoothness of a Rasch function. If $\mathrm{G}(\theta)$ is bimodal, there will be a 'bump' in the middle of the corresponding ERF.

## 1.6 Some Comparisons

The maximum information of a single binary Rasch item equals 1/4. Therefore the information obtainable with the total score on 2 independent binary Rasch items can never exceed 1/2. Bearing this in mind, and also the values given in (1.13) - (1.16) and (1.20), some remarks apply:

1. For trinary PCM items, the value of the maximum information is decreasing in $\delta_2 \Leftrightarrow \delta_1$. This follows from the fact that these maxima are, first, decreasing in $\delta_2 \Leftrightarrow \delta_1$ for both unimodal and bimodal information functions (see equations 1.12 and 1.19), and, second, equal for $\delta_2 \Leftrightarrow \delta_1 = 4 \ln 2$.

2. Hence, for trinary PCM items the maximum of every unimodal information function is larger than the maximum of every bimodal information function.

3. By a similar argument, the maximum information of every indecomposable trinary PCM item is larger than the maximum information of every decomposable trinary PCM item.

4. Consequently, the maximum information of every indecomposable trinary PCM item is larger than the maximum information obtainable with 2 independent binary Rasch items. Therefore indecomposable trinary PCM items are in a sense more efficient than the total score on two independent binary Rasch items can ever be.

5. For trinary PCM items with $\delta_2 \geq \delta_1$, the maximum information never exceeds 2/3.

6. If $\delta_2 \gg \delta_1$, the maximum information obtainable with a trinary PCM item is equal to the maximum information for just a single binary Rasch item. PCM items with $\delta_2 \gg \delta_1$ are perhaps best understood upon examining their expected score distributions. Assuming that $\delta_2 \gg \delta_1$ implies both $\overline{\delta} \gg \delta_1$ and $\delta_2 \gg \overline{\delta}$, then for $\theta \ll \overline{\delta}$ the probability of obtaining a score 2 is nearly 0, and for these values of $\theta$ the PCM item behaves as a binary Rasch item with parameter approximately equal to $\delta_1$. For $\theta \gg \overline{\delta}$, the probability of obtaining a score 0 will be nearly 0, and for these values of $\theta$ the PCM

item behaves as a binary Rasch item with parameter approximately equal to $\delta_2$. For 'average' values of $\theta$, which may be quite a substantial part of the $\theta$-axis if $\delta_2$ really is much larger than $\delta_1$, the probability of obtaining a score 1 on the PCM item will be nearly equal to 1. So in the limit, for $\delta_2 \gg \delta_1$, the decomposable trinary Rasch item is equivalent to two independent binary Rasch items with parameters approximately equal to $\delta_1$ and $\delta_2$. Its information function will resemble the information function of the total score on two widely separated Rasch items, with information approximately equal to zero for $\theta$ in the neighborhood of $\overline{\delta}$.

7. The other limit, that is, $\delta_2 \ll \delta_1$, is also interesting. Note that $\mathrm{P}_1(\theta)$ reaches its maximum at $\theta = \overline{\delta}$. Now if $\delta_2 \ll \delta_1$, the maximum $\mathrm{P}_1(\overline{\delta}) \approx 0$, and hence $\mathrm{P}_1(\theta) \approx 0$ for all $\theta$. Consequently, in this case $\mathrm{P}_0(\theta) + \mathrm{P}_2(\theta) \approx 1$ for all values of $\theta$; and because $\mathrm{P}_0(\theta)/\mathrm{P}_2(\theta) = 1/\{1+\exp[2(\theta \Leftrightarrow \overline{\delta})]\}$, it follows that $\mathrm{P}_0(\theta) \approx 1/\{1 + \exp[2(\theta \Leftrightarrow \overline{\delta})]\}$ and $\mathrm{P}_2(\theta) \approx \exp[2(\theta \Leftrightarrow \overline{\delta})]/\{1 + \exp[2(\theta \Leftrightarrow \overline{\delta})]\}$. So for $\delta_2 \ll \delta_1$ the PCM item reduces to a 2 parameter logistic item (Birnbaum, 1968) with location parameter $\overline{\delta}$ and discrimination equal to 2. The maximum information for such an item is equal to 1 (see next section).

## 1.7 Discrimination Parameter

Under the generalized partial credit model (GPCM; see Muraki, 1992; Muraki, 1993), the category response functions are given by

$$\mathrm{P}_k(\theta) = \frac{\exp[\alpha \sum_{p=0}^{k} (\theta \Leftrightarrow \delta_p)]}{\sum_{r=0}^{M} \exp[\alpha \sum_{p=0}^{r} (\theta \Leftrightarrow \delta_p)]},$$

again with $\sum_{p=0}^{0} (\theta \Leftrightarrow \delta_p) \equiv 0$. In this model the discrimination parameter $\alpha$ varies over items. If in this case we let $\xi = \exp(\alpha\theta)$, $\varepsilon_1 = \exp(\Leftrightarrow \alpha\delta_1)$ and $\varepsilon_2 = \exp(\Leftrightarrow \alpha\delta_2)$, all derivations will be analogous to the ones given in the previous sections, resulting in unimodal item information and discrimination functions if $\alpha(\delta_2 \Leftrightarrow \delta_1) < 4 \ln 2$ and bimodal functions otherwise. The values of the item information and discrimination are now equal to $\alpha^2 \mathrm{Var}(X;\theta)$ and $(\alpha/\mathrm{M})\mathrm{Var}(X;\theta)$, respectively. The single maximum is also located at $\theta = \overline{\delta}$; the location of the bimodal maximum can be adapted from (1.17):

$$\theta_{1,2} = \overline{\delta} \pm \frac{1}{\alpha} \ln \left\{ \frac{\varepsilon_1 \Leftrightarrow 8\varepsilon_2 + [(\varepsilon_1 \Leftrightarrow 8\varepsilon_2)^2 \Leftrightarrow 4\varepsilon_1\varepsilon_2]^{1/2}}{2(\varepsilon_1\varepsilon_2)^{1/2}} \right\}.$$

Furthermore, note that the number $4 \ln 2$ is invariant under linear transformation of the $\theta$-scale: setting $\theta^* = k\theta + m$ and offsetting this in the usual way by $\alpha^* = \alpha/k$ and $\delta^* = k\delta + m$ will yield

$$\frac{\varepsilon_1^*}{\varepsilon_2^*} = \frac{\exp(\Leftrightarrow\alpha^*\delta_1^*)}{\exp(\Leftrightarrow\alpha^*\delta_2^*)} = \frac{\exp[\Leftrightarrow\frac{\alpha}{k}(k\delta_1 + m)]}{\exp[\Leftrightarrow\frac{\alpha}{k}(k\delta_2 + m)]} = \frac{\exp(\Leftrightarrow\alpha\delta_1)}{\exp(\Leftrightarrow\alpha\delta_2)} = \frac{\varepsilon_1}{\varepsilon_2}.$$

Hence in any linearly transformed metric $\theta^*$, I$(\theta^*)$ will be unimodal if $\alpha^*(\delta_2^* \Leftrightarrow \delta_1^*) < 4 \ln 2$, and bimodal otherwise.

## 1.8 Conclusion and Discussion

The necessary and sufficient conditions have been stated under which both the item information and the item discrimination function of a trinary PCM item are unimodal, and the location and value of these maxima were derived. It was furthermore ascertained that for trinary PCM items the maximum of a unimodal item information function is always larger than the maximum of a bimodal one; and similarly for the item discrimination function. As a consequence, the ERF of a trinary PCM item with a unimodal discrimination function is both steeper and smoother than the ERF of a trinary PCM item with a bimodal discrimination function. The smoothness of the former is comparable to that of a binary Rasch item. Although the condition for bimodality of the item information function seems rather strong ($\delta_2 \Leftrightarrow \delta_1 > 4 \ln 2 = 2.77$), these items do occur in practice. For example, in the 1994 survey of the National Assessment of Educational Progress (NAEP), students of age 13 were administered 449 items covering Geometry, History and Reading. Of these items, 86 were trinary PCM items, and 11 of them had a bimodal information function. This is about 13 percent. At the time of writing, the analyses of the 1994 NAEP data is still in progress; however, a general overview of the scaling procedures used can be found in Mislevy, Johnson and Muraki (1992).

The results obtained in this paper may have some practical relevance. Matching a target information function is a commonly used criterion in test design. This may require, at some point in the test construction process, finding an item with a pre-specified information at a certain theta level, say at $\theta_0$. For binary Rasch items this is easy: under the Rasch model item information functions differ only in their location, and hence the item located closest to $\theta_0$ will have higher information at $\theta_0$ than all other items. For polytomous items, however, finding the best item is no trivial task. In order to avoid having to calculate $I(\theta_0)$ for many items, knowledge of the location and value of the maxima might be helpful. At least for trinary PCM items these are now available.

Another practical question is that of the optimal value of M. Assume an infinite item pool, and consider a situation in which an item with high information at $\theta = \theta_0$ is required. As an example, one might think of computerized adaptive testing, where a provisional ability estimate $\hat{\theta}^{(t)} = \theta_0$ is available. From an infinite item pool it will be possible to select an indecomposable trinary PCM item, with its maximum information located at $\theta = \theta_0$. As has been demonstrated, the maximum information of an indecomposable trinary PCM item is larger than the maximum information obtainable with any two independent Rasch items. Theoretically, therefore, (infinite item pools do not exist), in this case administration of a suitable, that is, indecomposable trinary item is more efficient than administration of two independent binary Rasch items can ever be. This does not generalize to $M > 2$: the likelihood of every PCM item with $M > 2$ is equivalent to the likelihood of the total score on a number of independent binary and indecomposable trinary PCM items, and therefore the two information functions will be equal. It follows that, even in an infinite item pool, there exist no PCM items with $M > 2$, having larger maximum information at $\theta = \theta_0$, than would be obtainable with the equivalent combination of independent binary and trinary PCM items.

This may have implications for test construction and item banking. It could be argued that the construction of more items with small M (i.e., $M = 1$ and $M = 2$) might be more profitable than the construction of less items, each of these with large M. First, in the infinite item pool, for every PCM item with $M > 2$ there will exist an equivalent combination of independent binary and trinary PCM items having the same information function. Second, from a practical perspective, there will be more items available to choose from. Third, the maximum information for items with M $\leq 2$ can now be calculated and hence these items may become more easily manageable in item banks than items with large M. It should be remarked, however, that not all trinary items are equally efficient: in particular, those with $\delta_2 \gg \delta_1$ have approximately zero information for average values of $\theta$.

The value of the maximum information is decreasing in $\delta_2 \Leftrightarrow \delta_1$. This would seem to suggest that it is desirable to have items with $\delta_2$ as small as possible, compared to the value of $\delta_1$. However, for rating scale items it might be inappropriate to have $\delta_2 < \delta_1$ (Andrich, 1982). Furthermore, although, of all trinary PCM items, the maximum information of items with $\delta_2$ much larger than $\delta_1$ is largest, it was pointed out that these items do not really behave as trinary items: they are equivalent to binary 2 parameter logistic items with $\alpha = 2$.

Finally, in the context of multiple choice items there are several reasons for arguing that the optimal number of choices be 3 (e.g. Lord, 1980). Some of these reasons have to do with discrimination and reliability. It would be interesting to examine whether there is any relationship with 3 seeming the most efficient number of categories for graded items, too.

## Appendix

Calculation of the maxima for a bimodal information function

In this appendix three shorthand symbols will be used:

$$\begin{aligned} \mathrm{f}(\theta) &= \exp(\theta \Leftrightarrow \delta_1), \\ \mathrm{g}(\theta) &= \exp(2\theta \Leftrightarrow \delta_1 \Leftrightarrow \delta_2), \\ y &= \exp(\delta_2 \Leftrightarrow \delta_1). \end{aligned}$$

Note that $\mathrm{g}(\theta)$ can be expressed in $\mathrm{f}(\theta)$ and $y$:

$$\mathrm{g}(\theta) \;=\; [\mathrm{f}(\theta)]^2/y. \tag{1.21}$$

Using (1.3) and the shorthand notations $\mathrm{f}(\theta)$ and $\mathrm{g}(\theta)$, for a trinary PCM item $\mathrm{I}(\theta) = \mathrm{P}_1(\theta) + 4\mathrm{P}_2(\theta) \Leftrightarrow [\mathrm{P}_1(\theta) + 2\mathrm{P}_2(\theta)]^2$ can be written as

$$\begin{aligned}
\mathrm{I}(\theta) &= \mathrm{P}_1(\theta)[1 \Leftrightarrow \mathrm{P}_1(\theta)] + 4\mathrm{P}_2(\theta)[1 \Leftrightarrow \mathrm{P}_2(\theta)] \Leftrightarrow 4\mathrm{P}_1(\theta)\mathrm{P}_2(\theta) \\
&= \frac{\mathrm{f}(\theta)[1+\mathrm{g}(\theta)] + 4\mathrm{g}(\theta)[1+\mathrm{f}(\theta)] \Leftrightarrow 4\mathrm{f}(\theta)\mathrm{g}(\theta)}{[1+\mathrm{f}(\theta)+\mathrm{g}(\theta)]^2} \\
&= \frac{\mathrm{f}(\theta)+\mathrm{f}(\theta)\mathrm{g}(\theta)+4\mathrm{g}(\theta)}{[1+\mathrm{f}(\theta)+\mathrm{g}(\theta)]^2}. \qquad (1.22)
\end{aligned}$$

Substituting (1.21) into (1.22) it is possible to express $\mathrm{I}(\theta)$ in $\mathrm{f}(\theta)$ and $y$ only:

$$\begin{aligned}
\mathrm{I}(\theta) &= \frac{\mathrm{f}(\theta)+\mathrm{f}(\theta)[\mathrm{f}(\theta)]^2/y + 4[\mathrm{f}(\theta)]^2/y}{\{1+\mathrm{f}(\theta)+[\mathrm{f}(\theta)]^2/y\}^2} \\
&= \frac{y\mathrm{f}(\theta)\left\{y+[\mathrm{f}(\theta)]^2+4\mathrm{f}(\theta)\right\}}{\{y+y\mathrm{f}(\theta)+[\mathrm{f}(\theta)]^2\}^2}. \qquad (1.23)
\end{aligned}$$

For a bimodal information function the locations of the maxima, expressed in $\xi = \exp(\theta)$ and $\varepsilon_k = \exp(\Leftrightarrow\delta_k)$, may be rewritten from the expression just above (1.17):

$$\xi_{1,2} = \frac{1}{2\varepsilon_1}\left\{\frac{\varepsilon_1}{\varepsilon_2} \Leftrightarrow 8 \pm \left[\left(\frac{\varepsilon_1}{\varepsilon_2} \Leftrightarrow 8\right)^2 \Leftrightarrow 4\frac{\varepsilon_1}{\varepsilon_2}\right]^{1/2}\right\}.$$

Noting that $\varepsilon_1/\varepsilon_2 = \exp(\delta_2 \Leftrightarrow \delta_1) = y$, the locations of the maxima expressed in $\theta$ are given by

$$\theta_{1,2} = \ln\left\{\frac{1}{2\exp(\Leftrightarrow\delta_1)}\left\{y \Leftrightarrow 8 \pm \left[(y \Leftrightarrow 8)^2 \Leftrightarrow 4y\right]^{1/2}\right\}\right\}.$$

Concentrate for the moment on the maximum located at $\theta_1$. Evaluating $\mathrm{f}(\theta)$ for $\theta = \theta_1$ will give, with $\mathrm{f}(\theta_1) = \exp(\theta_1 \Leftrightarrow \delta_1)$ :

$$\mathrm{f}(\theta_1) = \frac{1}{2}\left\{y \Leftrightarrow 8 + \left[(y \Leftrightarrow 8)^2 \Leftrightarrow 4y\right]^{1/2}\right\}. \tag{1.24}$$

Instead of directly trying to evaluate $\mathrm{I}(\theta)$ at $\theta = \theta_1$, note that with the help of (1.24) it is possible to express $[\mathrm{f}(\theta_1)]^2$ in $\mathrm{f}(\theta_1)$ and $y$:

$$\begin{aligned}
[\mathrm{f}(\theta_1)]^2 &= \frac{1}{4}\left\{(y \Leftrightarrow 8)^2 + 2(y \Leftrightarrow 8)\left[(y \Leftrightarrow 8)^2 \Leftrightarrow 4y\right]^{1/2} + (y \Leftrightarrow 8)^2 \Leftrightarrow 4y\right\} \\
&= \frac{1}{2}(y \Leftrightarrow 8)\left\{y \Leftrightarrow 8 + \left[(y \Leftrightarrow 8)^2 \Leftrightarrow 4y\right]^{1/2}\right\} \Leftrightarrow y \\
&= (y \Leftrightarrow 8)\mathrm{f}(\theta_1) \Leftrightarrow y. \qquad (1.25)
\end{aligned}$$

Now in order to evaluate $\mathrm{I}(\theta)$ at $\theta = \theta_1$, first, in (1.23), substitute $[\mathrm{f}(\theta_1)]^2$ by the expression derived in (1.25):

$$\begin{aligned}
\mathrm{I}(\theta_1) &= \frac{y\mathrm{f}(\theta_1)\left[y + (y \Leftrightarrow 8)\mathrm{f}(\theta_1) \Leftrightarrow y + 4\mathrm{f}(\theta_1)\right]}{\left[y + y\mathrm{f}(\theta_1) + (y \Leftrightarrow 8)\mathrm{f}(\theta_1) \Leftrightarrow y\right]^2} \\
&= \frac{y\mathrm{f}(\theta_1)[y\mathrm{f}(\theta_1) \Leftrightarrow 4\mathrm{f}(\theta_1)]}{[2y\mathrm{f}(\theta_1) \Leftrightarrow 8\mathrm{f}(\theta_1)]^2} \\
&= \frac{y}{4(y \Leftrightarrow 4)}.
\end{aligned}$$

For the maximum located at $\theta_2$ the derivation is similar. Therefore the value of the maximum of a bimodal information function follows upon replacing $y$ by $\exp(\delta_2 \Leftrightarrow \delta_1)$ in the above expression, yielding

$$\mathrm{I}(\theta_1) = \mathrm{I}(\theta_2) = \frac{\exp(\Leftrightarrow \delta_1)}{4[\exp(\Leftrightarrow \delta_1) \Leftrightarrow 4\exp(\Leftrightarrow \delta_2)]}.$$

# Chapter 2

# Polytomous item scores and Guttman dependence[1]

Abstract

Some theoretical relations are established between the score on a polytomous item and the total score on a set of Guttman dependent binary items. Conditions are derived under which these two scores are identically distributed. Application of the theoretical results to three well-known models for polytomous data yields, among others, that the score on a graded response item (Samejima, 1969) is never distributed as the total score on a set of independent binary Rasch items (Rasch, 1960).

Key words: Guttman dependence, graded response model, partial credit model, sequential model.

## 2.1 Introduction

There are several models for polytomous items, between which it is sometimes hard to decide. Breaking down polytomous items into dichotomous item steps, as is for example done by Van Engelenburg (1997, chapter 2), may facilitate the process of choosing an appropriate model for a particular set of polytomous items.

[1] This paper has been accepted by the *British Journal of Mathematical and Statistical Psychology*; it is scheduled for the May 1999 issue.

Bearing this in mind, the present paper aims at the establishment of some theoretical relations between item response models for polytomous, and models for binary data. The interest in differences and similarities between item response models has been growing during the last decade. Some papers focus on models for polytomous items only, such as Andrich' (1996) recent reformulation of the Likert scale (Likert, 1932) as an unfolding model, which clarifies its relation to the Thurstone (1927) procedure. Many authors, however, are concerned with relations between, on the one hand, models for binary, and on the other hand, models for polytomous data. Molenaar (1983) for example explored the mathematical relation of three different models for polytomous items with the Rasch model (Rasch, 1960). Jansen and Roskam (1986) investigated the possibility of dichotomizing graded responses, under several different models for polytomous data. Mellenbergh (1995) takes Bock's nominal response model (Bock, 1972) as a starting point and then distinguishes three different order preserving mechanisms leading to three different types of models for ordinal polytomous data. Van Engelenburg (1997) investigates formal subtask features in order to determine the kind of model that is called for by a particular kind of item. In all these contributions several models for polytomous items are considered. More specific comparisons however, focusing on one particular model, are also made. The polytomous Rasch model, one version of which has been thoroughly investigated by Fischer (1974), has received particular interest. Huynh (1994) for example showed that, under some conditions on the item parameters, the score on a partial credit item (Andersen,1977;Andrich,1978;Masters,1982) is distributed as the total score on a set of independent binary Rasch items; and subsequently he discovers that the score on every partial credit item is distributed as to the total score on a set of independent binary and trinary Rasch items (Huynh, 1996).

The focus in the present paper will be on the relation between response probabilities for a polytomous item, and response probabilities for a set of Guttman dependent binary variables (Guttman, 1950). In the context of item response theory, it is possible to make a distinction between deterministic and stochastic Guttman scales. In a deterministic Guttman scale, the response probability for the binary item $j$ is given by

$$\Pr(Y_j = 1; \theta) = \begin{cases} 0 & \text{for} \quad \theta < \delta_j \\ 1 & \text{for} \quad \theta \geq \delta_j \end{cases} ,$$

where $\delta_j$ could be thought of as an item parameter. Deterministic Guttman scales are very restrictive and therefore of little practical use. Three approaches have been taken to relax the severe restrictions in this model. The practical approach is the construction of an index for the deviation of the empirical data from the perfect Guttman pattern, and the formulation of a rule for the acceptability of this deviation. The literature about these scalability coefficients has been reviewed by e.g. Mokken (1970) and by Cliff (1983). A more fundamental approach was taken by Mokken (1970). Mokken relaxed the assumption of the deterministic response probabilities, and in doing so, his work formed the starting point for a growing body of literature about nonparametric item response theory (see e.g. Mokken and Lewis, 1982; Ellis and van den Wollenberg, 1993; Sijtsma and Junker, 1996; Hemker et al., 1996, 1997). In these nonparametric models however the maximal covariance property, which is an essential part of the Guttman scale, is lost. Tutz (1990, 1997) and Verhelst, Glas and de Vries (1997) could be considered as suggesting a third approach. These authors relaxed the deterministic response probabilities, whilst at the same time retaining the property of maximal covariances. This approach will be taken in the present paper as well. The resulting stochastic Guttman scale of course is still restrictive in that it admits of perfect response patterns only. However, the primary purpose in this paper is not the derivation of a model that offers a realistic description of empirical data. The purpose will rather be to reveal, in Section 2.2, some mathematical relations between the response probabilities for the score on a polytomous item and a for the score on a set of stochastic Guttman dependent binary variables. When these relations have been established, in Section 2.3 their implications for several parametric models for polytomous responses will be investigated. The interest will only be in relationships between the probability mass functions of polytomous and of binary variables. No claims are made concerning relationships between substantive item contents.

It will for example be shown that the score on a graded response item (Samejima,1969) is never distributed as to the total score on a set of independent binary

Rasch items. Comparisons like these may contribute towards an understanding of the kind of data to which models can be profitably applied.

### 2.1.1 Preliminaries: notation and definitions

Consider a set of M binary variables $\boldsymbol{Y} = (Y_1 \ldots Y_j \ldots Y_M)$, where each $Y_j$ takes a value of either 0 or 1. Let $\theta$ be a fixed latent ability. Then,

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}; \theta) = \Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_M = y_M; \theta)$$

will denote the simultaneous probability of observing a the vector $\boldsymbol{Y} = \boldsymbol{y}$ as a function of $\theta$.

The marginal probability of a certain response on one variable, say variable $j$, can be obtained from the simultaneous probability by summing over the probabilities of all other possible responses:

$$\Pr(Y_j = k; \theta) = \sum_{y_1=0}^{1} \cdots \sum_{y_{j-1}=0}^{1} \sum_{y_j=k}^{k} \sum_{y_{j+1}=0}^{1} \cdots \sum_{y_M=0}^{1} \Pr(\boldsymbol{Y} = \boldsymbol{y}; \theta),$$

for $k = 0, 1$. When these marginal probabilities are considered as a function of $\theta$, they will be called operating characteristics (OCs). This term, which is due to Samejima (1969), is preferred here over the term 'item characteristic curve', to stress the emphasis on variables rather than items. Next,

$$\Pr(Y_i = y_i \mid Y_j = y_j; \theta)$$

will denote the conditional probability of the score on variable $i$, given the score on variable $j$, evaluated as a function of $\theta$. The symbol $T$ will be used for the total score on the set of binary variables:

$$T = \sum_{j=1}^{M} Y_j,$$

and the total score functions (TSFs) give the probability, considered as a function of $\theta$, of obtaining a particular total score:

$$\mathrm{TSF}_t(\theta) \quad = \quad \Pr(T = t; \theta), \qquad \text{for} \quad t = 0, 1, \ldots, \mathrm{M}.$$

The ETSF, finally, describes the expectation of the total score as a function of $\theta$:

$$\mathrm{ETSF}(\theta) \quad = \quad E(T;\theta).$$

**Definition 1** *Let a set of M binary variables be indexed such that* $\Pr(Y_i = 1;\theta) \geq \Pr(Y_j = 1;\theta)$ *for all* $\theta$ *and all* $i < j$*. This set of variables will be called Guttman dependent if*

$$\Pr(Y_i = 1 \mid Y_j = 1;\theta) = 1$$

*for* $j = 2, 3, \ldots, M$ *and all* $i < j$*, and for all values of* $\theta$*.*

As will be shown below, the OCs are nonintersecting. Furthermore, if $\Pr(Y_k = 1;\theta) = \Pr(Y_r = 1;\theta)$ for some $k$ and $r$ and for all $\theta$, it is possible to use either $k = r \Leftrightarrow 1$ or $k = r + 1$. In a set of Guttman dependent binary variables only perfect response patterns occur: with the indexing convention from Definition 1, and noting that if $\Pr(Y_k = 1;\theta) = \Pr(Y_{k+1} = 1;\theta)$ then $\Pr(T = k;\theta) = 0$, a total score $T = t$ on a set of Guttman dependent binary variables implies that the first $t$ variables are equal to 1 and the last M $\Leftrightarrow t$ variables are equal to 0. From this property it follows that, for a set of Guttman dependent binary variables, $\Pr(Y_j = 1;\theta) = \Pr(T \geq j;\theta)$, for $j = 1, 2, \ldots, \mathrm{M}$. Hence, the total score functions (TSFs) for a set of Guttman dependent binary variables are given by:

$$\begin{aligned} \Pr(T = t;\theta) &= \Pr(T \geq t;\theta) \Leftrightarrow \Pr(T \geq t+1;\theta) \\ &= \Pr(Y_t = 1;\theta) \Leftrightarrow \Pr(Y_{t+1} = 1;\theta), \quad \text{for} \quad t = 0, 1, \ldots, \mathrm{M}, \end{aligned} \quad (2.1)$$

where, for notational convenience, $\Pr(Y_0 = 1;\theta) \equiv 1$ and $\Pr(Y_{M+1} = 1;\theta) \equiv 0$. Consequently, the OCs are nonintersecting otherwise $\Pr(T = t;\theta)$ would be negative for some $\theta$. It may be noted that a set of Guttman dependent binary variables may, but need not, consist of variables with deterministic OCs.

A polytomous variable is defined as a variable with possible scores $0, 1, 2, \ldots, \mathrm{N}$. Below, only one polytomous variable will be considered at a time, so that there is no need here for a separate variable index. The index $k$ can therefore be reserved for the categories of and the score on a polytomous variable; he index $j$ will be used to indicate variables from a set of binary variables. In order to be able to distinguish scores on binary from scores on polytomous variables, from now on the symbol $X$ will be used to indicate the score on a polytomous variable; the symbol $Y$ will be continued for use with binary variables. The marginal response probabilities on a polytomous variable, considered as a function of $\theta$, will be called category response functions or CRFs:

$$\mathrm{CRF}_k(\theta) \quad = \quad \Pr(X = k; \theta), \qquad \text{for} \quad k = 0, 1, \ldots, \mathrm{N}.$$

Cumulative category response functions (CCRFs) give the probability of obtaining a score of $k$ or higher on a polytomous variable, as a function of $\theta$:

$$\mathrm{CCRF}_k(\theta) \quad = \quad \Pr(X \geq k; \theta) \qquad \text{for} \quad k = 0, 1, \ldots, \mathrm{N}.$$

The expected value of the score on a polytomous variable, as a function of $\theta$, will be referred to as the expected-response function ERF:

$$\mathrm{ERF}(\theta) \quad = \quad E(X; \theta).$$

The term 'response probability' will be used in the context both of binary and of polytomous variables, to indicate the probability of obtaining a particular score.

In the sequel, F will be a set of OCs for M binary variables, and the separate OCs in the set will be denoted as $\mathrm{f}_1 \ldots \mathrm{f}_j \ldots \mathrm{f}_M$, which are functions of $\theta$. For notational purposes it will also be convenient to define

$$\mathrm{f}_0(\theta) \;\equiv\; 1 \quad \text{and} \quad \mathrm{f}_{M+1}(\theta) \;\equiv\; 0,$$

for all $\theta$. Furthermore, G will be a set of N + 1 CRFs for a polytomous variable, and the separate CRFs in the set will be denoted as $\mathrm{g}_0 \ldots \mathrm{g}_k \ldots \mathrm{g}_N$.

Assuming that the response probabilities on both the binary and the polytomous variables are governed by the same trait $\theta$, this gives:

$$\mathrm{F} \quad = \quad \{\mathrm{f}_1 \ldots \mathrm{f}_j \ldots \mathrm{f}_M\}\,, \qquad \text{where } \mathrm{f}_j(\theta) \;=\; \Pr(Y_j = 1; \theta)$$

$$\mathrm{G} \quad = \quad \{\mathrm{g}_0 \ldots \mathrm{g}_k \ldots \mathrm{g}_N\}\,, \qquad \text{where } \mathrm{g}_k(\theta) \;=\; \Pr(X = k; \theta).$$

Below, unless explicitly stated otherwise, the only assumptions which the curves in sets F and G are required to meet are the following:

1. $0 \leq \mathrm{f}_j(\theta) \leq 1$ for all $\theta$ and $j = 1, \ldots, \mathrm{M}$;

2a. $0 \leq \mathrm{g}_k(\theta) \leq 1$ for all $\theta$ and $k = 0, 1, \ldots, \mathrm{N}$;

2b. $\sum_{k=0}^{N} \mathrm{g}_k(\theta) = 1$;

3. $\mathrm{f}_j(\theta)$ nor $\mathrm{g}_k(\theta)$ is equal to 0 for all $\theta$, for any $j \in 1 \ldots \mathrm{M}$ or $k \in 0 \ldots \mathrm{N}$.

These assumptions will in the sequel be referred to as Assumptions 1 - 3.

Finally, the set of all possible F will be denoted as $\mathcal{U}_F$, the set of all possible G as $\mathcal{U}_G$. In Section 2.2.3 a restriction will be placed on $\mathcal{U}_F$ and $\mathcal{U}_G$, but until then all sets F and G considered will just be elements of $\mathcal{U}_F$ and $\mathcal{U}_G$.

**Definition 2** *Two variables $P$ and $Q$ will be called identically distributed for every value of $\theta$ if*

$$\Pr(P = k; \theta) \quad = \quad \Pr(Q = k; \theta) \qquad \textit{for all } k \textit{ and all } \theta.$$

Hence, if $X$ is the score on a polytomous variable and $T$ the total score on a set of binary variables, $X$ and $T$ are identically distributed if and only if the CRFs of the polytomous variable are pairwise identical with the TSFs of the set of binary variables. If $X$ and $T$ are identically distributed then obviously, by Assumption 3, M is equal to N. In the sequel, therefore, only polytomous variables with N = M are considered. Furthermore, it will be convenient to have a term for the relation between a polytomous variable and a set of binary variables, whose (total) scores are identically distributed. In this case the polytomous variable and the set of binary variables will be called distributionally identical.

It must be stressed at this point that the marginal distribution of $T = \sum_j Y_j$ does not, in general, reveal anything about the simultaneous probability distribution $\Pr(\boldsymbol{Y} = \boldsymbol{y}; \theta)$. Therefore it is possible for the total score on several sets of binary variables having different OCs and different covariance structures, to be identically distributed. This is demonstrated by the following two examples, where $\varepsilon$ is short for $\exp(\Leftrightarrow\delta)$ and $\xi$ for $\exp(\theta)$:

**Example A** Let variables 1 and 2 have $\mathrm{f}_1(\theta) = [(\varepsilon_1 + \varepsilon_2)\xi + \varepsilon_1\varepsilon_2\xi^2]/\mathrm{D}$, and $\mathrm{f}_2(\theta) = \varepsilon_1\varepsilon_2\xi^2/\mathrm{D}$, where $\mathrm{D} = 1 + (\varepsilon_1 + \varepsilon_2)\xi + \varepsilon_1\varepsilon_2\xi^2$. Assume the two variables are Guttman dependent. Then, using (2.1), the TSFs for these two variables are given by $\Pr(T = 1; \theta) = \mathrm{f}_1(\theta) \Leftrightarrow \mathrm{f}_2(\theta) = (\varepsilon_1 + \varepsilon_2)\xi/\mathrm{D}$, and $\Pr(T = 2; \theta) = \mathrm{f}_2(\theta) = \varepsilon_1\varepsilon_2\xi^2/\mathrm{D}$. The probability of a zero score equals $1 \Leftrightarrow \Pr(T = 1; \theta) \Leftrightarrow \Pr(T = 2; \theta)$.

**Example B** Consider two binary variables with scores $Y_1'$ and $Y_2'$, and assume the scores on these variables are independent, given $\theta$. In this case let the OCs be given by $\mathrm{f}_1'(\theta)$ and $\mathrm{f}_2'(\theta)$, respectively. Assume $\mathrm{f}_1'(\theta) = \varepsilon_1\xi/(1 + \varepsilon_1\xi)$ and $\mathrm{f}_2'(\theta) = \varepsilon_2\xi/(1 + \varepsilon_2\xi)$, where $\varepsilon_1$ and $\varepsilon_2$ have the same value as in example A. Using the independence assumption, this gives $\Pr(T' = 1; \theta) = (\varepsilon_1 + \varepsilon_2)\xi/\mathrm{D}$; and $\Pr(T' = 2; \theta) = \varepsilon_1\varepsilon_2\xi^2/\mathrm{D}$.

Clearly, the TSFs in Example A are equal to the ones in Example B, although both the OCs and the covariance structures are different.

A final definition concerns a relation between two sets of response curves F and G:

**Definition 3** *Consider a set $F \in \mathcal{U}_F$ and a set $G \in \mathcal{U}_G$, satisfying Assumptions 1 - 3. F is a set of M curves $f_1 \dots f_j \dots f_M$, and G is a set of $M + 1$ curves $g_0 \dots g_k \dots g_M$; and recall that $f_0 \equiv 1$ and $f_{M+1} \equiv 0$. Between the sets F and G there exists an additive relation (an AR-S) if*

$$f_j(\theta) = \sum_{k=j}^{M} g_k(\theta), \quad \textit{for all } \theta \textit{ and for } j = 1, \dots, M, \tag{2.2a}$$

*which can also be written as*

$$g_k(\theta) = f_k(\theta) \Leftrightarrow f_{k+1}(\theta), \quad \textit{for all } \theta \textit{ and for } k = 0, 1, \dots, M. \tag{2.2b}$$

*Two sets F and G related by an AR-S will be called 'additively related sets'.*

Note the following: (a) the final S in AR-S indicates that the relation is between two sets of curves; it is there to distinguish it from another additive relation which will be defined below, and which will be denoted AR-F; (b) this definition is about sets of curves only and therefore does not involve any covariance assumptions; (c) from formulation (2.2b) it follows that if an AR-S exists between two sets F and G, the curves $\mathrm{f}_1 \ldots \mathrm{f}_M$ must be nonintersecting, otherwise the curves $\mathrm{g}_k$ would become negative; (d) from formulation (2.2a) it can be deduced that $\mathrm{f}_i(\theta) \geq \mathrm{f}_j(\theta)$ for all $i < j$; (e) finally, note that an AR-S is impossible between two sets F and G having $\mathrm{M} \neq \mathrm{N}$, as for example $\mathrm{M} < \mathrm{N}$ would give $\mathrm{g}_N(\theta) = \mathrm{f}_N(\theta) \Leftrightarrow \mathrm{f}_{N+1}(\theta) = 0 \Leftrightarrow 0 = 0$ for all $\theta$, which does not obey Assumption 3.

## 2.2 Polytomous variables and Guttman dependence

Below some lemmas and theorems will be proved, concerning the relation of the distribution of the score on a polytomous variable, and the distribution of the total score on a set of binary variables, both considered as a function of $\theta$. The emphasis in this section will be on the mathematical derivations; a substantive interpretation of the results will be given in Section 2.3. The content of the first lemma is not very surprising, but it will serve to structure the discussion in this section.

**Lemma 1** *Consider*

1. *the score $X$ on a polytomous variable with response curves $g_0 \ldots g_k \ldots g_M$ forming a set $G \in \mathcal{U}_G$; and*
2. *the total score $T = \sum_{j=1}^{M} Y_j$ on a set of $M$ binary variables with response probabilities $f_1 \ldots f_j \ldots f_M$ forming a set $F \in \mathcal{U}_F$.*

*Then the following implication holds:*

$$\left.\begin{array}{ll} (a). & f_j(\theta) = \sum_{k=j}^{M} g_k(\theta) \textit{ for } j = 1 \ldots M; \\ (b). & \Pr(Y_i = 1 | Y_j = 1; \theta) = 1 \textit{ for all } i < j. \end{array}\right\} \Rightarrow \begin{array}{c} \Pr(X = k; \theta) = \Pr(T = k; \theta) \\ \textit{for } k = 0 \ldots M. \end{array}$$

*Proof.* The implication can be easily verified using (2.1) and Definition 3.

□

The lemma states that if the variables $Y_1 \ldots Y_M$ are Guttman dependent and if there is an AR-S between the sets F and G, then $X$ and $T$ will be identically distributed. The reverse of this statement, i.e. the implication from right to left, does not hold, as can be demonstrated by a counterexample: if $X$ is distributed as the total score in Example A from Section 2.1.1, then $X$ is also distributed as the total score in Example B, although in this latter situation the left hand side properties (a) and (b) do not hold.

Each of the two left hand side conditions (a) and (b) of this lemma will now in turn be considered not as a condition, but as an assumption, and it will be demonstrated that in each case this changes the implication into an equivalence. In Section 2.2.3 another assumption will be introduced, which will also change the implication into an equivalence.

### 2.2.1 Unique OCs

Let $X$ be the score on a polytomous variable, and $T = \sum_{j=1}^{M} Y_j$ be the total score on a set of Guttman dependent binary variables. Hence, the covariance structure of this set of binary variables is known. No assumptions are made about the OCs of the binary variables, except the ones stated in Assumptions 1 - 3. In this situation it can be shown that an AR-S between sets F and G is a necessary and sufficient condition for the distributions of the score $X$ on a polytomous variable and the total score and $T$ on a set of Guttman dependent binary variables to be identical:

**Theorem 1** *Consider $X$ and $T$ as defined in the preamble to Lemma 1, and assume that $\Pr(Y_i = 1|Y_j = 1;\theta) = 1$ for all $i < j$. Then the following equivalence holds:*

$$\left.\begin{array}{c}\Pr(X = k;\theta) = \Pr(T = k;\theta)\\ \textit{for } k = 0\ldots M\end{array}\right\} \quad \Leftrightarrow \quad \left\{\begin{array}{c}f_j(\theta) = \sum_{k=j}^{M} g_k(\theta)\\ \textit{for } j = 1\ldots M\end{array}\right.$$

*Proof.* Right to left: using the Guttman dependence assumption, this proof is equal to the proof of Lemma 1.

Left to right. The OCs for variables $1, 2, \ldots, \mathrm{M}$ are given by $\mathrm{f}_1 \ldots \mathrm{f}_M$. Under the assumed Guttman dependence property it holds, using (2.1), that $\Pr(T = k; \theta) = \mathrm{f}_k(\theta) \Leftrightarrow \mathrm{f}_{k+1}(\theta)$, for $k = 0 \ldots \mathrm{M}$. The left hand side condition $\Pr(X = k; \theta) = \Pr(T = k; \theta)$ now gives

$$\mathrm{g}_k(\theta) = \mathrm{f}_k(\theta) \Leftrightarrow \mathrm{f}_{k+1}(\theta) \quad \text{for} \quad k = 0, 1, \ldots, \mathrm{M}.$$

Because $\mathrm{f}_{M+1} \equiv 0$, it follows that $\mathrm{f}_M = \mathrm{g}_M$ for all values of $\theta$. Substituting this into $\mathrm{g}_{M-1} = \mathrm{f}_{M-1} \Leftrightarrow \mathrm{f}_M =$ will yield that $\mathrm{f}_{M-1} = \mathrm{g}_{M-1} + \mathrm{g}_M$. Repeating the argument all the way down to the requirement $\mathrm{f}_1 = \sum_{k=1}^{M} \mathrm{g}_k$ for all $\theta$ establishes the implication from left to right. Combining both parts of the proof completes the proof of the theorem.

□

### 2.2.2 Unique covariance structure

In the previous theorem the Guttman dependence from Lemma 1 was made an assumption; now a theorem will be stated in which, instead of the Guttman dependence, the shape of the OCs is drawn into the assumptions. So it will be shown that assuming an AR-S, $X$ and $T$ are identically distributed if and only if the binary variables are Guttman dependent.

**Theorem 2** *Consider again $X$ and $T$ as defined in the preamble to Lemma 1, and assume an AR-S between the sets $F$ and $G$. Then the following equivalence holds:*

$$\left.\begin{array}{c} \Pr(X = k; \theta) = \Pr(T = k; \theta) \\ \text{for } k = 0, 1, \ldots, M \end{array}\right\} \Leftrightarrow \left\{\begin{array}{c} \Pr(Y_i = 1 | Y_j = 1; \theta) = 1 \\ \text{for } j = 2, \ldots, M \text{ and all } i < j. \end{array}\right.$$

*Proof.* Right to left: using the AR-S assumption, this proof is equal to the proof of Lemma 1.

Left to right. First define $\mathrm{B}_q$, for $q = 2, 3, \ldots, \mathrm{M}$, as

$$\mathrm{B}_q =$$

$$\Pr(Y_1 = 0, \ldots, Y_{q-1} = 0, Y_{q+1} = 0, \ldots, Y_M = 0 \mid Y_q = 0; \theta)$$

$$+ \sum_{r=1}^{q-2} \Pr(Y_1 = 1, \ldots, Y_r = 1, Y_{r+1} = 0, \ldots, Y_{q-1} = 0,$$

$$Y_{q+1} = 0, \ldots, Y_M = 0 \mid Y_q = 0; \theta)$$

$$+ \Pr(Y_1 = 1, \ldots, Y_{q-1} = 1, Y_{q+1} = 0, \ldots, Y_M = 0 \mid Y_q = 0; \theta). \quad (2.3)$$

Next, consider a set of M binary variables in which only perfect response patterns occur for $T = 0, 1, \ldots, q \Leftrightarrow 1$. In this situation the probability $\Pr(T \leq q \Leftrightarrow 1; \theta)$ can be written using the quantity $\mathrm{B}_q$:

$$\begin{aligned} \Pr(T \leq q \Leftrightarrow 1; \theta) &= \mathrm{B}_q[\Pr(Y_q = 0; \theta)] \\ &= \mathrm{B}_q[1 \Leftrightarrow \mathrm{f}_q(\theta)], \text{ for } q = 2, \ldots, \mathrm{M}. \qquad (2.4) \end{aligned}$$

The proof of the implication from left to right will now be given by means of mathematical induction. First note that obviously for $k = 0$, where $k$ refers to the index in the left hand side of the theorem, the only pattern possible is to have $Y_j = 0$ for all $j$.

*Proof for* $k = 1$, where again $k$ refers to the index in the left hand side of the theorem.

$$\begin{aligned} \Pr(T \leq 0; \theta) &= \Pr(Y_1 = 0, \ldots, Y_M = 0; \theta) \\ &= \Pr(Y_2 = 0, \ldots, Y_M = 0 \mid Y_1 = 0; \theta) \Pr(Y_1 = 0; \theta) \\ &= \Pr(Y_2 = 0, \ldots, Y_M = 0 \mid Y_1 = 0; \theta)[1 \Leftrightarrow \mathrm{f}_1(\theta)]. \qquad (2.5) \end{aligned}$$

However, it also holds that

$$\begin{aligned}
\Pr(T \leq 0; \theta) &= \Pr(X \leq 0; \theta) \quad \text{(by the left hand side condition)} \\
&= 1 \Leftrightarrow \Pr(X > 1; \theta) = 1 \Leftrightarrow \sum_{r=1}^{M} \mathrm{g}_r(\theta) \\
&= 1 \Leftrightarrow \mathrm{f}_1(\theta) \quad \text{(by the AR-S)}\ , \qquad (2.6)
\end{aligned}$$

so that, from (2.5) and (2.6):

$$\Pr(Y_2 = 0, \ldots, Y_M = 0 \mid Y_1 = 0) = 1.$$

Assuming the left hand side to hold, therefore, any response pattern having the combination $(Y_1 = 0, T \geq 1)$ is impossible, which will be referred to as Consequence 1 in the sequel. In particular, the combination $(Y_1 = 0, T = 1)$ is impossible so that $T = 1$ implies $Y_1 = 1$. This establishes the first result in the induction chain: the only pattern possible for $k = 1$ is a Guttman pattern.

*Proof for* $k = q$. By the induction assumption the Guttman property now holds for $k = 0, 1, \ldots, q \Leftrightarrow 1$, so that it is possible to apply (2.4) and write

$$\Pr(T \leq q \Leftrightarrow 1; \theta) \quad = \quad \mathrm{B}_q[1 \Leftrightarrow \mathrm{f}_q(\theta)].$$

However, reasoning as in the proof for $k = 1$, it also holds that

$$\begin{aligned}
\Pr(T \leq q \Leftrightarrow 1; \theta) &= \Pr(X \leq q \Leftrightarrow 1; \theta) \text{ (by the left hand side condition)} \\
&= 1 \Leftrightarrow \mathrm{f}_q(\theta) \quad \text{(by the AR-S)}\ ,
\end{aligned}$$

so that

$$\mathrm{B}_q = 1.$$

Consequently, the response patterns implied in (2.3) are the only response patterns with $Y_q = 0$ and therefore the combination $(Y_q = 0, T \geq q)$ is impossible (Consequence $q$). In particular, the combination $(Y_q = 0, T = q)$ is impossible so that any pattern leading to $T = q$ has to have $Y_q = 1$. However, Consequences $1, \ldots, q \Leftrightarrow 1$ taken together imply that $(Y_r = 0, T =$

$q$) is impossible for all $r \leq q \Leftrightarrow 1$ as well. It follows that the only pattern for $T = q$ can be the pattern with 1s on the first $q$ positions, which is the $q$'th pattern in a Guttman sequence.

The induction chain can be carried through as far as $k = \mathrm{M} \Leftrightarrow 1$. Then finally the only pattern possible for $T = \mathrm{M}$ is $(11 \ldots 11)$, which completes the proof from left to right.

□

### 2.2.3 Families of sets of response curves

Until now, the emphasis has been on sets $\mathrm{F} \in \mathcal{U}_F$ and $\mathrm{G} \in \mathcal{U}_G$. In Section 2.3 however, several item response models will be considered. Most item response models entail a restriction on the universe of sets of response curves. Let $\mathcal{F} \subset \mathcal{U}_F$ and $\mathcal{G} \subset \mathcal{U}_G$ represent such restrictions, and consider, in this section, $\mathrm{F} \in \mathcal{F} \subset \mathcal{U}_F$ and $\mathrm{G} \in \mathcal{G} \subset \mathcal{U}_G$. The subsets $\mathcal{F}$ and $\mathcal{G}$, which are sets of sets, will be called families of sets of curves, or just short families. As an example, $\mathcal{G}$ could be the family of sets of response curves obeying the graded response model (Samejima, 1969), or the partial credit model (Masters, 1982).

In the subsequent derivations it will be assumed that all $\mathrm{F} \in \mathcal{F}$ and $\mathrm{G} \in \mathcal{G}$ are unique: there are no $\mathrm{F}_1$ and $\mathrm{F}_2$ in family $\mathcal{F}$, nor $\mathrm{G}_1$ and $\mathrm{G}_2$ in family $\mathcal{G}$, whose curves completely coincide.

Assuming two families $\mathcal{F}$ and $\mathcal{G}$ to be defined, a relation between them can be defined:

**Definition 4** *Consider*

1. *a family $\mathcal{F} \subset \mathcal{U}_F$ of unique sets F, and*
2. *a family $\mathcal{G} \subset \mathcal{U}_G$ of unique sets G.*

*There exists an additive relation between the families $\mathcal{F}$ and $\mathcal{G}$ (an AR-F) if both*

- *for every set of curves in family $\mathcal{F}$, say the set $F_0$, there exists a set of curves in family $\mathcal{G}$, say the set $G_0$, and*
- *for every set of curves in family $\mathcal{G}$, say the set $G_1$, there exists a set of curves in family $\mathcal{F}$, say the set $F_1$,*

*such that an AR-S exist between the sets $F_0$ and $G_0$, and the sets $G_1$ and $F_1$.*

It was previously remarked that an AR-S is impossible between two sets F and G having M $\neq$ N. Similarly, an AR-F is only possible between two families $\mathcal{F}$ and $\mathcal{G}$ consisting of the same numbers of sets with the same number of curves. As an example, if family $\mathcal{G}$ consists of, say, 100 sets, of which 70 have N = 2 and 30 have N = 4, then, in order for the AR-F to be defined, family $\mathcal{F}$ will consist of 100 sets of which 70 have M = 2 and 30 have M = 4 as well.

Any two sets F and G between which an AR-S holds, were called 'additively related sets' in Definition 3. Because all sets F $\in \mathcal{F}$ and G $\in \mathcal{G}$ are unique, it can be deduced from the first formulation in Definition 3 that if the set $\mathrm{F}_0$ is additively related to the set $\mathrm{G}_0$, there is no other set in $\mathcal{F}$ that can be additively related to $\mathrm{G}_0$. From the second formulation in Definition 3 it can be similarly concluded that if set $\mathrm{G}_0$ is additively related to $\mathrm{F}_0$, no other set in $\mathcal{G}$ can be additively related to $\mathrm{F}_0$. Hence the AR-F relation is a bijection.

Compare Definition 4 to Definition 3, which describes a relation between two single sets of curves. The present definition describes a relation between two families of sets of curves, where for every member of one family there exists a member of the other family such that between these two members the AR-S from Definition 3 holds.

Next, the property of unique correspondence will be defined. This concept was defined by Chang and Mazzeo (1994). In the present paper, it will be used in two different contexts, therefore its definition consists of two parts. Recall that the

ERF of a polytomous variable is the expectation function $E(X;\theta)$, and that the ETSF of a set of binary variables is the expectation function of the total score $E(T;\theta)$.

**Definition 5** *A. There is a unique correspondence between the expected-response functions ERF and the sets of category response functions CRF in family $\mathcal{G}$, if, in this family, each ERF can be the result of one and only one set G of CRFs.*

*B. There is a unique correspondence between the expectation functions for the total score ETSF and the sets of operating characteristics OC in family $\mathcal{F}$, if, in this family, each ETSF can be the result of one and only one set of OCs.*

As $E(T;\theta) = \sum_{t=0}^{M} t \Pr(T = t)$, one might expect a covariance structure in the second part of the definition. However, there is no need for that, as $E(T;\theta)$ can also be written as $E(\Sigma Y_j;\theta) = \Sigma[E(Y_j;\theta)]$, which is independent of the covariances. Furthermore, it may be noted that the size of the sets (i.e. their number of curves) is not included in the definition: if the unique correspondence property holds in a family consisting of sets of differing size, then still each ERF (or ETSF) is the result of one and only one set of CRFs (or OCs).

It may be noted that neither the AR-F nor the unique correspondence property require that N (or M) be equal for different sets in the family.

The probability of obtaining a score $k$ on a polytomous variable following the response curves in a particular set G from family $\mathcal{G}$ will from now on be written as $\Pr(X = k;\theta,\mathrm{G}) = \mathrm{g}_k(\theta;\mathrm{G})$. Likewise, the probability of a correct response on variable $j$ from a set F in family $\mathcal{F}$ can be written as $\Pr(Y_j = 1;\theta,\mathrm{F}) = \mathrm{f}_j(\theta;\mathrm{F})$.

Using Definitions 4 and 5, another lemma can be proved. Prior to stating the lemma, note that 'a set of CCRFs' is meant as the entire set of curves $\sum_{k=r}^{N} \mathrm{g}_k$, for $r = 0, 1, \ldots, \mathrm{N}$.

**Lemma 2** *Consider families $\mathcal{F}$ and $\mathcal{G}$ as defined in the preamble to Definition 4* *and assume an additive relation AR-F between these two families. Then the following equivalence holds:*

$$\left.\begin{array}{c}\textit{A unique correspondence between}\\ \textit{ERF and CRFs}\\ \textit{in family } \mathcal{G}\end{array}\right\} \Leftrightarrow \left\{\begin{array}{c}\textit{A unique correspondence between}\\ \textit{ETSF and OCs}\\ \textit{in family } \mathcal{F}.\end{array}\right.$$

*Proof.* Before demonstrating the two implications, three preliminary remarks will be made.

Remark 1. If there is a unique correspondence between ERF and CRFs in family $\mathcal{G}$, then in this family there also is a unique correspondence between ERF and CCRFs, and vice versa.

Remark 2. For any two additively related sets $\mathrm{F}_0$ and $\mathrm{G}_0$, it holds that

$$E(X;\theta,\mathrm{G}_0) = E(T;\theta,\mathrm{F}_0), \tag{2.7}$$

i.e. the ERF for set $\mathrm{G}_0$ exactly coincides with the ETSF for set $\mathrm{F}_0$ and vice versa. Therefore, if there is an AR-F between the families $\mathcal{F}$ and $\mathcal{G}$, any two ERFs in $\mathcal{F}$ are distinct if and only if the ETSFs in the additively related sets are distinct.

Remark 3. Consider two additively related sets $\mathrm{F}_0$ and $\mathrm{G}_0$, and deduce from the definition of the AR-S that the CCRFs of $\mathrm{G}_0$ exactly coincide with the OCs of $\mathrm{F}_0$, and vice versa. Therefore, if there is an AR-F (which is a bijection) between the families $\mathcal{F}$ and $\mathcal{G}$, then since all sets $\mathrm{F} \in \mathcal{F}$ and $\mathrm{G} \in \mathcal{G}$ are unique, every set of CCRFs in $\mathcal{G}$ coincides with exactly one set of OCs in $\mathcal{F}$, and vice versa.

Using these three preliminary observations, the implication from left to right is shown as follows. By the left hand side condition there is a unique correspondence between ERF and CRFs in family $\mathcal{G}$, and hence there also is a unique correspondence between ERF and CCRFs in family $\mathcal{G}$ (remark 1). Furthermore, every distinct ETSF $E(T;\theta,\mathrm{F})$ in family $\mathcal{F}$ coincides with a distinct ERF in family $\mathcal{G}$ (remark 2), and every distinct set of OCs in family $\mathcal{F}$ coincides with a distinct set of CCRFs in family $\mathcal{G}$ (remark 3). Consequently there must be a unique correspondence between ETSF and OCs in family $\mathcal{F}$.

The proof of the implication from right to left is more or less analogous: every distinct ERF in family $\mathcal{G}$ coincides with a distinct ETSF in family $\mathcal{F}$ (remark 2), and every distinct set of CCRFs in family $\mathcal{G}$ coincides with a distinct set of OCs in family $\mathcal{F}$ (remark 3). By the right hand side condition there is only one set of OCs that can possibly lead to a certain ETSF $E(T; \theta, \mathrm{F})$ in family $\mathcal{F}$, and therefore there is also only one set of CCRFs that can lead to the additively related ERF in family $\mathcal{G}$. Because of remark 1, finally, there is a unique correspondence between ERF and CRFs in family $\mathcal{G}$.

□

Below a theorem will be stated, in which an additional assumption renders the implication from Lemma 1 into a full equivalence. One final preliminary lemma will facilitate the proof in this theorem.

**Lemma 3** *Consider a family $\mathcal{F} \subset \mathcal{U}_F$ and a family $\mathcal{G} \subset \mathcal{U}_G$. In both families the unique correspondence property is assumed to hold. Furthermore, assume an additive relation AR-F between the families $\mathcal{F}$ and $\mathcal{G}$, and*

1. *let $X$ be the score on a polytomous variable with response curves G, and*
2. *let $T = \sum_{j=1}^{M} Y_j$ be the total score on a set of M binary variables with response curves F.*

*Then for all $F \in \mathcal{F}$ and all $G \in \mathcal{G}$ it holds that*

$$\left.\begin{array}{c} \Pr(X = k; \theta, G) = \Pr(T = k; \theta, F) \\ \textit{for } k = 0 \ldots M \end{array}\right\} \Rightarrow \left\{\begin{array}{c} f_j(\theta; F) = \sum_{k=j}^{M} g_k(\theta; G) \\ \textit{for } j = 1 \ldots M. \end{array}\right.$$

*Proof.* Let the set G on the left hand side be the set $\mathrm{G}_0$, and let $\mathrm{F}_0$ be the set that is additively related to $\mathrm{G}_0$. If the set F on the left hand side is the set $\mathrm{F}_0$, the right hand side follows trivially. Now suppose, however, that there is another set in $\mathcal{F}$, say the set $\mathrm{F}_1$, for which $\Pr(X = k; \theta, \mathrm{G}_0) = \Pr(T = k; \theta, \mathrm{F}_1)$. Then the left hand side condition implies

$$E(X; \theta, \mathrm{G}_0) = E(T; \theta, \mathrm{F}_1).$$

However, from the assumed AR-F it follows, using Equation 2.7, that also

$$E(X; \theta, \mathrm{G}_0) = E(T; \theta, \mathrm{F}_0),$$

so that

$$E(T; \theta, \mathrm{F}_1) = E(T; \theta, \mathrm{F}_0). \tag{2.8}$$

Because of the assumed unique correspondence property, in (2.8), the set of OCs leading to $E(T; \theta, \mathrm{F}_1)$ is the same set of OCs as the one leading to $E(T; \theta, \mathrm{F}_0)$. Consequently, $\mathrm{f}_j(\theta; \mathrm{F}_1) = \mathrm{f}_j(\theta; \mathrm{F}_0)$ for all $j$ so that $\mathrm{F}_1$ must be equal to $\mathrm{F}_0$, which establishes the AR-S between the sets G and F.

□

This lemma states the following. In general, if $X$ and $T$ are identically distributed, this does not, without additional assumptions, imply anything about the OCs or the covariance structure of the binary variables $Y_1, \ldots, Y_M$, as was demonstrated in Examples A and B and Lemma 1. However, if the unique correspondence property holds in either one of two additively related families $\mathcal{F}$ and $\mathcal{G}$, then distributional identity of $X$ and $T$ implies that the OCs in F are equal to the CCRFs in G.

**Corollary 1** *Consider $X, F, \mathcal{F}, \mathcal{U}_F, T, G, \mathcal{G}$ and $\mathcal{U}_G$ as in the preamble to Lemma 3. Then for all $F \in \mathcal{F}$ and $G \in \mathcal{G}$ it holds that*

$$E(X; \theta, G) = E(T; \theta, F) \quad \Rightarrow \quad f_j(\theta; F) = \sum_{k=j}^{M} g_k(\theta; G), \quad \textit{for } j = 1 \ldots M.$$

*Proof.* Equal to the proof of Lemma 3, because of the unique correspondence.

□

It is now possible to proceed with the last theorem.

**Theorem 3** *Consider $\mathcal{F}, \mathcal{G}, X$ and $T$ as in the preamble to Lemma 3 and again assume a unique correspondence in $\mathcal{F}$ and $\mathcal{G}$, and an AR-F between $\mathcal{F}$ and $\mathcal{G}$. Then for all $F \in \mathcal{F}$ and $G \in \mathcal{G}$ the following equivalence holds:*

$$\left.\begin{array}{c}\Pr(X = k; \theta, G) = \Pr(T = k; \theta, F) \\ \textit{for } k = 0 \ldots M\end{array}\right\} \Leftrightarrow$$

$$\left\{\begin{array}{ll}(a) & f_j(\theta; F) = \sum_{k=j}^{M} g_k(\theta; G) \textit{ for } j = 1 \ldots M; \\ (b) & \Pr(Y_i = 1 | Y_j = 1; \theta, F) = 1 \textit{ for } j = 2 \ldots M \textit{ and all } i < j.\end{array}\right.$$

*Proof.* Right to left: equal to the proof of Lemma 1.

Left to right: right hand side consequence (a) was established in Lemma 3. Using this result, and the left hand side condition $\Pr(X = k; \theta, \mathrm{G}) = \Pr(T = k; \theta, \mathrm{F})$, Theorem 2 can be applied, and it may be inferred that $\Pr(Y_i = 1 | Y_j = 1; \theta, \mathrm{F})$ must be equal to 1 for all $i < j$, which concludes the proof of the implication from left to right, because this is right hand side consequence (b).

Combining both parts of the proof completes the proof of the theorem.

□

Consequently, if two families $\mathcal{F}$ and $\mathcal{G}$ both defined by the unique correspondence property are AR-F related, the implication sign from Lemma 1 becomes a full equivalence.

The explanation for the fact that the equivalence obtained in Theorem 3 would not hold in Lemma 1, may be clarified as follows. In general, for a given set of TSFs, or an ETSF, the OCs are not uniquely determined (see Examples A and B). In Theorem 3 however, the TSFs belong to a family $\mathcal{F}$ defined by the unique correspondence property between ETSF and OCs. Therefore, in this family, any ETSF is uniquely tied to only a single set of OCs. Because of the AR-F and the left hand side condition, which implies $E(X;\theta,\mathrm{G}) = E(T;\theta,\mathrm{F})$, this is the set having an AR-S with G (see Corollary 1), so that we are now really in the conditions and assumptions not of Lemma 1 but of Theorem 2, in which also a full equivalence was obtained. So although, in the situation of Theorem 3, there are still no implication arrows between OCs and TSFs, there are other assumptions in this theorem which make it possible to infer something about the OCs from a condition on the TSFs.

## 2.3 Applications

The results obtained will now be investigated with respect to their consequences for three well-known parametric models for polytomous data: the graded response model, the sequential model and the partial credit model. Before starting this investigation, define a set F of M one parameter logistic (1-PL) functions as

$$\mathrm{h}_j(\theta;\boldsymbol{\lambda}) = \frac{\exp(\theta \Leftrightarrow \lambda_j)}{1+\exp(\theta \Leftrightarrow \lambda_j)}, \quad \text{for } j = 1,\ldots,\mathrm{M}. \tag{2.9}$$

In this formula $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_M)$ denotes the parameter for an entire set of M 1-PL curves. It may be noted that two 1-PL curves with different parameter values are nonintersecting. Again it will be assumed, for notational convenience, that $\mathrm{h}_0(\theta;\boldsymbol{\lambda}) \equiv 1$ and $\mathrm{h}_{M+1}(\theta;\boldsymbol{\lambda}) \equiv 0$.

### 2.3.1 Graded Response Model

The graded response model (GRM) developed by Samejima (1969) is a parametric model for responses on polytomous variables. The GRM parameter for a variable with maximum score M will be denoted by $\boldsymbol{\gamma} = (\gamma_1 \ldots \gamma_k \ldots \gamma_M)$. The elements of the parameter vector are ordered such that $\gamma_1 < \gamma_2 < \ldots < \gamma_M$. In the GRM the CRFs are given by

$$\Pr(X = k;\theta,\boldsymbol{\gamma}) = \mathrm{h}_k(\theta;\boldsymbol{\gamma}) \Leftrightarrow \mathrm{h}_{k+1}(\theta;\boldsymbol{\gamma}) \qquad \text{for } k = 0,\ldots,\mathrm{M}, \tag{2.10}$$

where $\mathrm{h}_k$ is the 1-PL function defined in (2.9). A variable whose distribution follows the GRM will be called a GRM variable, and similarly for the PCM and the SM below. If the score on a GRM variable is distributed as the total score on a set of Guttman dependent binary variables, then by Theorem 1 these binary variables have the following unique OCs:

$$\begin{aligned}
\Pr(Y_j = 1; \theta, \boldsymbol{\gamma}) &= \sum_{k=j}^{M} \Pr(X = k; \theta, \boldsymbol{\gamma}) \\
&= \sum_{k=j}^{M} [\mathrm{h}_k(\theta; \boldsymbol{\gamma}) \Leftrightarrow \mathrm{h}_{k+1}(\theta; \boldsymbol{\gamma})] \\
&= \mathrm{h}_j(\theta; \boldsymbol{\gamma}), \quad \text{for } j = 1 \ldots \mathrm{M},
\end{aligned}$$

which is a set of 1-PL OCs having $\boldsymbol{\lambda} = \boldsymbol{\gamma}$. On the other hand, if the score on a GRM variable with parameter vector $\boldsymbol{\gamma}$ is distributed as the total score on a set of 1-PL binary variables with parameter vector $\boldsymbol{\lambda} = \boldsymbol{\gamma}$, this set of binary variables must be Guttman dependent by Theorem 2.

Now let $\mathcal{G}$ be the family of sets of response curves for the GRM, each set $\mathrm{G} \in \mathcal{G}$ having a different parameter $\boldsymbol{\gamma}$, and let $\mathcal{F}$ be the family of sets of 1-PL OCs, each with a different parameter vector $\boldsymbol{\lambda}$. It may then be inferred from (2.10) that there exists an AR-S (Definition 3) between the sets G with parameter vector $\boldsymbol{\gamma}$ and F with parameter vector $\boldsymbol{\lambda} = \boldsymbol{\gamma}$. Furthermore, note, first, that all sets of curves in families $\mathcal{F}$ and $\mathcal{G}$ are unique, and second, that it is possible to find an additively related set $\mathrm{G} \in \mathcal{G}$ for every set $\mathrm{F} \in \mathcal{F}$, and vice versa. It then follows that there exists an AR-F (Definition 4) between the families $\mathcal{F}$ and $\mathcal{G}$. Chang and Mazzeo (1994) proved that in the GRM there is a unique correspondence between CRFs and ERF. Consequently, by Lemma 2 there also is a unique correspondence in family $\mathcal{F}$. Finally, therefore, applying Theorem 3, any set of binary 1-PL variables distributionally identical to a GRM variable with parameter vector $\boldsymbol{\gamma}$, has to be Guttman dependent and has to be additively related to it with parameter vector $\boldsymbol{\lambda} = \boldsymbol{\gamma}$. It is impossible for a GRM variable to be distributionally identical to a set of binary 1-PL variables with any other parameter vector, or with any other dependence structure. In particular, no set of independent binary 1-PL variables (i.e. no set of Rasch variables) can be distributionally identical to a GRM variable. Huynh (1994) derived a condition under which PCM variables are distributed as the total score on a set of independent binary 1-PL variables. It follows that no PCM variable satisfying Huynh's condition, is distributionally identical to a GRM variable.

### 2.3.2 Partial Credit Model

In the partial credit model (PCM) (Andrich, 1978; Masters, 1982) the response probabilities on a variable with maximum score M are given by

$$\Pr(X = k; \theta, \boldsymbol{\delta}) \propto \frac{1}{\mathrm{D}_M}\left[\exp\left(k\theta \Leftrightarrow \sum_{s=1}^{k}\delta_s\right)\right], \quad k = 0, 1, \ldots, \mathrm{M},$$

with the proportionality constant $\mathrm{D}_M$ being given by $\sum_{k=0}^{M}[\exp(k\theta \Leftrightarrow \sum_{s=1}^{k}\delta_s)]$, and $\sum_{s=1}^{0}(\Leftrightarrow\delta_s) \equiv 0$.

The marginal response probabilities for a set of Guttman dependent binary variables distributionally identical to a PCM variable are found upon application of Theorem 1:

$$\begin{aligned}\Pr(Y_j = 1; \theta, \boldsymbol{\delta}) &= \sum_{k=j}^{M}\Pr(X = k; \theta, \boldsymbol{\delta}) \\ &= \frac{1}{\mathrm{D}_M}\sum_{k=j}^{M}\left[\exp\left(k\theta \Leftrightarrow \sum_{s=1}^{k}\delta_s\right)\right], \quad j = 1, 2, \ldots, \mathrm{M}. \quad (2.11)\end{aligned}$$

As an example for M = 4, and using $\xi = \exp(\theta)$ and $\varepsilon_k = \exp(\Leftrightarrow\delta_k)$, this gives:

$$\begin{aligned}\Pr(Y_1 = 1; \theta, \boldsymbol{\delta}) &= (\xi\varepsilon_1 + \xi^2\varepsilon_1\varepsilon_2 + \xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4, \\ \Pr(Y_2 = 1; \theta, \boldsymbol{\delta}) &= (\xi^2\varepsilon_1\varepsilon_2 + \xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4, \\ \Pr(Y_3 = 1; \theta, \boldsymbol{\delta}) &= (\xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4, \\ \Pr(Y_4 = 1; \theta, \boldsymbol{\delta}) &= (\xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4.\end{aligned}$$

Next, any set of binary variables with OCs as in (2.11) and distributionally identical to a PCM variable with the same parameter vector has to be Guttman dependent by Theorem 2. Finally, because of the unique correspondence between

ERF and CRFs in the PCM, which was proved by Chang and Mazzeo (1994), again let $\mathcal{G}$ be the family of sets of CRFs in the PCM, and let $\mathcal{F}$ be the family of sets of OCs defined in (2.11). Applying Lemma 2 and Theorem 3, it can be concluded that no set of binary variables with OCs as in (2.11) other than a Guttman dependent set with parameter vector equal to $\boldsymbol{\delta}$, can be distributionally identical to a PCM variable with parameter vector $\boldsymbol{\delta}$.

It is interesting to note here that up to a certain point an early derivation of the PCM (Andrich, 1978) was along these same lines of thought. Andrich assumes 1-PL functions for the OCs of the two binary variables distributionally identical to a trinary rating scale variable, and starts from an 'imagined' local independence on these two binary variables. He then observes that logically a response pattern 01 on the two ordered binary variables should have zero probability and resolves this conflict by reweighing the probabilities of the other three response patterns such that they sum up to 1. Molenaar (1983) calls this 'post hoc conditioning on the Guttman property'. The parameter vector in this model will be denoted as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$. As an example, for M = 2, before the post hoc conditioning, i.e. still assuming local independence, the simultaneous response probabilities are given by

$$\begin{aligned}
\Pr(Y_1 = 0, Y_2 = 0; \theta, \boldsymbol{\beta}) &= 1/\mathrm{D} \\
\Pr(Y_1 = 1, Y_2 = 0; \theta, \boldsymbol{\beta}) &= \exp(\theta \Leftrightarrow \beta_1)/\mathrm{D} \\
\Pr(Y_1 = 0, Y_2 = 1; \theta, \boldsymbol{\beta}) &= \exp(\theta \Leftrightarrow \beta_2)/\mathrm{D} \\
\Pr(Y_1 = 1, Y_2 = 1; \theta, \boldsymbol{\beta}) &= \exp(2\theta \Leftrightarrow \beta_1 \Leftrightarrow \beta_2)/\mathrm{D}
\end{aligned}$$

in which $\mathrm{D} = 1 + \exp(\theta \Leftrightarrow \beta_1) + \exp(\theta \Leftrightarrow \beta_2) + \exp(2\theta \Leftrightarrow \beta_1 \Leftrightarrow \beta_2)$. After the conditioning they become

$$\begin{aligned}
\Pr{}^*(Y_1 = 0, Y_2 = 0; \theta, \boldsymbol{\beta}) &= 1/\mathrm{D}^* \\
\Pr{}^*(Y_1 = 1, Y_2 = 0; \theta, \boldsymbol{\beta}) &= \exp(\theta \Leftrightarrow \beta_1)/\mathrm{D}^* \\
\Pr{}^*(Y_1 = 1, Y_2 = 1; \theta, \boldsymbol{\beta}) &= \exp(2\theta \Leftrightarrow \beta_1 \Leftrightarrow \beta_2)/\mathrm{D}^*
\end{aligned} \tag{2.12}$$

with $\mathrm{D}^* = 1 + \exp(\theta \Leftrightarrow \beta_1) + \exp(2\theta \Leftrightarrow \beta_1 \Leftrightarrow \beta_2)$. A similar reasoning is applied to polytomous rating scale variables with M > 2. As a result, only perfect response patterns will appear. Using the starred probabilities as response probabilities for a polytomous variable, essentially the CRFs of the partial credit model (Masters, 1982) have been obtained, although Andrich applies a notation more suited to the needs of a rating scale model. Note however that because of the conditioning the (starred) response probabilities of the PCM variable with parameters $(\beta_1, \beta_2)$, given in (2.12) do not coincide with the TSFs of a set of Guttman dependent binary 1-PL variables with parameters $(\beta_1, \beta_2)$, nor do they coincide with the TSFs of a set of independent binary Rasch variables with parameters $(\beta_1, \beta_2)$. They may, however, coincide with the TSFs of a set of two independent binary 1-PL variables with parameters other than $(\beta_1, \beta_2)$ (see Huynh, 1994).

### 2.3.3 Sequential Model

Molenaar (1983) has pointed out the possibility of a so-called conditional model for polytomous responses. This model has subsequently become known as the sequential model (SM) and it has been worked out by Tutz (1990, 1997) and by Verhelst, Glas and de Vries (1997). Let the parameter vector for the polytomous variable with maximal score M in the SM be $\boldsymbol{\sigma} = (\sigma_1 \ldots \sigma_k \ldots \sigma_M)$. The $\sigma_k$'s need not be ordered. The response probabilities in this model are given by

$$\Pr(X = k; \theta, \boldsymbol{\sigma}) = \prod_{r=1}^{k} \mathrm{h}_r(\theta; \boldsymbol{\sigma}) \Leftrightarrow \prod_{r=1}^{k+1} \mathrm{h}_r(\theta; \boldsymbol{\sigma}) \quad \text{for } k = 0 \ldots \mathrm{M}, \tag{2.13}$$

where again $\mathrm{h}_k$ is the 1-PL function and for notational convenience $\prod_{r=1}^{0} \mathrm{h}_r(\theta; \boldsymbol{\sigma}) \equiv 1$ and $\prod_{r=1}^{M+1} \mathrm{h}_r(\theta; \boldsymbol{\sigma}) \equiv 0$.

Consider the score on an SM variable with parameter vector $\boldsymbol{\sigma}$. If this score is distributed as the total score on a set of Guttman dependent binary variables, then by Theorem 1, the OCs of these binary variables are given by

$$\begin{aligned}
\Pr(Y_j = 1; \theta, \boldsymbol{\sigma}) &= \sum_{k=j}^{M} \Pr(X = k; \theta, \boldsymbol{\sigma}) \\
&= \sum_{k=j}^{M} \left[ \prod_{r=1}^{k} \mathrm{h}_r(\theta; \boldsymbol{\sigma}) \Leftrightarrow \prod_{r=1}^{k+1} \mathrm{h}_r(\theta; \boldsymbol{\sigma}) \right] \\
&= \prod_{r=1}^{j} \mathrm{h}_r(\theta; \boldsymbol{\sigma}), \quad \text{for } j = 1, 2, \ldots, \mathrm{M}.
\end{aligned} \tag{2.14}$$

On the other hand, if the total score on a set of binary variables with OCs as in (2.14) is distributed as the score on a polytomous SM variable with the same parameter vector, then by Theorem 2, these binary variables must be Guttman dependent. Finally, let $\mathcal{G}$ be the family of sets G of CRFs for the SM, and let $\mathcal{F}$ be the family of sets F of OCs as in (2.14). Note that there is an AR-F relation between the families $\mathcal{F}$ and $\mathcal{G}$. If it can be shown that in the SM there is a unique correspondence between ERF and CRFs, Lemma 2 and Theorem 3 can be applied, and it may be inferred that any set of binary variables having OCs as in (2.14) and distributionally identical to a polytomous SM variable with parameter vector $\boldsymbol{\sigma}$, has to be a Guttman dependent set, with parameter vector equal to $\boldsymbol{\sigma}$ as well. The proof of unique correspondence is given in appendix 2.4; it is modeled after the proofs given by Chang and Mazzeo (1994).

The sequential model was expressly formulated to deal with Guttman dependent binary 'subtasks' or 'steps'. It might be argued that, for modeling Guttman dependent item steps, the conditional probabilities $\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta)$ are more relevant than the marginal probabilities $\Pr(Y_j = 1; \theta)$. In the SM these conditional probabilities have a particularly simple form:

$$\begin{aligned}
\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta, \boldsymbol{\sigma}) &= \frac{\Pr(Y_j = 1 \text{ and } Y_{j-1} = 1; \theta, \boldsymbol{\sigma})}{\Pr(Y_{j-1} = 1; \theta, \boldsymbol{\sigma})} \\
&= \frac{\Pr(Y_j = 1; \theta, \boldsymbol{\sigma})}{\Pr(Y_{j-1} = 1; \theta, \boldsymbol{\sigma})} \\
&= \mathrm{h}_j(\theta; \boldsymbol{\sigma}), \qquad (2.15)
\end{aligned}$$

which only depends on $\sigma_j$. The second line is a direct consequence of the assumed Guttman property. In both the GRM and the PCM these conditional probabilities are much more complex, and in these models each conditional probability depends on several parameters. Therefore, although theoretically for every polytomous variable the distribution of its score is identical to the distribution of the total score on a set of Guttman dependent binary variables, the SM probably fits the Guttman assumption best.

### 2.3.4 Generalization

Samejima (1969) and Muraki (1992) present generalized versions of the GRM and the PCM, respectively. In these generalized models every polytomous variable $j$ has an item discrimination parameter $\alpha_j$. A similar generalization can be formulated for the SM. The proofs of unique correspondence between ERF and CRFs for the GRM and PCM, given by Chang and Mazzeo (1994), and for the SM, given in appendix 2.4 of this paper, extend to this case. Hence all results obtained in Section 2.3 can be extended to the generalized forms of PCM, GRM and SM.

As an example, let $\mathcal{G}^*$ be the family of sets of CRFs for the generalized GRM, with parameter vector $\boldsymbol{\gamma}^* = (\alpha, \gamma_1, \ldots \gamma_k, \ldots, \gamma_N)$, and let $\mathcal{F}^*$ be the family of sets of 1-PL curves with a common discrimination parameter per set, with parameter vector $\boldsymbol{\lambda}^* = (\alpha, \lambda_1, \ldots, \lambda_j, \ldots, \lambda_M)$. Note that $\mathcal{F}^*$ is not the family of sets of two parameter logistic or 2PL items (Birnbaum, 1968). Then the only set of OCs from family $\mathcal{F}^*$ that can coincide with a set of CRFs from family $\mathcal{G}^*$ will be a Guttman dependent set having $\boldsymbol{\lambda}^* = \boldsymbol{\gamma}^*$.

## 2.4 Conclusion and discussion

The score on a polytomous item is sometimes assumed to come about by means of a process of solving a number of binary subtasks. In that case, the score on the polytomous item is equal to the total score obtained on the binary subtasks. This perspective points to two interesting routes of investigation. On the one hand, if this perspective is taken, it may be possible to find a suitable model for the score on the polytomous item from the assumptions made on the subtasks. This is done by Van Engelenburg (1997). On the other hand, starting from this perspective it is also possible to examine whether there are mathematical relations between models for binary and models for polytomous variables, that either permit or forbid such an interpretation.

It was shown in Theorem 1 that for every polytomous item a set of OCs can be found, such that if T is the total score on a set of Guttman dependent binary variables (subtasks) having these OCs, then the score on the polytomous item and T are identically distributed. Therefore the score on any polytomous item is distributionally identical to the total score on a set of Guttman dependent binary variables (subtasks).

However, this does not mean that there is a substantive equivalence between the polytomous item and an empirical set of Guttman dependent binary items or subtasks: the bare fact that two sets of curves coincide, does not, in general, imply anything about the cognitive processes involved in getting at a particular polytomous response. This is clearly demonstrated by the fact that, although all polytomous items are distributionally identical to a set of Guttman dependent binary variables, some of these items are also distributionally identical to a set of independent binary variables, as can be seen in examples A and B in Section 2.1.1 and in the work of Huynh (1994). It would be interesting to investigate this matter further, and in particular to find out whether any relationships can be established with substantive item contents (Wilson, 1988; Rosenbaum, 1988).

It was observed that the score on a GRM item is never distributed as the total score on a set of independent binary 1-PL or Rasch variables. This contrasts the GRM with the PCM (see Huynh, 1994). It is also interesting to compare this result with Jansen and Roskam (1986), who note that when the score on a GRM variable is dichotomized, the GRM for M = 1 results, which is the Rasch model.

A final question which presents itself is the following. Suppose one has a set of Guttman dependent binary variables, for example the scores obtained on a set of binary items under a sequential design. Let $T$ be the total score on this set of binary variables. Furthermore, let $X$ be a polytomous item response. As it has been shown that all $X$ are distributionally identical to the total score on a set of Guttman dependent binary variables, one might feel tempted to apply just any model for polytomous item responses to $T$. However, this conclusion may not be justified. There might be other requirements, besides distributional identity, that should be satisfied if one wants to apply a model for polytomous item responses to

$T$. This question can probably best be investigated upon a closer inspection not of the marginal probabilities $\Pr(Y_j = 1; \theta)$, but of the conditional probabilities $\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta)$. As was already noted in Equation 2.15, the SM is well suited to model maximal covariances. For the GRM and the PCM this question will be addressed in a separate paper.

## Appendix

Proof of unique correspondence in the SM

The proof will be given for a generalized sequential model, i.e. for a model having a discrimination parameter $\alpha_j$ for each polytomous variable $j$.

Let there be two SM variables. The first variable has maximum score M, and parameter vector $\boldsymbol{\sigma} = (\alpha, \sigma_1, \ldots, \sigma_M)$; the second variable has maximum score N, and parameter vector $\boldsymbol{\tau} = (\beta, \tau_1, \ldots, \tau_N)$. Here $\alpha$ and $\beta$ are discrimination parameters which are constant within each SM variable, but may vary over variables. Let, in this appendix, $\mathrm{h}(\theta; \alpha, \delta)$ denote the 2PL function $\exp[\alpha(\theta \Leftrightarrow \delta)]/\{1 + \exp[\alpha(\theta \Leftrightarrow \delta)]\}$.

From (2.13) it can be derived that the ERF of an SM variable with parameter vector $\boldsymbol{\sigma} = (\alpha, \sigma_1, \ldots, \sigma_M)$ is given by

$$E(X; \theta, \boldsymbol{\sigma}) = \sum_{k=1}^{M} \prod_{r=1}^{k} \mathrm{h}(\theta; \alpha, \sigma_r).$$

Hence there is equality of ERFs for the two variables if

$$\sum_{k=1}^{M} \prod_{r=1}^{k} \mathrm{h}(\theta; \alpha, \sigma_r). = \sum_{j=1}^{N} \prod_{p=1}^{j} \mathrm{h}(\theta; \beta, \tau_p). \tag{2.16}$$

If we let $\mathrm{x}^{\alpha} = \exp(\alpha\theta)$, $\mathrm{x}^{\beta} = \exp(\beta\theta)$, $\mathrm{s}_r = \exp(\Leftrightarrow\alpha\sigma_r)$ and $\mathrm{t}_p = \exp(\Leftrightarrow\beta\tau_p)$, it is possible to write

$$\sum_{k=1}^{M}\prod_{r=1}^{k} h(\theta;\alpha,\sigma_r)$$

$$= \frac{s_1x^{\alpha}}{1+s_1x^{\alpha}} + \frac{s_1s_2x^{2\alpha}}{(1+s_1x^{\alpha})(1+s_2x^{\alpha})} + \ldots + \frac{s_1s_2\ldots s_Mx^{M\alpha}}{(1+s_1x^{\alpha})\ldots(1+s_Mx^{\alpha})}$$

$$= \frac{1}{(1+s_1x^{\alpha})(1+s_2x^{\alpha})\ldots(1+s_Mx^{\alpha})}\{s_1x^{\alpha}\left[(1+s_2x^{\alpha})(1+s_3x^{\alpha})\ldots(1+s_Mx^{\alpha})\right]$$

$$+s_1s_2x^{2\alpha}\left[(1+s_3x^{\alpha})\ldots(1+s_Mx^{\alpha})\right] + \quad\ldots\quad + s_1s_2\ldots s_Mx^{M\alpha}\Big\}$$

$$= \frac{1}{1+c_1x^{\alpha}+c_2x^{2\alpha}+\ldots+c_Mx^{M\alpha}}\Big\{s_1x^{\alpha}[1+c_{11}x^{\alpha}+c_{21}x^{2\alpha}+\ldots+c_{M-1,1}x^{(M-1)\alpha}]$$

$$+s_1s_2x^{2\alpha}[1+c_{12}x^{\alpha}+\ldots+c_{M-2,2}x^{(M-2)\alpha}] + \quad\ldots\quad + s_1s_2\ldots s_Mx^{M\alpha}[c_{M-M,M}]\Big\}.$$

In this last expression $c_1\ldots c_M$ are elementary symmetric functions of the vector $\mathbf{s} = (s_1,\ldots,s_M)$, and $c_{rk}$ is the $r$th symmetric function of $(s_{k+1}, s_{k+2},\ldots,s_M)$.

The ERF can be rewritten as

$$\sum_{k=1}^{M}\prod_{r=1}^{k} h(\theta;\alpha,\sigma_r) = \frac{C_1x^{\alpha}+C_2x^{2\alpha}+\ldots+C_Mx^{M\alpha}}{1+c_1x^{\alpha}+c_2x^{2\alpha}+\ldots+c_Mx^{M\alpha}}, \tag{2.17}$$

where

$$\begin{aligned}
&C_1 = s_1\\
&C_2 = s_1c_{11} + s_1s_2\\
&C_3 = s_1c_{21} + s_1s_2c_{12} + s_1s_2s_3\\
&C_4 = s_1c_{31} + s_1s_2c_{22} + s_1s_2s_3c_{13} + s_1s_2s_3s_4\\
&\vdots\\
&C_M = s_1c_{M-1,1} + s_1s_2c_{M-2,2} + \ldots + s_1s_2\ldots s_M,
\end{aligned}$$

or, alternatively,

$$C_k = \sum_{r=1}^{k}\left[\left(\prod_{v=1}^{r} s_v\right)(c_{k-r,r})\right],$$

where $c_{0r} \equiv 1$. It may be verified that $C_k > 0$ for all $k$, and in particular for $k = 1$, as $C_1 = s_1 = \exp(\Leftrightarrow\alpha\sigma_1) > 0$. Similarly, with $d_1\ldots d_N$ being the elementary symmetric functions of $\mathbf{t} = (t_1,\ldots,t_N)$, $d_{pj}$ the $p$th symmetric function of

$(t_{j+1}, t_{j+2}, \ldots, t_N)$, and

$$D_j = \sum_{p=1}^{j} \left[ \left( \prod_{w=1}^{p} t_w \right) (d_{j-p,p}) \right],$$

the ERF of the second variable can be written as

$$\sum_{j=1}^{N} \prod_{p=1}^{j} h(\theta; \beta, \tau_p) = \frac{D_1 x^{\beta} + D_2 x^{2\beta} + \ldots + D_N x^{N\beta}}{1 + d_1 x^{\beta} + d_2 x^{2\beta} + \ldots + d_N x^{N\beta}}. \tag{2.18}$$

Because of the equality of the ERFs, the expressions in equations (2.17) and (2.18) are equal. Hence

$$\frac{C_1 x^{\alpha} + C_2 x^{2\alpha} + \ldots + C_M x^{M\alpha}}{1 + c_1 x^{\alpha} + c_2 x^{2\alpha} + \ldots + c_M x^{M\alpha}} = \frac{D_1 x^{\beta} + D_2 x^{2\beta} + \ldots + D_N x^{N\beta}}{1 + d_1 x^{\beta} + d_2 x^{2\beta} + \ldots + d_N x^{N\beta}},$$

and therefore, multiplying the numerators with the denominators,

$$\begin{aligned}
& C_1 x^{\alpha} + C_1 d_1 x^{\alpha+\beta} + C_1 d_2 x^{\alpha+2\beta} + \ldots + C_1 d_N x^{\alpha+N\beta} \\
& +C_2 x^{2\alpha} + C_2 d_1 x^{2\alpha+\beta} + \ldots + C_2 d_N x^{2\alpha+N\beta} + \\
& \vdots \\
& +C_M x^{M\alpha} + C_M d_1 x^{M\alpha+\beta} \ldots + C_M d_N x^{M\alpha+N\beta} \\
= \; & D_1 x^{\beta} + D_1 c_1 x^{\beta+\alpha} + D_1 c_2 x^{\beta+2\alpha} + \ldots + D_1 c_M x^{\beta+M\alpha} \\
& +D_2 x^{2\beta} + D_2 c_1 x^{2\beta+\alpha} + \ldots + D_2 c_M x^{2\beta+M\alpha} + \\
& \vdots \\
& +D_N x^{N\beta} + D_N c_1 x^{N\beta+\alpha} \ldots + D_N c_M x^{N\beta+M\alpha}.
\end{aligned} \tag{2.19}$$

Assuming that $\alpha < \beta$, of all these terms $C_1 x^{\alpha}$ is the one with the smallest exponent. In order for the equality to hold for all x, therefore, $C_1$ should be 0 because all $x > 0$. This however would contradict $C_1 = s_1 = \exp(\Leftrightarrow\alpha\sigma_1) > 0$.

Hence it must be concluded that it is impossible to have $\alpha < \beta$. By a similar reasoning it can be demonstrated that it is impossible to have $\alpha > \beta$ as well, and hence $\alpha$ has to be equal to $\beta$. The equality from (2.19) can now be rewritten as

$$\begin{aligned}
& \mathrm{C}_1\mathrm{x}^{\alpha} + \mathrm{C}_1\mathrm{d}_1\mathrm{x}^{2\alpha} + \mathrm{C}_1\mathrm{d}_2\mathrm{x}^{3\alpha} + \ldots + \mathrm{C}_1\mathrm{d}_N\mathrm{x}^{(1+N)\alpha} \\
& \quad +\mathrm{C}_2\mathrm{x}^{2\alpha} + \mathrm{C}_2\mathrm{d}_1\mathrm{x}^{3\alpha} + \ldots + \mathrm{C}_2\mathrm{d}_N\mathrm{x}^{(2+N)\alpha} + \\
& \quad \vdots \\
& \quad +\mathrm{C}_M\mathrm{x}^{M\alpha} + \mathrm{C}_M\mathrm{d}_1\mathrm{x}^{(M+1)\alpha} \ldots + \mathrm{C}_M\mathrm{d}_N\mathrm{x}^{(M+N)\alpha} \\
= {} & \quad \mathrm{D}_1\mathrm{x}^{\alpha} + \mathrm{D}_1\mathrm{c}_1\mathrm{x}^{2\alpha} + \mathrm{D}_1\mathrm{c}_2\mathrm{x}^{3\alpha} + \ldots + \mathrm{D}_1\mathrm{c}_M\mathrm{x}^{(1+M)\alpha} \\
& \quad +\mathrm{D}_2\mathrm{x}^{2\alpha} + \mathrm{D}_2\mathrm{c}_1\mathrm{x}^{3\alpha} + \ldots + \mathrm{D}_2\mathrm{c}_M\mathrm{x}^{(2+M)\alpha} + \\
& \quad \vdots \\
& \quad +\mathrm{D}_N\mathrm{x}^{N\alpha} + D_N\mathrm{c}_1\mathrm{x}^{(N+1)\alpha} \ldots + \mathrm{D}_N\mathrm{c}_M\mathrm{x}^{(N+M)\alpha}. \qquad (2.20)
\end{aligned}$$

In order for this equality to hold, $(\mathrm{C}_1 \Leftrightarrow \mathrm{D}_1)\mathrm{x}^{\alpha}$ has to be equal to 0 for all $x$, which is only achieved if $\mathrm{C}_1 = \mathrm{D}_1$, or alternatively if $\mathrm{s}_1 = \mathrm{t}_1$, and hence if $\sigma_1 = \tau_1$, as it was already shown that $\alpha = \beta$. Now however both $\sigma_1 = \tau_1$ and $\alpha = \beta$, whence $\mathrm{h}(\theta; \alpha, \sigma_1) = \mathrm{h}(\theta; \beta, \tau_1)$ for all $\theta$. Dividing both sides of (2.16) by this common factor will yield another equation to solve:

$$1 + \sum_{k=2}^{M} \prod_{r=2}^{k} \mathrm{h}(\theta; \alpha, \sigma_r) = 1 + \sum_{j=2}^{N} \prod_{p=2}^{j} \mathrm{h}(\theta; \alpha, \tau_p).$$

Similarly as before, it will be found that $\sigma_2 = \tau_2$; by mathematical induction this then holds for all other values of the two parameter vectors as well; in the end yielding $M = N$, $\alpha = \beta$ and $\sigma_k = \tau_k$ for all $k$.

# Chapter 3

# Modeling sequentially scored item responses

### Abstract

The sequential model was developed for describing the score resulting from a sequential process. In this paper the appropriateness of the partial credit model and the graded response model for sequential scoring is investigated. Mathematical reasons are given for the inapplicability of these models to sequentially scored variables.

Key words: Sequential scoring, sequential model, partial credit model, graded response model.

## 3.1 Introduction

The number of successes before the occurrence of the first failure in a sequence of M Bernoulli trials, $X$, follows a truncated geometric distribution with parameters $\pi$ and M:

$$\Pr(X = k; \pi, \mathrm{M}) = \begin{cases} \pi^k(1 \Leftrightarrow \pi) & \text{for} \quad k = 0, 1, \ldots, \mathrm{M} \Leftrightarrow 1; \\ \pi^M & \text{for} \quad k = \mathrm{M}. \end{cases}$$

The random experiment considered in this paper consists of two generalizations of the above: the parameter $\pi$ is allowed to vary over trials, and each $\pi_j$ is itself a function of a variable $\theta$, so that the experiment has parameter vector

$\boldsymbol{\pi}(\theta) = [\pi_1(\theta), \dots, \pi_j(\theta), \dots, \pi_M(\theta)]$ instead of parameter $\pi$. $X$ then follows a generalized truncated geometric distribution:

$$\Pr[X = k; \boldsymbol{\pi}(\theta), \mathrm{M}] = \begin{cases} \left[\prod_{j=1}^{k} \pi_j(\theta)\right] [1 \Leftrightarrow \pi_{k+1}(\theta)] & \text{for} \quad k = 0, 1, \dots, \mathrm{M} \Leftrightarrow 1; \\ \prod_{j=1}^{M} \pi_j(\theta) & \text{for} \quad k = \mathrm{M}. \end{cases}$$

**Definition 1** *Counting, in a sequence of M Bernoulli trials, the number of successes until the occurrence of the first failure, will be called sequential scoring. The trials are assumed to be performed in a fixed order.*

The substantive field to which this framework will be applied is that of item response theory (IRT). In this field (a) the trials are trials to solve a binary item, (b) the trials are sometimes referred to as steps, (c) the result of a trial is called a response, (d) $\theta$ is a latent ability, (e) the functions $\pi_j(\theta)$ are called item characteristic curves (ICCs), (f) the variable $X$ is known as the score, and (g) the functions $\Pr(X = k; \boldsymbol{\pi}(\theta), \mathrm{M})$ are commonly written as $\Pr(X = k; \theta)$. In this paper the functions $\Pr(X = k; \theta)$ will be called score functions. Furthermore, the binary variables $Y_1 \dots Y_j \dots Y_M$ are used for the results on the trials, where a 1 denotes success, and a 0 failure.

In accordance with common usage in IRT the ICCs will be denoted as $\mathrm{f}_j(\theta)$ instead of $\pi_j(\theta)$, so that

$$\mathrm{f}_j(\theta) \equiv \pi_j(\theta) = \Pr_B(Y_j = 1; \theta), \qquad (3.1)$$

where the subscript B is meant to indicate that, for each fixed value of $\theta$, the function $\mathrm{f}_j(\theta)$ is considered as the parameter in a Bernoulli trial. The need for the subscript B will become clear in Section 3.2.

Throughout it will be assumed, for notational convenience, that $\mathrm{f}_0(\theta) \equiv 1$ and $\mathrm{f}_{M+1}(\theta) \equiv 0$. With this convention it is possible to reformulate the expression for the distribution of the number correct score in a process of sequential scoring:

$$\Pr(X = k; \theta) = \left[\prod_{j=0}^{k} \mathrm{f}_j(\theta)\right] [1 \Leftrightarrow \mathrm{f}_{k+1}(\theta)] \quad \text{for } k = 0, 1, \dots, \mathrm{M}. \qquad (3.2)$$

If trial $k$ results in a failure, the results of all trials $k + j$, for $j \geq 1$, remain unobserved. Consequently there is dependence in the observed scores.

Sequential scoring in IRT may occur in two situations:

**1** In the first situation a set of binary items is involved. These items are tried in a fixed order, and the score $X$ is the number of correct responses until the first failure. It is immaterial whether all items are tried by the examinee, or whether a stop-if-fail presentation procedure is followed. It is the scoring rule (count the number of correct responses until the first failure) which makes the process sequential. This kind of sequential scoring might be encountered in the testing of psycho-motor skills.

**2** In the second situation a more or less complex problem is presented as a polytomous item consisting of a number of subtasks. These subtasks are called steps, or item steps. A judge evaluates the result of each step, in a fixed order. The score $X$ consists of the number of correct responses on the subtasks until the occurrence of the first failure. The position is taken here that in this second situation too, it is the scoring rule, and not substantive item contents or some mental or cognitive process, that makes the process sequential.

In both situations, therefore, a set of binary variables is subjected to sequential scoring. The two situations will be treated as equivalent. In Section 3.4 the reason for the distinction will become clear.

The prominence of the scoring rule, in the second situation, is demonstrated in the following example, which is originally due to Masters (1982):

$$\sqrt{7.5/0.3 \Leftrightarrow 16} = ?$$

In order to gain full credit for this item, three calculations have to be performed: first $7.5/0.3 = 25$ has to be calculated, then $25 \Leftrightarrow 16 = 9$ should be found, and finally $\sqrt{9} = 3$ has to be obtained. If the item is scored sequentially, 1 point would be earned by finding 25, another by finding both 25 and 9, and the maximum score of 3 is gained only if also the last step is carried out correctly. Now suppose, however, that the first fraction were incorrectly calculated as $7.5/0.3 = 20$; but starting from this incorrect number the next two steps are carried through correctly, i.e. both $20 \Leftrightarrow 16 = 4$ and $\sqrt{4} = 2$ are obtained. If sequential scoring were in effect, a score of 0 would be obtained. On the other hand, it would also be possible to give a credit of 2 points for the 2 steps that have been correctly performed, given the failure on the first step. This however would not be sequential scoring, and the score on the very same polytomous item, scored in this different way, would not be the result of a sequential process. Hence, the scoring rule is the decisive factor in declaring a process to be sequential.

The interest in this paper is in one sequentially scored polytomous item, or in one set of sequentially scored binary items only. The obvious way to model sequential scoring in IRT would be to assume a suitable function for the ICCs and then to derive the functional form for the distribution of the score. Samejima (1972) and Molenaar (1983) have pointed out the possibility of a 1-parameter logistic curve for the ICCs in this situation. This model has been worked out by Tutz (1990; 1997), and it has become known as the sequential model (SM). The estimation of the parameters in this model has been studied by Verhelst, Glas and de Vries (1997). One might wonder, however, whether other models for polytomous item responses, such as the graded response model (GRM; Samejima,1969) or the partial credit model (PCM: Masters, 1982; Andrich, 1978; Andersen, 1977), could not be used equally well with sequential scoring. It can for example be shown (see Section 3.2) that for each polytomous item response model a set of functions can be derived such that, if these functions were ICCs in a sequential process, the score probabilities would be described by the original model. Furthermore, both the PCM and the GRM have been presented as models for sequential processes (Masters, 1982; Samejima, 1972, 1995 ).

De Vries (1988) developed an algorithm to find a set of PCM curves, given a set of SM curves, such that the area between the two sets of curves is minimized. The results of a study with this algorithm showed that response curves under these two models can be very close. Furthermore, Verhelst, Glas and de Vries (1997) found a comparable fit when they applied either the SM or the PCM to the same data set. This might lead one to the conclusion that using the PCM with sequential scoring would not necessarily lead to severe problems. On the other hand, Molenaar (1983) already noted conceptual difficulties with the PCM as a model for sequential scoring. Considering the probability of observing a score $k$ rather than $k+1$, i.e. the probability $\Pr(X = k \mid X = k \text{ or } X = k+1; \theta)$, Molenaar observed that, although under this model these odds depend on only one parameter, the ICCs for step $k$ must depend on several additional parameters. The relation between these parameters and the ICCs will be established in Section 3.4. Andrich (1995) too argues that the PCM is conceptually unsuited for sequential scoring, because under this model each probability of obtaining a particular score depends on all parameters.

A possible argument against the use of the GRM with sequential scoring is that this model describes a different process: the GRM emerges upon assuming a 1-parameter logistic curve for the process of giving one overall judgement or grade (see e.g. Molenaar, 1983; Mellenbergh, 1995; Van Engelenburg, 1997). This will be called a graded scoring. However, Tutz (1997) observed that if the extreme value distribution function with parameter value $\delta_j$ is chosen for the ICCs in a sequential process, that is, if the ICCs are given by

$$\mathrm{f}_j(\theta) = \exp\{\Leftrightarrow \exp[\Leftrightarrow(\theta \Leftrightarrow \delta_j)]\}, \tag{3.3}$$

then the generalized truncated geometric distribution induced by this choice is given by

$$\begin{aligned}
\Pr(X = k; \theta) &= \left[\prod_{j=0}^{k} \mathrm{f}_j(\theta)\right] [1 \Leftrightarrow \mathrm{f}_{k+1}(\theta)] \\
&= \left[\prod_{j=0}^{k} \mathrm{f}_j(\theta)\right] \Leftrightarrow \left[\prod_{j=0}^{k+1} \mathrm{f}_j(\theta)\right] \\
&= \mathrm{f}'_k(\theta) \Leftrightarrow \mathrm{f}'_{k+1}(\theta), \qquad \text{for } k = 0, \ldots, \mathrm{M},
\end{aligned} \tag{3.4}$$

where the functions $\mathrm{f}'_k(\theta)$ are again extreme value functions but with parameter values $\delta'_k = \ln[\sum_{j=1}^{k} \exp(\delta_j)]$. Equation 3.4 however is an expression for the GRM with the usual logistic or normal ogive functions replaced by extreme value functions. This model will henceforth be referred to as extreme value GRM, to distinguish it from the GRM with logistic functions, which will be called the logistic GRM. If there is any danger of confusion, the SM will also be given a suffix of logistic or extreme value, to identify the nature of the ICC employed. The extreme value GRM, or, equivalently, the extreme value SM, would be suitable for sequential *and* for graded scoring. So before a model is declared unsuitable to describe a particular kind of process, a careful examination seems justified.

The purpose of the present paper is to investigate what happens if either the PCM or the (logistic) GRM is applied to data which are known to have been generated by a sequential process. It will be shown that the use of these models leads to fundamental problems. The investigation will be carried out by considering the consequences of inserting or removing trials into or from the sequence of Bernoulli trials, without changing any of the other trials.

With polytomous items, it is impossible to remove or insert a trial (step or subtask) somewhere in the middle of the item, without also changing other item steps. Changing $\sqrt{(7.5/0.3 \Leftrightarrow 16)}$ into $\sqrt{(7.5/0.3 \Leftrightarrow 16) \times 4}$ would not only add a new step between the second and the third ones, but it would also change the last step from $\sqrt{9}$ into $\sqrt{36}$. Consequently, if the sequential process under consideration is the sequential scoring of a polytomous item, and if the existing item steps should remain unchanged, subtasks can only be removed from or added to the beginning or the end of the polytomous item. For example, if the item mentioned above were changed into

$$\sqrt{\frac{3^2 \Leftrightarrow 1.5}{0.3} \Leftrightarrow 16} \quad \text{and then to} \quad \sqrt{\frac{(1+2)^2 \Leftrightarrow 1.5}{0.3} \Leftrightarrow 16}.$$

two steps would have been added in front, but the new steps 3,4 and 5 are still the same ones as the original steps 1,2, and 3.

Removing a particular item or item step should not be confused with the joining of categories, which was studied by Jansen and Roskam (1986) and Andrich (1995). In this paper, no categories are joined, but an item or item step is altogether removed. However, there is a great similarity between the objective of this study and the objective of Jansen and Roskam. These authors specify a criterion which models for rating scales should satisfy, and they investigate models with respect to their criterion. In the present paper a criterion will be formulated which models for sequential scoring should satisfy, and several models are investigated with respect to this criterion.

The outline of the paper is as follows. First, in Section 3.2, some equalities are derived which will be of great help later on. In Section 3.3 a mathematical requirement for models describing sequential scoring is formulated. The question whether this requirement is satisfied under the SM, the GRM and the PCM is investigated in Section 3.4.

## 3.2 Preliminaries

In this section some equalities are derived which will facilitate the exposition in Section 3.4. Recall that the variables $Y_j$, $j = 1 \ldots$ M, are used for the results on the Bernoulli trials. These variables will be collected in the vector $\boldsymbol{Y} = (Y_1 \ldots Y_M)$. As an example, for M $= 4$ the variable $\boldsymbol{Y}$ can have the following values: $0ccc, 10cc, 110c, 1110, 1111$, where $c$ is a symbol to denote 'unobserved'. It can be easily verified that the distribution of $\boldsymbol{Y}$ is given by

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}; \theta) = \Pr[X = t(\boldsymbol{Y}); \theta],$$

where $t(\boldsymbol{Y})$ is the number of correct responses in score pattern $\boldsymbol{Y}$. The marginal probability of $Y_j$ in a sequential process is obtained through the distribution of $\boldsymbol{Y}$:

$$\begin{aligned}\Pr{}_{\mathrm{m}}(Y_j = 1; \theta) &= \sum_{\boldsymbol{Y}:Y_j=1} \Pr(\boldsymbol{Y} = \boldsymbol{y}; \theta) \\ &= \sum_{k=j}^{M} \Pr(X = k; \theta).\end{aligned}$$

The second equality follows from the sequential scoring rule. The subscript m serves to explicitly distinguish this marginal probability from the ICC, which is also a probability of a correct response on item $j$, but which has been given the subscript B for Bernoulli in Equation 3.1.

The following two equations are obviously equivalent:

$$\mathrm{Pr}_{\mathrm{m}}(Y_j = 1; \theta) \;=\; \sum_{k=j}^{M} \mathrm{Pr}(X = k; \theta), \qquad j = 1 \ldots \mathrm{M}; \tag{3.5a}$$

$$\mathrm{Pr}(X = k; \theta) \;=\; \mathrm{Pr}_{\mathrm{m}}(Y_k = 1; \theta) \Leftrightarrow \mathrm{Pr}_{\mathrm{m}}(Y_{k+1} = 1; \theta), \quad k = 0 \ldots \mathrm{M}, \tag{3.5b}$$

with $Y_0 \equiv 1$ and $Y_{M+1} \equiv 0$. With a sequential scoring rule, if $Y_j = 1$, then $Y_{j-1}$ is equal to 1 as well, so that $\mathrm{Pr}_{\mathrm{m}}(Y_j = 1 \text{ and } Y_{j-1} = 1; \theta) = \mathrm{Pr}_{\mathrm{m}}(Y_j = 1; \theta)$. For $\mathrm{Pr}(Y_{j-1} = 1; \theta) \neq 0$, two more equivalent equations are therefore given by

$$\mathrm{Pr}(Y_j = 1 \mid Y_{j-1} = 1; \theta) \;=\; \frac{\mathrm{Pr}_{\mathrm{m}}(Y_j = 1; \theta)}{\mathrm{Pr}_{\mathrm{m}}(Y_{j-1} = 1; \theta)}; \tag{3.6a}$$

$$\mathrm{Pr}_{\mathrm{m}}(Y_j = 1; \theta) \;=\; \prod_{r=1}^{j} \mathrm{Pr}(Y_r = 1 \mid Y_{r-1} = 1; \theta), \tag{3.6b}$$

both for $j = 1 \ldots \mathrm{M}$, and again with $Y_0 \equiv 1$ for notational convenience. Equation (3.6b) follows from (3.6a) first for $j = 2$; then by mathematical induction for $j = 3$ and further.

A very useful expression for the conditional probability $\mathrm{Pr}(Y_j = 1 \mid Y_{j-1} = 1; \theta)$ is obtained as follows:

$$\begin{aligned}
\mathrm{Pr}(Y_j = 1 \mid Y_{j-1} = 1; \theta) &= \frac{\mathrm{Pr}_{\mathrm{m}}(Y_j = 1; \theta)}{\mathrm{Pr}_{\mathrm{m}}(Y_{j-1} = 1; \theta)} && \text{(by Eq. 3.6a)} \\
&= \frac{\sum_{k=j}^{M} \mathrm{Pr}(X = k; \theta)}{\sum_{k=j-1}^{M} \mathrm{Pr}(X = k; \theta)} && \text{(by Eq. 3.5a)} \\
&= \frac{\sum_{k=j}^{M} \left\{ \left[ \prod_{r=0}^{k} \mathrm{f}_r(\theta) \right] [1 \Leftrightarrow \mathrm{f}_{k+1}(\theta)] \right\}}{\sum_{k=j-1}^{M} \left\{ \left[ \prod_{r=0}^{k} \mathrm{f}_r(\theta) \right] [1 \Leftrightarrow \mathrm{f}_{k+1}(\theta)] \right\}} && \text{(by Eq. 3.2)}
\end{aligned}$$

$$= \frac{\prod_{r=0}^{j} \mathrm{f}_r(\theta)}{\prod_{r=0}^{j-1} \mathrm{f}_r(\theta)}$$

$$= \mathrm{f}_j(\theta). \tag{3.7}$$

The transition to the one but last line can be verified upon expanding both numerator and denominator. For example, for $j = 2$ and $\mathrm{M} = 4$ the numerator would be $\mathrm{f}_1\mathrm{f}_2[1 \Leftrightarrow \mathrm{f}_3] + \mathrm{f}_1\mathrm{f}_2\mathrm{f}_3[1 \Leftrightarrow \mathrm{f}_4] + \mathrm{f}_1\mathrm{f}_2\mathrm{f}_3\mathrm{f}_4 = \mathrm{f}_1\mathrm{f}_2$.

If, instead of sequential scoring, a complete design with local independence and response functions $\mathrm{f}_j(\theta)$ were considered, the marginal probability of obtaining $Y_j = 1$ would be equal to the ICC. Under the sequential scoring rule, however, it is the conditional probability $\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta)$ which is equal to the ICC.

The equality of $\Pr_{\mathrm{m}}(Y_j = 1)$ in a complete design and $\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta)$ under a sequential scoring rule can also be established using Rubin's (1976) ignorability principle. Let $\boldsymbol{Z} = (Z_1 \ldots Z_M)$ be a design vector, i.e. $Z_j = 1$ indicates that $Y_j$ is observed, and $Z_j = 0$ indicates that $Y_j$ is not observed. With sequential scoring $Z_j = 1$ iff $Y_{j-1} = 1$, therefore $Y_{j-1}$ is the design variable for $Y_j$. As the design does not depend upon the value of the unobserved data, it follows that these data are missing at random, whence the design can be ignored (Rubin, 1976). The ignorability principle states that if the design can be ignored, then

$$\Pr(Y_j = y_j \mid Z_j = 1; \theta) = \Pr(Y_j = 1; \theta, \text{no design}).$$

Rephrasing this into terms of a sequential scoring rule gives

$$\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta) = \Pr_B(Y_j = 1; \theta),$$

which is Equation 3.7.

Finally, because with sequential scoring $\mathrm{f}_j(\theta) = \Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta)$, Equations 3.6a and 3.6b can also be written as

$$\mathrm{f}_j(\theta) = \frac{\Pr_{\mathrm{m}}(Y_j = 1; \theta)}{\Pr_{\mathrm{m}}(Y_{j-1} = 1; \theta)}; \tag{3.8a}$$

$$\Pr_{\mathrm{m}}(Y_j = 1; \theta) = \prod_{r=1}^{j} \mathrm{f}_r(\theta), \tag{3.8b}$$

revealing that, with sequential scoring, the division of two consecutive marginal probabilities results in the ICCs. This explains why ICCs in a sequential process are also known as continuation ratios.

It may be noted that any model for sequential scoring is fully defined by the specification of either the ICCs or conditional probabilities $\Pr(Y_j = 1|Y_{j-1} = 1; \theta)$, or the marginal probabilities $\Pr_{\mathrm{m}}(Y_j = 1; \theta)$, or the score probabilities $\Pr(X = k; \theta)$. Equations (3.5a) - (3.6b) suffice to derive any one of these probabilities from any other.

If a model for polytomous item responses is used to describe the variable resulting from a sequential scoring rule, several requirements have to be fulfilled. A first requirement is that the score probabilities under the model comply with a set of ICCs and the particular covariance structure resulting from sequential scoring. Using Equations 3.5a through 3.6b, it can be shown that this first condition is always met: consider any model for $\Pr(X = k; \theta)$, say model A. Now assume that this model is used with sequential scoring. Applying Equation 3.5a will lead to formulae for $\Pr_{\mathrm{m}}(Y_j = 1; \theta)$; and using Equation 3.6a the formulae for $\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta)$ are obtained. If these ICCs are assumed, then the distribution of the number of successes upon applying a sequential scoring rule will be exactly equal to the distribution of $X$ under model A, as can be easily verified.

In Section 3.4 several item response models are presented by their score probabilities $\Pr(X = k; \theta)$, and the above manipulations will be performed to derive the ICCs that would, under a sequential scoring rule, return the original score probabilities. Then the ICCs thus derived will be investigated with respect to a second requirement. This second requirement is formulated in Section 3.3. It will appear that the second requirement is not met by all models.

## 3.3 A requirement for models describing sequential scoring

Consider two sequential processes, say processes A and B. Let the trials in process B be the same as those in process A, except for one trial, say trial q, which is not included in process B. Let process A consist of sequentially scoring the responses on a set of binary items $1, \ldots, \mathrm{M}$, and process B of sequentially scoring the responses on the binary items $1, \ldots, \mathrm{q} \Leftrightarrow 1, \mathrm{q}+1, \ldots, \mathrm{M}$. These sets of items will also be denoted as sets A and B. The index $j$ actually identifies the items, that is, if set A consists of items $1, 2, 3, 4, 5$ and set B of items $1, 2, 4, 5$, then the item labelled $'4'$ is meant to be the same item in both sets (and it is not meant to indicate the 4th item). To further avoid confusion, an asterisk will be added to the variables arising in the context of process B: $Y_j^*$ is the score on item $j$ when this item is tried in process B, and $Y_j$ is the score on the same item $j$ when it is tried in process A.

If a sequential process is modified by leaving out one of the trials, and if the resulting set of trials is again subjected to a sequential rule, then, using the ignorability principle, the parameters (ICCs) of the trials that figure in both sequences are equal. Similarly, if a sequence of Bernoulli trials is modified by inserting a new trial between two of the existing trials, and if both sets of trials are subjected to sequential scoring, the parameters (ICCs) of the trials that figure in both sequences are equal as well.

This implication of the ignorability principle will be called ICC-invariance for sequential scoring, or, short, ICC invariance. As ICC-invariance is an essential property of sequential scoring, any model for sequential scoring must be able to accommodate it. That is, any model for sequential scoring must be able to describe the score probabilities for processes A and B, allowing the ICCs for corresponding binary items (trials) to remain unchanged.

Recalling Equation 3.7, it holds both in set A and in set B that

$$\Pr(Y_j = 1 \mid Y_{j-1} = 1; \theta) \equiv \mathrm{f}_j(\theta),$$

for all $j$ in the set. In particular for $j = \mathrm{q} + 1$ this means that

$$\text{in set A:} \qquad \Pr(Y_{q+1} = 1 \mid Y_q = 1; \theta) \equiv \mathrm{f}_{q+1}(\theta);$$

$$\text{and in set B:} \qquad \Pr(Y^*_{q+1} = 1 \mid Y^*_{q-1} = 1; \theta) \equiv \mathrm{f}_{q+1}(\theta).$$

The two right hand sides are equal. Therefore the two left hand sides must also be equal. Consequently, for any two sequentially scored item response processes A and B, where the items in set B are the same ones as in set A, except for item q, which is not in set B, it holds that

$$\underset{\text{in set A}}{\Pr(Y_{q+1} = 1 \mid Y_q = 1; \theta)} = \underset{\text{in set B}}{\Pr(Y^*_{q+1} = 1 \mid Y^*_{q-1} = 1; \theta)} . \tag{3.9}$$

Any model for sequential scoring must be able to accommodate this equality. This second requirement can be used to investigate the suitability of a model for sequential scoring.

## 3.4 Suitability of several models for sequential scoring

In this section the possibility of the SM, the GRM, and the PCM for accommodating the equality in Equation 3.9 is investigated. The emphasis will be on removing a binary item, but the reasoning for inserting an item is analogous.

Consider the two sets A and B described in Section 3.3 and assume, for ease of presentation, that M = 6 and q = 3. That is, set A consists of items $1, 2, \ldots, 6$, which are tried in the order 123456, and set B consists of the same items except for item 3. The items in set B are tried in the order 12456.

In all models the parameter vector will be denoted as $\boldsymbol{\delta} = (\delta_1 \ldots \delta_M)$, although of course the interpretation of the parameters differs between models.

### 3.4.1 Sequential Model

As was noted in Section 3.1, the sequential model (Molenaar, 1983; Tutz, 1990; Tutz, 1997; Verhelst, Glas and de Vries, 1997) has been explicitly developed to model the score in a sequential process. The logistic SM is defined by assuming a 1-PL curve $\mathrm{h}(\theta;\delta)$ for the ICCs. This curve is given by

$$\mathrm{h}(\theta;\delta) = \frac{\exp(\theta \Leftrightarrow \delta)}{1+\exp(\theta \Leftrightarrow \delta)}. \tag{3.10}$$

Let the parameter vector for a single polytomous SM item with maximal score M, or alternatively, for a set of M sequentially scored binary items, be $\boldsymbol{\delta} = (\delta_1 \ldots \delta_j \ldots \delta_M)$, where the $\delta_j$'s need not be ordered. Then

$$\Pr(Y_j = 1|Y_{j-1} = 1; \theta, \boldsymbol{\delta}) = \mathrm{h}(\theta; \delta_j), \qquad \text{for } j = 1, \ldots, \mathrm{M}, \tag{3.11}$$

with, for notational convenience, $Y_0 \equiv 1$. This model can be used for sequential scoring without any problems. The extreme value SM too, whose ICCs are given by the extreme value function from Equation 3.3, is appropriate for sequential scoring.

### 3.4.2 Graded Response Model

Under the logistic graded response model (GRM) developed by Samejima (1969) the score probabilities are given by

$$\Pr(X = k; \theta, \boldsymbol{\delta}) = \mathrm{h}_k(\theta) \Leftrightarrow \mathrm{h}_{k+1}(\theta), \qquad \text{for } k = 0, 1, \ldots, \mathrm{M}, \tag{3.12}$$

where $\mathrm{h}_k(\theta) = \mathrm{h}(\theta; \delta_k)$ is the 1-parameter logistic function defined in (3.10). The elements of the parameter vector $\boldsymbol{\delta}$ are ordered such that $\delta_1 \leq \delta_2 \leq \ldots \leq \delta_M$.

The marginal probabilities for the binary variables in a sequential process which would be described by the GRM are found applying (3.5a) and (3.12):

$$\begin{aligned} \Pr\mathrm{m}(Y_j = 1; \theta, \boldsymbol{\delta}) &= \sum_{k=j}^{M} \Pr(X = k; \theta, \boldsymbol{\delta}) \\ &= \mathrm{h}_j(\theta), \quad \text{for } j = 1 \ldots \mathrm{M}. \end{aligned} \tag{3.13}$$

The ICCs are given by the conditional probabilities of succeeding on the next trial. For the GRM this gives, applying (3.6a) and (3.13):

$$\begin{aligned} \mathrm{f}_j(\theta) &= \Pr(Y_j = 1 | Y_{j-1} = 1; \theta, \boldsymbol{\delta}) = \frac{\mathrm{Prm}(Y_j = 1; \theta, \boldsymbol{\delta})}{\mathrm{Prm}(Y_{j-1} = 1; \theta, \boldsymbol{\delta})} \\ &= \frac{\mathrm{h}_j(\theta)}{\mathrm{h}_{j-1}(\theta)}. \end{aligned} \tag{3.14}$$

Having established the functional forms for the ICCs and the marginal probabilities, the possibility of accommodating the equality $\Pr(Y_{q+1} = 1 \mid Y_q = 1; \theta) = \Pr(Y^*_{q+1} = 1 \mid Y^*_{q-1} = 1; \theta)$ will now be investigated.

Assume that, under the logistic GRM, the third item is removed from a set of binary items with $\mathrm{M} = 6$. As $\mathrm{h}_0(\theta) = \mathrm{h}(\theta; \delta_0) \equiv 1$, the first line of Equation 3.14 implies that the model must allow

$$\mathrm{Prm}(Y^*_1 = 1; \theta) = \mathrm{Prm}(Y_1 = 1; \theta, \boldsymbol{\delta}). \tag{3.15}$$

Similarly, the model must allow $\Pr(Y^*_2 = 1 | Y^*_1 = 1; \theta) = \Pr(Y_2 = 1 | Y_1 = 1; \theta, \boldsymbol{\delta})$. Again applying (3.14) gives:

$$\frac{\mathrm{Prm}(Y^*_2 = 1; \theta)}{\mathrm{Prm}(Y^*_1 = 1; \theta)} = \frac{\mathrm{Prm}(Y_2 = 1; \theta, \boldsymbol{\delta})}{\mathrm{Prm}(Y_1 = 1; \theta, \boldsymbol{\delta})},$$

which, by virtue of (3.15), yields

$$\mathrm{Prm}(Y^*_2 = 1; \theta) = \mathrm{Prm}(Y_2 = 1; \theta, \boldsymbol{\delta}). \tag{3.16}$$

Until now there are no problems. Item 3 is not in set B, so it is not necessary to demand anything for item 3. For item 4 however, the model must allow $\Pr(Y^*_4 = 1 | Y^*_2 = 1; \theta) = \Pr(Y_4 = 1 | Y_3 = 1; \theta, \boldsymbol{\delta})$, which is equivalent to demanding

$$\frac{\mathrm{Prm}(Y^*_4 = 1; \theta)}{\mathrm{Prm}(Y^*_2 = 1; \theta)} = \frac{\mathrm{Prm}(Y_4 = 1; \theta, \boldsymbol{\delta})}{\mathrm{Prm}(Y_3 = 1; \theta, \boldsymbol{\delta})}.$$

Substituting (3.16) into this expression and rewriting it gives

$$\mathrm{Prm}(Y^*_4 = 1; \theta) = \frac{\mathrm{Prm}(Y_2 = 1; \theta, \boldsymbol{\delta})}{\mathrm{Prm}(Y_3 = 1; \theta, \boldsymbol{\delta})} \mathrm{Prm}(Y_4 = 1; \theta, \boldsymbol{\delta}),$$

which is not a 1-PL curve. Consequently $\Pr(X^* = 2; \theta)$, which by (3.5b) is equal to $\mathrm{Prm}(Y_2^* = 1; \theta) \Leftrightarrow \mathrm{Prm}(Y_4^* = 1; \theta)$, is not the difference between two 1-PL functions which it should be because of (3.12). Hence, for $X^* \geq 2$ the logistic GRM cannot, in set B, at the same time accommodate (a) ICCs equal to corresponding ICCs in set A, and (b) the function in (3.12) for the score probabilities. And therefore the logistic GRM is not suited for modeling the probabilities in a sequential scoring process.

It may be noted that the logistic GRM does allow removal of the last item. If it would also allow removal of the first item, it could still be useful for sequential processes of the second kind distinguished in the Introduction; but removing the first item will lead to similar problems as removing a third item.

For the extreme value GRM, the logistic curves have to be replaced by extreme value curves. Using the reparameterization mentioned in the Introduction, it can be shown that it is no problem to apply the extreme value GRM to sequential scoring.

### 3.4.3 Partial Credit Model

Under the PCM (Masters, 1982; Andrich, 1978) the probability of obtaining a score $X = k$ is given by

$$\Pr(X = k; \theta, \boldsymbol{\delta}) = \frac{1}{\mathrm{D}_M}\left[\exp\left(k\theta \Leftrightarrow \sum_{r=1}^{k} \delta_r\right)\right], \quad \text{for } k = 0, 1, \ldots, \mathrm{M}, \qquad (3.17)$$

with $\sum_{r=1}^{0}(\Leftrightarrow\delta_r) \equiv 0$ and $\mathrm{D}_M = 1 + \exp(\theta \Leftrightarrow \delta_1) + \exp(2\theta \Leftrightarrow \delta_1 \Leftrightarrow \delta_2) + \ldots + \exp(M\theta \Leftrightarrow \delta_1 \Leftrightarrow \delta_2 \Leftrightarrow \ldots \Leftrightarrow \delta_M)$.

Under the PCM as a model for sequential scoring, the marginal probabilities for $j = 1 \ldots \mathrm{M}$ are found applying (3.5a):

$$\mathrm{Prm}(Y_j = 1; \theta, \boldsymbol{\delta}) \;=\; \sum_{k=j}^{M} \Pr(X = k; \theta, \boldsymbol{\delta})$$

$$= \frac{1}{\mathrm{D}_M} \sum_{k=j}^{M} \left[ \exp\left( k\theta \Leftrightarrow \sum_{r=1}^{k} \delta_r \right) \right], \tag{3.18}$$

with $\mathrm{D}_M$ as above. As an example, for M = 4, and using $\xi = \exp(\theta)$ and $\varepsilon_k = \exp(\Leftrightarrow \delta_k)$, the marginal probabilities are given by:

$$\begin{array}{rcr}
\mathrm{Prm}(Y_1 = 1; \theta, \boldsymbol{\delta}) & = & (\xi\varepsilon_1 + \xi^2\varepsilon_1\varepsilon_2 + \xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4, \\
\mathrm{Prm}(Y_2 = 1; \theta, \boldsymbol{\delta}) & = & (\xi^2\varepsilon_1\varepsilon_2 + \xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4, \\
\mathrm{Prm}(Y_3 = 1; \theta, \boldsymbol{\delta}) & = & (\xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4, \\
\mathrm{Prm}(Y_4 = 1; \theta, \boldsymbol{\delta}) & = & (\xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4)/\mathrm{D}_4.
\end{array} \tag{3.19}$$

The ICCs, or, alternatively, the conditional probabilities, are found from (3.6a) and (3.18):

$$\begin{aligned}
\mathrm{f}_j(\theta) &= \Pr(Y_j = 1 | Y_{j-1} = 1; \theta, \boldsymbol{\delta}) \\
&= \frac{\sum_{k=j}^{M} \exp\left(k\theta \Leftrightarrow \sum_{r=1}^{k} \delta_r\right)}{\sum_{k=j-1}^{M} \exp\left(k\theta \Leftrightarrow \sum_{r=1}^{k} \delta_r\right)} \\
&= \frac{\sum_{k=j}^{M} \exp\left[(k \Leftrightarrow j + 1)\theta \Leftrightarrow \sum_{r=j}^{k} \delta_r\right]}{1 + \sum_{k=j}^{M} \exp\left[(k \Leftrightarrow j + 1)\theta \Leftrightarrow \sum_{r=j}^{k} \delta_r\right]}.
\end{aligned} \tag{3.20}$$

Writing this out, again as an example for M = 4, gives:

$$\mathrm{f}_1(\theta) = \Pr(Y_1 = 1 | Y_0 = 1; \theta, \boldsymbol{\delta}) = \frac{\xi\varepsilon_1 + \xi^2\varepsilon_1\varepsilon_2 + \xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4}{1 + \xi\varepsilon_1 + \xi^2\varepsilon_1\varepsilon_2 + \xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4},$$

$$\mathrm{f}_2(\theta) = \Pr(Y_2 = 1 | Y_1 = 1; \theta, \boldsymbol{\delta}) = \frac{\xi\varepsilon_2 + \xi^2\varepsilon_2\varepsilon_3 + \xi^3\varepsilon_2\varepsilon_3\varepsilon_4}{1 + \xi\varepsilon_2 + \xi^2\varepsilon_2\varepsilon_3 + \xi^3\varepsilon_2\varepsilon_3\varepsilon_4},$$

$$\mathrm{f}_3(\theta) = \Pr(Y_3 = 1 | Y_2 = 1; \theta, \boldsymbol{\delta}) = \frac{\xi\varepsilon_3 + \xi^2\varepsilon_3\varepsilon_4}{1 + \xi\varepsilon_3 + \xi^2\varepsilon_3\varepsilon_4},$$

$$\mathrm{f}_4(\theta) = \Pr(Y_4 = 1 | Y_3 = 1; \theta, \boldsymbol{\delta}) = \frac{\xi\varepsilon_4}{1 + \xi\varepsilon_4}.$$

Apparently, if the next trial is trial $j$, the probability of succeeding on that next trial under the PCM is equal to the probability of succeeding, under the PCM, on the first trial in a 'smaller' sequential process with $\mathrm{M}' = \mathrm{M} \Leftrightarrow (j \Leftrightarrow 1)$ and parameters $(\delta_j \ldots \delta_M)$.

To simplify the notation, let, in the previous example,

$$\begin{aligned}
\mathrm{G}_1 &= \xi\varepsilon_1 + \xi^2\varepsilon_1\varepsilon_2 + \xi^3\varepsilon_1\varepsilon_2\varepsilon_3 + \xi^4\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4,\\
\mathrm{G}_2 &= \xi\varepsilon_2 + \xi^2\varepsilon_2\varepsilon_3 + \xi^3\varepsilon_2\varepsilon_3\varepsilon_4,\\
\mathrm{G}_3 &= \xi\varepsilon_3 + \xi^2\varepsilon_3\varepsilon_4,\\
\mathrm{G}_4 &= \xi\varepsilon_4;
\end{aligned} \tag{3.21}$$

and note that it is possible to define the PCM by the following two equations, where $\mathrm{G}_{M+1} \equiv 0$ for notational convenience:

$$\begin{cases} \mathrm{G}_j = \xi\varepsilon_j(1 + \mathrm{G}_{j+1}) & \text{for } j = \mathrm{M} \ldots 1, \\ \\ \Pr(Y_j = 1|Y_{j-1} = 1; \theta) = \mathrm{G}_j/(1 + \mathrm{G}_j), & \text{for } j = 1 \ldots \mathrm{M}. \end{cases} \tag{3.22}$$

The necessary preliminaries having been established, the PCM can now be investigated as a model for a sequential scoring. Removing an item other than the first one, will cause trouble. Combining Equation 3.9 and the second line of (3.22), it is necessary, if for example item 3 is deleted, that:

$$\begin{aligned}
\Pr(Y_1^* = 1|Y_0^* = 1; \theta) &= \Pr(Y_1 = 1|Y_0 = 1; \theta, \boldsymbol{\delta}) &&= \mathrm{G}_1/(1 + \mathrm{G}_1),\\
\Pr(Y_2^* = 1|Y_1^* = 1; \theta) &= \Pr(Y_2 = 1|Y_1 = 1; \theta, \boldsymbol{\delta}) &&= \mathrm{G}_2/(1 + \mathrm{G}_2),\\
\Pr(Y_4^* = 1|Y_2^* = 1; \theta) &= \Pr(Y_4 = 1|Y_3 = 1; \theta, \boldsymbol{\delta}) &&= \mathrm{G}_4/(1 + \mathrm{G}_4).
\end{aligned}$$

In this sequence $\mathrm{G}_2$ will in general not be equal to $\xi\varepsilon_2(1 + \mathrm{G}_4)$. Therefore the above is not a sequence of PCM probabilities, as the first line of (3.22) is not satisfied.

It must be concluded that under the PCM as a model for sequential scoring, it is impossible to satisfy both ICC invariance and the definition of the PCM as it was formulated in (3.22): a specification error is made if the PCM were applied to a sequential process.

Only if the first item were deleted, the new sequence of conditional probabilities $\mathrm{G}_2/(1+\mathrm{G}_2),\ldots,\mathrm{G}_M/(1+\mathrm{G}_M)$ would still comply with the assumption of equal ICCs in both processes; the resulting new PCM item would have parameters $\boldsymbol{\delta}^* = (\delta_2,\ldots,\delta_M)$. However, removal of the last item is again problematic so that the PCM is unsuited for sequential processes of both the kinds that were distinguished in the Introduction.

## 3.5 Discussion

The results obtained in this paper can be used when a model has to be chosen for describing the variable resulting from the application of a sequential scoring rule. As the SM was developed for modeling sequential processes, it is not surprising that this model does a good job on the criterion investigated. However, until now it was not known whether the mistakes made by applying a different model were of a practical or of a fundamental nature. It has been demonstrated that the application of the PCM or the logistic GRM to a variable resulting from sequential scoring, amounts to making a specification error.

Tutz (1997) pointed out that the extreme value SM and the extreme value GRM are the same models. Using this equivalence, another very practical warning can be formulated: it is not necessary to believe that a model that is formulated to describe one kind of random experiment, is *therefore* unsuited to describe another random experiment.

A topic for future research would be to further explore the similarity in structures of the SM and the GRM. In (3.14) the ICC for the GRM was shown to be equal to the ratio of two 1-PL curves; the formula for the 1-PL curve will now be substituted into this equation:

$$\Pr(Y_j = 1|Y_{j-1} = 1; \theta, \boldsymbol{\delta}) = \frac{\Pr(Y_j = 1; \theta, \boldsymbol{\delta})}{\Pr(Y_{j-1} = 1; \theta, \boldsymbol{\delta})} = \frac{\mathrm{h}(\theta; \delta_j)}{\mathrm{h}(\theta; \delta_{j-1})}$$

$$= \frac{\exp(\theta \Leftrightarrow \delta_j)}{1 + \exp(\theta \Leftrightarrow \delta_j)} \frac{1 + \exp(\theta \Leftrightarrow \delta_{j-1})}{\exp(\theta \Leftrightarrow \delta_{j-1})}$$

$$
\begin{aligned}
&= \exp(\delta_{j-1} \Leftrightarrow \delta_k) \frac{1 + \exp(\theta \Leftrightarrow \delta_{j-1})}{1 + \exp(\theta \Leftrightarrow \delta_j)} \\
&= \frac{\exp(\delta_{j-1} \Leftrightarrow \delta_j) + \exp(\theta \Leftrightarrow \delta_j)}{1 + \exp(\theta \Leftrightarrow \delta_j)} \\
&= \frac{\exp(\theta \Leftrightarrow \delta_j)}{1 + \exp(\theta \Leftrightarrow \delta_j)} + \frac{\exp(\delta_{j-1} \Leftrightarrow \delta_j)}{1 + \exp(\theta \Leftrightarrow \delta_j)} \\
&= \mathrm{h}(\theta; \delta_j) + \exp(\delta_{j-1} \Leftrightarrow \delta_j)[1 \Leftrightarrow \mathrm{h}(\theta; \delta_j)]. \qquad (3.23)
\end{aligned}
$$

Apparently, under the GRM the ICC $\Pr(Y_j = 1|Y_{j-1} = 1; \theta, \boldsymbol{\delta})$ follows a three parameter logistic (3PL) curve (Birnbaum, 1968) with discrimination parameter equal to 1, location parameter equal to $\delta_j$ and guessing parameter equal to $\exp(\delta_{j-1} \Leftrightarrow \delta_j)$ for $j = 2, 3, \ldots, \mathrm{M}$. The guessing parameter for the first item is 0, as $\Pr(Y_0 = 1; \theta) \equiv 1$. So assuming for the ICCs in the SM a 3PL function with the constraint $\delta_j = \exp(\delta_{j-1} \Leftrightarrow \delta_j)$ for $j = 2, \ldots, \mathrm{M}$ and $\delta_1 = 0$, will result in the GRM. This too reveals a similarity in structures of the GRM and the SM.

As a final remark: it can be shown that every 3-PL curve with discrimination parameter $\alpha$, guessing parameter $\gamma$, and location parameter $\delta$, can be written as a ratio of two 2-PL curves, both with discrimination parameter equal to $\alpha$, and with location parameters given by $\delta$ and $\delta + \ln(\gamma)$, respectively. The formula for each 3-PL curve can therefore be rewritten as a formula containing one discrimination parameter, and two location parameters. It would be an interesting question whether this could be overparameterization, and hence might contribute towards the difficulties that are commonly encountered in estimating the parameters in the 3PL model.

# Chapter 4

# Distinguishing between models for polytomous item responses[1]

Abstract

This paper reports on two studies concerning the possibility of empirically distinguishing between item response models. The first study is a modified replication of an experiment performed by Maydeu-Olivares et al. (1994). First, a procedure is described for selecting the models that will be actually compared. An observer is then is presented with two data sets and two completely specified item response models. Each data set has been generated by one of the models. The observer has to decide which model generated which data set. The item parameters are treated as known, and they are used in the decision. When a most powerful test is used for the decision, the percentage of correct classifications is known as the ideal observer index (IOI). The IOI can be used as an indicator for the difference between two models. In the present paper a criterion based on the IOI is investigated: it is suggested to evaluate the difference between two models by the sample size that is needed to have the IOI exceed .95. For those models that were selected for the investigation, it is found that, when the two models being compared are from the same family, this sample size has to be about twice as large as when the models being compared are from different families.
In the second part of the paper one data set and several families of item response models are considered. The question is which of these families can best be used to describe the data. The parameters are assumed to be unknown: they have to be estimated from the data. A decision theoretic approach is employed and that model is decided on that has the largest likelihood. The procedure performs well, in the sense that with simulated data the correct model was always recognized, and the type I error and the power under the alternatives are satisfactory. The procedure is also applied to an empirical data set. The type I error is larger here, but this is a consequence of the smaller sample size. The power under the alternatives is again high.

Key words: Response pattern classification, ideal observer index, model choice, bootstrap, partial credit model, graded response model, sequential model.

[1] The author wishes to thank dr. H. Vorst and dr. B. van Engelenburg for kindly providing her with the data used in Sections 4.3 and 4.4.

## 4.1 Introduction

This paper is concerned with the question which item response model, or which family of models, can best be used to describe a data set. The models considered are the partial credit model (Masters, 1982; Andrich, 1978; Andersen, 1977), the graded response model (Samejima, 1969), and the sequential model (Tutz 1990, 1997; Verhelst, Glas and de Vries, 1997). These models are defined in Section 4.2.

The paper consists of two main parts, which deal with some practical and statistical aspects of model choice in item response theory (IRT), respectively. The first part of the paper, Section 4.3, focuses on the possibility of distinguishing between data generated under the three item response models mentioned above. If it appears to be difficult to distinguish data generated under different models, the practical consequences of using a 'wrong' model may not be very large. In the second part of the paper, a decision theoretic approach to deciding on a family of models is proposed and examined.

To investigate the first question, a method used by Maydeu-Olivares et al. (1994), and proposed by Levine et al. (1992), is employed. Basically, this method is an application of the two-alternative forced choice experiment, which is itself an extension of the yes/no experiment in signal detection theory (Green and Swets, 1966). In the two-alternative forced choice experiment there are two sources, say sources A and B, both emitting a signal. An observer knowing all the relevant characteristics of the two sources is presented with the two signals in a random order; in 50 percent of the presentations the signal from source A is presented first. The observer has to decide which signal comes from source A, and which signal comes from source B. The procedure is repeated a large number of times. By chance alone, there would be a probability of .50 of correctly classifying the two signals. If the two sources emit very similar signals, it will be difficult to distinguish between them and the observer will not be able to perform much better than by chance alone. If the sources emit rather different signals, the classification task will be easier and the result can be more often expected to be correct. The rate of correct classification is known as the ideal observer index (IOI). It can be used as an index for the similarity of the signals emitted by the two sources. Maydeu-Olivares et al. (1994) used this method to compare item response models. The procedure for selecting the models to be compared will be described in detail in Section 4.3.2. Response patterns generated under different models from the PCM and the GRM families were treated as signals. In all cases, the IOI for the comparison of two models from a single family was found to be significantly higher than the IOI for the comparison of two models from different families. However, in value the differences between these IOIs were small. In the present paper an alternative criterion for the difference between two models is investigated. This alternative criterion is derived from the IOI: it is the sample size needed to obtain an IOI larger than .95. Furthermore, the comparison includes the SM as well as the PCM and the GRM. The results are very clear: it is found that for the comparison of two models from different families, this sample size has to be about twice as large as for the comparison of two models from a single family. These results pertain to the models that were selected for the comparison; it is not claimed that they hold in general.

In the above experiment, the parameter values used to generate the data are used in the classification decision. That is, in the IOI study the true parameter values are used. The second part of the paper deals with a situation which is more realistic, in that the parameter values are not assumed to be a priori known. The problem here is to decide upon the best model for the data at hand. Most of the literature concerning statistical tests for model choice deals with situations where the choice is restricted to models from one family, such as for example the number of factors in a factor analysis, or the predictors in a regression analysis. In such cases, a likelihood ratio statistic can be used, or an information criterion, such as e.g. Akaikes AIC (Akaike, 1973; Akaike, 1974) or Schwarz' BIC (Schwarz, 1978). Gelfand and Ghosh (1998) showed that many criteria can be rewritten as a combination of a goodness-of-fit term and a penalty term for the number of parameters. In Section 4.4 the possibility of deciding between models from different families is investigated. It is assumed that the numbers of parameters under each family are equal. Therefore, criteria with a penalty term for the number of parameters are not really called for. The procedure proposed is as follows. First, for each of the three families of item response models under consideration, that member is identified that best fits the data. Then a zero/one loss function and a maximum likelihood decision rule are applied in order to decide upon a model. The size of the type I error of this procedure will be investigated, and also the power under the alternatives. The procedure is applied both to simulated and to empirical data. The results appear to be satisfactory.

## 4.2 Preliminaries: definitions and notation

The variable $X_j = k$, for $k = 0, \dots, \mathrm{M}$, will denote the score on a polytomous item $j$, for $j = 1 \dots \mathrm{L}$. Note that the symbol M refers to the maximum score, and L is used for the test length. Furthermore, $\theta$ will represent a latent variable which is measured by the items. Let $\mathrm{g}(\theta)$ be the density function of $\theta$; this density is assumed to be standard normal. The subjects are supposed to be randomly sampled from $\mathrm{g}(\theta)$. Given $\theta$, under the partial credit model (PCM) (Masters, 1982; Andrich, 1978; Andersen, 1977), the probability of obtaining a score $k$ on item $j$, with parameters $\alpha$ and $\boldsymbol{\delta}_j = (\delta_{1j} \dots \delta_{kj} \dots \delta_{Mj})$, as a function of $\theta$, is assumed to be given by

$$\mathrm{Pr}(X_j = k; \theta, \alpha, \boldsymbol{\delta}_j) = \frac{\exp\left[\alpha(k\theta \Leftrightarrow \sum_{p=1}^{k} \delta_{pj})\right]}{\sum_{r=0}^{M} \exp\left[\alpha(r\theta \Leftrightarrow \sum_{p=1}^{r} \delta_{pj})\right]}, \qquad j = 1 \dots \mathrm{L}, \quad k = 0 \dots \mathrm{M},$$

with $\sum_{p=1}^{0} \delta_{pj} \equiv 0$. The parameter $\alpha$ is a scale parameter; it is constant over all items within a test, and it enters into the model if the variance for $\theta$ is fixed. In this paper, the probabilities of interest are the marginal probabilities integrated over the range of $\theta$. Letting $\boldsymbol{X}$ be the vector $(X_1, \dots, X_L)$, then for a test of length L, and assuming that scores on different items are independent given $\theta$, these marginal probabilities are given by

$$\begin{aligned}\mathrm{Pr}(\boldsymbol{X} = \boldsymbol{x}; \alpha, \boldsymbol{\delta}_1 \dots \boldsymbol{\delta}_L) &= \int_{\Theta} \mathrm{Pr}(\boldsymbol{X} = \boldsymbol{x}; \theta, \alpha, \boldsymbol{\delta}_1 \dots \boldsymbol{\delta}_L) \mathrm{g}(\theta) \mathrm{d}\theta \\ &= \int_{\Theta} \prod_{j=1}^{L} \mathrm{Pr}(X_j = x_j; \theta, \alpha, \boldsymbol{\delta}_j) \mathrm{g}(\theta) \mathrm{d}\theta. \qquad (4.1)\end{aligned}$$

The graded response model (GRM) (Samejima, 1969) has parameters $\beta$ and

$\boldsymbol{\gamma}_j = (\gamma_{1j} \ldots \gamma_{kj} \ldots \gamma_{Mj})$ for item $j$, where $\beta$ is again a scale parameter. The elements of the parameter vector are ordered such that $\gamma_{1j} \leq \gamma_{2j} \leq \ldots \leq \gamma_{Mj}$. Given $\theta$, in this model the score probabilities are assumed to be

$$\Pr(X_j = k; \theta, \beta, \boldsymbol{\gamma}_j) = \mathrm{h}_k(\theta; \beta, \boldsymbol{\gamma}_j) \Leftrightarrow \mathrm{h}_{k+1}(\theta; \beta, \boldsymbol{\gamma}_j), \qquad j = 1 \ldots \mathrm{L}, \;\; k = 0 \ldots \mathrm{M},$$

where $\mathrm{h}_k$ is given by

$$\mathrm{h}_k(\theta; \beta, \boldsymbol{\gamma}_j) = \frac{\exp[\beta(\theta \Leftrightarrow \gamma_{kj})]}{1 + \exp[\beta(\theta \Leftrightarrow \gamma_{kj})]}, \qquad k = 1, \ldots, \mathrm{M}. \tag{4.2}$$

For notational convenience, $\mathrm{h}_0(\theta; \beta, \boldsymbol{\gamma}_j) \equiv 1$ and $\mathrm{h}_{M+1}(\theta; \beta, \boldsymbol{\gamma}_j) \equiv 0$. For constant $\beta$, the function in (4.2) is the item characteristic curve for a binary item in the well-known Rasch model (Rasch, 1960). Again assuming independence between scores on different items, the marginal probability of the score vector follows as

$$\Pr(\boldsymbol{X} = \boldsymbol{x}; \beta, \boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_L) = \int_\Theta \prod_{j=1}^{L} \Pr(X_j = x_j; \theta, \beta, \boldsymbol{\gamma}_j) \mathrm{g}(\theta) \mathrm{d}\theta. \tag{4.3}$$

Again the scale parameter $\beta$ has to be estimated if the density of $\theta$ is assumed to be standard normal.

In the sequential model or SM (Tutz 1990, 1997; Verhelst, Glas and de Vries, 1997) the parameter vector for item $j$ is $\boldsymbol{\sigma}_j = (\sigma_{1j} \ldots \sigma_{kj} \ldots \sigma_{Mj})$, and the scale parameter will be denoted as $\tau$. The $\sigma_{kj}$'s need not be ordered. The score probabilities in this model are given by

$$\Pr(X_j = k; \theta, \tau, \boldsymbol{\sigma}_j) = \prod_{r=1}^{k} \mathrm{h}_r(\theta; \tau, \boldsymbol{\sigma}_j) \Leftrightarrow \prod_{r=1}^{k+1} \mathrm{h}_r(\theta; \tau, \boldsymbol{\sigma}_j) \qquad j = 1 \ldots \mathrm{L}, \;\; k = 0 \ldots \mathrm{M},$$

where the functions $\mathrm{h}_r$ are defined as in (4.2), and here $\prod_{r=1}^{0} \mathrm{h}_r(\theta; \tau, \boldsymbol{\sigma}_j) \equiv 1$ and $\prod_{r=1}^{M+1} \mathrm{h}_r(\theta; \tau, \boldsymbol{\sigma}_j) \equiv 0$. The marginal probabilities of the score patterns are given by

$$\Pr(\boldsymbol{X} = \boldsymbol{x}; \tau, \boldsymbol{\sigma}_1 \ldots \boldsymbol{\sigma}_L) = \int_\Theta \prod_{j=1}^{L} \Pr(X_j = x_j; \theta, \tau, \boldsymbol{\sigma}_j) \mathrm{g}(\theta) \mathrm{d}\theta. \tag{4.4}$$

## 4.3 Deciding between models when the parameter values are known: the ideal observer index

In this section the sample size needed for the IOI to exceed .95 is examined. Basically, the experiment described here is a modified replication of the Maydeu-Olivares et al. (1994) experiment (see also Levine et al., 1992). The difference is in the use of an alternative criterion variable. Below, first the IOI will be defined. Two topics associated with the use of the IOI are addressed in Section 4.3.2. In Section 4.3.3 a short overview is given of results previously obtained with the IOI. Section 4.3.4 focuses on the alternative criterion variable. The design of the present study is summarized in Section 4.3.5, and Section 4.3.6 contains the results.

### 4.3.1 The Ideal Observer Index

Levine et al. (1992) describe how the IOI can be used to distinguish between item response models. In their application, item response models act as sources and a score pattern is the equivalent of a signal. The part of the observer can be played by a computer, into which the decision rule has been programmed. The observer is presented with two simulated score patterns, say patterns $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, and has to decide whether pattern $\boldsymbol{u}_1$ was generated under model A (and hence pattern $\boldsymbol{u}_2$ under model B), or vice versa. The observer knows the models and the parameter values that were used to generate the score pattern under each model. The only thing the observer does not know is which pattern was generated under which model.

The observer is supposed to be an 'ideal observer', that is, an observer making statistically optimal decisions. A statistically optimal decision in this classification task would be a most powerful test of the hypothesis

H$_1$: $\boldsymbol{u}_1$ was generated under model A and $\boldsymbol{u}_2$ under model B: the order is AB;

H$_2$: $\boldsymbol{u}_1$ was generated under model B and $\boldsymbol{u}_2$ under model A: the order is BA.

Because the value of the parameter vectors is known, the probability mass functions under both $\mathrm{H}_1$ and $\mathrm{H}_2$ are completely specified. Consequently, for any given significance level $\alpha$ and sample size n, a most powerful test of the hypothesis $\mathrm{H}_1$ versus $\mathrm{H}_2$ can be obtained as follows: let $L_1 = L(\boldsymbol{u}_1$ from A and $\boldsymbol{u}_2$ from B), and $L_2 = L(\boldsymbol{u}_1$ from B and $\boldsymbol{u}_2$ from A). Then, using

$$\lambda = \frac{L_1}{L_2}, \tag{4.5}$$

the decision rule is formulated as:

> if $\lambda > \mathrm{k}$, for some k, decide that the order is AB; if $\lambda < \mathrm{k}$, decide that the order is BA; randomize if $\lambda = \mathrm{k}$.

The quantity $\lambda$ is a simple likelihood ratio statistic, and it can be shown by the Neyman Pearson Lemma (Kendall and Stuart, 1979; Lehmann, 1959) that this decision rule is a most powerful test of the simple hypothesis $\mathrm{H}_1$ versus the simple hypothesis $\mathrm{H}_2$. In this experiment there are no clear null and alternative hypothesis, so there is no reason to favor either $\mathrm{H}_1$ or $\mathrm{H}_2$. Therefore k is taken to be 1, so that the procedure reads: if $L_1 > L_2$ decide AB; if $L_1 < L_2$ decide BA; otherwise randomize. The percentage of correct classifications obtained through this rule is called the ideal observer index (IOI).

Let $L_A(\boldsymbol{u}_1)$ be the likelihood of $\boldsymbol{u}_1$ under model A. The quantities $L_B(\boldsymbol{u}_1)$, $L_A(\boldsymbol{u}_2)$ and $L_B(\boldsymbol{u}_2)$ are analogously defined. If independence can be assumed between $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, the likelihoods can be written out as follows: $L_1(\boldsymbol{u}_1, \boldsymbol{u}_2) = L_A(\boldsymbol{u}_1)L_B(\boldsymbol{u}_2)$ and $L_2(\boldsymbol{u}_1, \boldsymbol{u}_2) = L_B(\boldsymbol{u}_1)L_A(\boldsymbol{u}_2)$. This gives

$$\begin{aligned} L_1 > L_2 \quad &\Leftrightarrow \quad L_A(\boldsymbol{u}_1)L_B(\boldsymbol{u}_2) > L_B(\boldsymbol{u}_1)L_A(\boldsymbol{u}_2) \\ &\Leftrightarrow \quad \frac{L_A(\boldsymbol{u}_1)}{L_B(\boldsymbol{u}_1)} > \frac{L_A(\boldsymbol{u}_2)}{L_B(\boldsymbol{u}_2)}, \end{aligned}$$

so letting

$$\lambda_1 = \frac{L_A(\boldsymbol{u}_1)}{L_B(\boldsymbol{u}_1)}, \qquad \text{and} \qquad \lambda_2 = \frac{L_A(\boldsymbol{u}_2)}{L_B(\boldsymbol{u}_2)}, \tag{4.6}$$

the decision rule for independent $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ can also be written as: if $\lambda_1 > \lambda_2$, decide AB; if $\lambda_1 < \lambda_2$, decide BA; otherwise randomize.

### 4.3.2 Use of the IOI in IRT

Each parametric item response model in fact constitutes an entire family of models. Every value of the parameter vector corresponds with a different member of the family; and each member of the family is itself a distinct model. To select the models to be used in the comparison, Maydeu-Olivares et al. (1994) and Levine et al. (1992) started from an empirical data set,and for each family under consideration they identified the member fitting best to these data.

However, these two best-fitting models were not directly compared to each other, because when members from different families are compared, there may be need for a baseline value. Suppose that an IOI of, say, .75 were found for the difference between the two best-fitting members. Then what would this number actually mean? A baseline value might help facilitate the interpretation of the observed value of the IOI. Maydeu-Olivares et al. adopt the following procedure for the purpose of establishing a baseline IOI.

The first part in the baseline procedure consists of obtaining an empirical data set and identifying the best-fitting member from one of the families to it, say from family $\mathcal{A}$. The actual member of family $\mathcal{A}$ that has been fitted to the data will be denoted as model A. An artificial data set is then simulated from model A. Therefore model A will be called the simulation model. To this artificial data set both families $\mathcal{A}$ and $\mathcal{B}$ are fitted, resulting in two fitted members that are denoted as A$'$ and B$'$. Models A$'$ and B$'$ will be called the estimated models.

Now two different IOIs can be calculated, to establish (a) the degree to which it is possible to distinguish between models A and A$'$, and (b) the degree to which it is possible to distinguish between models A and B$'$. The former IOI will be denoted as IOI$_{AA'}$, the latter as IOI$_{AB'}$. In this procedure there is no direct comparison of models A$'$ and B$'$. However, this procedure does allow the comparison of IOI$_{AB'}$ with IOI$_{AA'}$, where IOI$_{AA'}$ acts as the baseline for the interpretation of IOI$_{AB'}$. As an example, suppose that IOI$_{AB'} = .75$ and IOI$_{AA'} = .73$. Then one would conclude that the models A and A' from family $\mathcal{A}$ can differ just as much from each other as model A differs from model B$'$. If, on the other hand, IOI$_{AB'} = .75$

and IOI$_{AA'}$ = .52, the conclusion would be that the difference between models A and B' is much larger than the difference between the two models from family $\mathcal{A}$.

The second part of the baseline procedure consists in repeating the above with family $\mathcal{B}$ in the role of family $\mathcal{A}$. This baseline procedure will be employed in the present paper as well.

### 4.3.3 Previous results obtained with the IOI

Maydeu-Olivares et al. (1994) compared the GRM and the PCM with the procedure described above. In their experiment the sample size N took values of $250, 500, 1000$ and 3000, and the number of items L varied from 5 through 10 to 25. Therefore there were 4 (N's) × 3 (L's) × 2 (simulation models) × 2 (estimated models) = 48 IOIs reported. These ranged in value from .527 to .714, with an average of .608. Maydeu-Olivares et al. performed a regression analysis with log[IOI/(1 ⇔IOI)] as the dependent variable, and L, N and L × N as predictors. They found (a) that the IOI increased with the number of items L; (b) that there was a negative regression coefficient for the interaction term L × N; (c) that there was no significant effect of the sample size N; and (d) that for both families investigated (GRM and PCM), the IOI$_{AB'}$ was slightly higher than IOI$_{AA'}$.

Van Engelenburg (1997, Chapter 1) investigated differences between PCM, GRM and SM. He used N = 300 and L = 8 throughout. On the whole, his results were comparable to those of Maydeu-Olivares et al. His maximum IOI was equal to .620. He furthermore investigated the effects of a rating scale restriction and an equality restriction on the discrimination parameter. In all cases models without restriction were more easily distinguished than models with restriction.

### 4.3.4 A criterion derived from the IOI

In the Maydeu-Olivares et al. experiment the variable N is the sample size of the data sets simulated from model A. Using this simulated data set, models A$'$ and B$'$

are estimated. Once the parameters of $A'$ and $B'$ have been estimated, Maydeu-Olivares et al. proceed to simulate N samples of size 1 both from $A'$ and $B'$, and to classify these simulated score vectors. The percentage of correct classifications of these N score vectors constitutes their IOI. Recall from Section 4.3.3 that there proved to be no marked effect of N on this IOI.

It is, however, also possible to calculate IOIs on the basis of a sample size larger than 1, that is, to calculate the likelihood ratio statistic $\lambda$ for two score matrices instead of for two score vectors. It is then possible to distinguish two different sample sizes, which will be denoted here as N and n: N is the size of the sample simulated from A, and used to estimate the parameters of $A'$ or $B'$; and n is the size of the samples simulated from $A'$ and $B'$, and used to calculate the IOIs. It can be shown that the power under the likelihood ratio test for a simple null and alternative hypothesis increases with n (see e.g. Serfling, 1980, Section 4.4.3). The sample size n needed to obtain an IOI of .95 is proposed here as an alternative criterion for assessing the difference between two models.

The following justification can be given for this alternative criterion. In designing a statistical test, usually a null hypothesis is formulated and a significance level $\alpha$ is determined. If the critical value k were given, one could select the sample size n such that $\Pr(\lambda \geq \mathrm{k} \mid \mathrm{n},\ \mathrm{H}_0) = \alpha$, where $\lambda$ is the test statistic and $\alpha$ refers to the type I error, i.e. $\alpha = \Pr(\text{incorrect decision} \mid \mathrm{H}_0)$. In the investigation at hand, however, there is no compelling reason for selecting either $\mathrm{H}_1$ or $\mathrm{H}_2$ as a null hypothesis. That is, there is no clear argument for choosing either $\alpha = \Pr(\text{incorrect decision} \mid \text{the presentation order is AB})$, or $\alpha = \Pr(\text{incorrect decision} \mid \text{the presentation order is BA})$. This symmetry was also the reason for selecting the critical value $\mathrm{k} = 1$.

As a way out of this dilemma, it is proposed to replace $\alpha$ by the quantity $\varepsilon$ which is defined here as the unconditional probability of an incorrect decision:

$$\begin{aligned} \varepsilon &= \Pr(\text{incorrect decision}) \\ &= \Pr(\text{reject } \mathrm{H}_1 \mid \mathrm{H}_1)\Pr(\mathrm{H}_1) + \Pr(\text{reject } \mathrm{H}_2 \mid \mathrm{H}_2)\Pr(\mathrm{H}_2), \end{aligned} \qquad (4.7)$$

where $\Pr(\mathrm{H}_1)$ should be read as $\Pr(\mathrm{H}_1$ true), and, analogously, $\Pr(\mathrm{H}_2)$ as $\Pr(\mathrm{H}_2$ true). The quantity $\varepsilon$ can be used for designing a test in exactly the same way as $\alpha$ was used above: determine a 'significance level' $\varepsilon$, say $\varepsilon = .05$, and let the test statistic be $\lambda$. In the experiment at hand, the prior probabilities $\Pr(\mathrm{H}_1)$ and $\Pr(\mathrm{H}_2)$ are both equal to .5. As the critical value k is given, a search can be undertaken for that value of n, for which $\Pr[\lambda \geq \mathrm{k} \mid \mathrm{n}, \Pr(\mathrm{H}_1) = .5] = \varepsilon$.

If, in (4.7), $\Pr(\mathrm{H}_0$ true$) = 1$, then $\varepsilon = \alpha$. Furthermore, in the experiment at hand, where $\Pr(\mathrm{H}_1) = \Pr(\mathrm{H}_2) = .5$, it follows from (4.7) that $\varepsilon = .5(\alpha + \beta)$, where $\alpha$ and $\beta$ are the type I and type II errors, respectively. As the IOI represents the probability of a correct decision, and $\varepsilon$ represents the probability of an incorrect decision, the sample size needed for $\varepsilon = .05$ is the same sample size that is needed for an IOI of .95.

Below, it will be investigated whether the sample size $\mathrm{n}_1$, needed to get an $\mathrm{IOI}_{AA'}$ of .95, would be larger than the sample size $\mathrm{n}_2$ needed to get an $\mathrm{IOI}_{AB'}$ of .95. These sample sizes are not analytically derived; they will be inferred by means of linear interpolation of the outcomes of several simulations.

### 4.3.5 Procedure

The sample size N is fixed and equal to 1000. The sample size n will be varying between 1 and 500. The IOI will be calculated for two test lengths ($\mathrm{L} = 4$ and $\mathrm{L} = 8$) and nine sample sizes ($\mathrm{n} = 1, 10, 25, 50, 100, 150, 200, 250$ and $500$).

Algorithmically the design of the entire simulation study is given by:

1. Collect an empirical data set.
2. Consider the PCM as family $\mathcal{P}$, and fit a member of this family to the data. The fitted member of $\mathcal{P}$ will be denoted as model P and it will be called the simulation model.
3. Simulate a sample of size $\mathrm{N} = 1000$ from model P.

4. Fit the PCM ($\mathcal{P}$), the GRM ($\mathcal{G}$) and the SM ($\mathcal{S}$) to the simulated data. The fitted models will be denoted as models P$'$, G$'$ and S$'$, respectively. They will be referred to as the estimated models.

5. Let L = 4.

6. Let n = 1.

7. Simulate a sample of size ($n = 1, L = 4$) from all 4 models P, P$'$, G$'$ and S$'$. Classify each pair of signals PP$'$, PG$'$ and PS$'$ according to the method described in Section 4.3.1.

8. Repeat Step 7 100 times; the percentages of correct classifications are point estimates of the three IOIs.

9. Repeat Steps 7⇔8 10 times and calculate the means and standard deviations of the three resulting sets of point estimates.

10. Repeat Steps 6 - 9 for the other eight values of n (i.e. n = 10, 25 through to 500).

11. Repeat Steps 5 - 10 for L = 8.

12. Perform Steps 3 - 11 once again and compare the two results.

13. Repeat Steps 2 - 12 with the GRM in the role of the PCM.

14. Repeat Steps 2 - 12 with the SM in the role of the PCM.

The results obtained in Step 11 depend heavily on the estimated models P$'$, G$'$ and S$'$, obtained in Step 4, which are based on the data simulated in Step 3. Step 12 is performed to get an indication of the robustness of the results with respect to the sampling of the data in Step 3. In total, 324 IOIs will be calculated: 2 (L's) × 9 (n's) × 3(simulation models) × 3(estimated models) × 2(repetitions).

The use of an empirical data set in Step 1 is to ensure the use of realistic parameter values in the simulation to follow. The data set used in this paper consists of the answers of 362 psychology students to 8 items of the Dominance scale of a Dutch revised version (Van der Ark, 1994) of the Personality Research Form-E (Jackson, 1984). The revised version is a rating scale and has 5 response categories

The items used in the test of length 4 were randomly selected from the 8 items constituting the original test: they were items 1,2,4 and 7.

The model estimations in Steps 2 and 4 are done by marginal maximum likelihood estimation (Bock and Aitkin, 1981), using a standard normal ability distribution. The GRM and the PCM were estimated with the MULTILOG program, version 5.1 (Thissen, 1988). The estimates for the SM were calculated using BILOG-MG (Zimowski et al., 1996). In order to estimate the SM with BILOG-MG each score on a polytomous item was first expanded as follows: scores $0, 1, 2, 3$ and 4 became $0ccc, 10cc, 110c, 1110$ and 1111, respectively, where the symbol $c$ indicates a missing observation. BILOG-MG allows for missing observations. It can be easily verified that fitting the Rasch model (Rasch, 1960), which was defined in Equation 4.2, to the 40 binary variables thus constructed, will yield the proper SM estimates. In the simulations in Steps 3 and 7 a standard normal ability distribution was used as well. The likelihoods necessary in the evaluation of the statistics $\lambda_1$ and $\lambda_2$ in Step 7 are also marginal likelihoods (see Equation 4.6); here too the density of $\theta$ is taken to be standard normal. The integrals are evaluated using Gauss-Hermite quadrature with 9 quadrature points (Press, Teukolsky, Vetterling and Flannery, 1992). Finally, all random numbers were generated with the generator developed by Wichmann and Hill (1982).

### 4.3.6 Results

The average IOIs and their standard errors obtained from the 10 replications performed in Step 9, are reported in Table 4.1. From this table the following can be concluded: (a) the IOI increases with L; (b) the IOI increases with n; (c) the standard errors decrease with n, which is probably due to both a ceiling effect and the increasing sample size.

Starting with a general look at the table, it seems that it is easiest to distin-

Table 4.1: **Average and standard deviation of the IOI over 10 replications of 100 classification trials** ( ×100).

| **Simulation model** | | **PCM** | | | **GRM** | | | **SM** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Estimated model** | | **P’** | **G’** | **S’** | **P’** | **G’** | **S’** | **P’** | **G’** | **S’** |
| **L=4** | n=1 | 54 | 54 | 55 | 55 | 56 | 55 | 54 | 56 | 53 |
| | | 7 | 4 | 5 | 6 | 4 | 4 | 2 | 5 | 5 |
| | 10 | 56 | 61 | 65 | 67 | 62 | 65 | 68 | 69 | 63 |
| | | 6 | 8 | 6 | 3 | 6 | 6 | 6 | 2 | 5 |
| | 25 | 60 | 70 | 74 | 72 | 65 | 71 | 75 | 73 | 63 |
| | | 6 | 4 | 5 | 6 | 4 | 6 | 7 | 6 | 6 |
| | 50 | 64 | 75 | 80 | 83 | 75 | 77 | 84 | 80 | 75 |
| | | 5 | 5 | 4 | 3 | 4 | 4 | 4 | 3 | 4 |
| | 100 | 68 | 84 | 87 | 88 | 82 | 89 | 93 | 86 | 82 |
| | | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 4 | 2 |
| | 150 | 71 | 90 | 93 | 93 | 86 | 91 | 96 | 92 | 88 |
| | | 4 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 2 |
| | 200 | 75 | 91 | 95 | 95 | 86 | 94 | 97 | 94 | 90 |
| | | 3 | 3 | 2 | 1 | 4 | 2 | 1 | 3 | 3 |
| | 250 | 77 | 94 | 97 | 96 | 92 | 95 | 98 | 96 | 92 |
| | | 3 | 3 | 2 | 2 | 3 | 2 | 2 | 2 | 2 |
| | 500 | 84 | 98 | 100 | 99 | 98 | 99 | 100 | 99 | 98 |
| | | 5 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| **L=8** | n=1 | 55 | 58 | 57 | 57 | 55 | 58 | 59 | 58 | 53 |
| | | 6 | 4 | 5 | 4 | 6 | 6 | 5 | 4 | 4 |
| | 10 | 68 | 71 | 75 | 70 | 64 | 71 | 75 | 70 | 65 |
| | | 6 | 5 | 4 | 3 | 4 | 3 | 4 | 4 | 6 |
| | 25 | 71 | 79 | 85 | 83 | 72 | 81 | 89 | 84 | 75 |
| | | 3 | 6 | 4 | 3 | 4 | 4 | 4 | 4 | 3 |
| | 50 | 80 | 90 | 93 | 89 | 78 | 89 | 95 | 91 | 82 |
| | | 2 | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 3 |
| | 100 | 88 | 96 | 98 | 96 | 86 | 96 | 99 | 98 | 91 |
| | | 3 | 2 | 1 | 2 | 3 | 2 | 1 | 1 | 3 |
| | 150 | 92 | 99 | 100 | 98 | 91 | 98 | 100 | 99 | 94 |
| | | 3 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 2 |
| | 200 | 94 | 99 | 100 | 99 | 94 | 100 | 100 | 100 | 97 |
| | | 2 | 2 | 1 | 1 | 3 | 1 | .3 | 1 | 2 |
| | 250 | 98 | 100 | 100 | 100 | 96 | 100 | 100 | 100 | 98 |
| | | 1 | 0 | 0 | 1 | 2 | .3 | .3 | 1 | 2 |
| | 500 | 99 | 100 | 100 | 100 | 99 | 100 | 100 | 100 | 100 |
| | | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | .4 |

P’, G’, S’: Fitted members from the PCM, the GRM and the SM, respectively.

L=test length. n=sample size. Numbers in each second row are standard deviations

Table 4.2: **Sample size needed for the IOI to exceed** .95 (interpolated values)

| **Simulation model:** | **PCM** | | | **GRM** | | | **SM** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Estimated model:** | **P'** | **G'** | **S'** | **P'** | **G'** | **S'** | **P'** | **G'** | **S'** |
| **Replication 1** | | | | | | | | | |
| L=4 | $> 500$ | 313 | 200 | 200 | 375 | 250 | 134 | 225 | 375 |
| L=8 | 213 | 92 | 70 | 93 | 225 | 93 | 50 | 79 | 167 |
| **Replication 2** | | | | | | | | | |
| L=8 | 250 | 92 | 70 | 92 | 249 | 99 | 49 | 84 | 167 |

P', G', S': Fitted members from the PCM, the GRM, and the SM, respectively.

guish between the SM and the PCM. The columns pertaining to this comparison (Columns 3 and 7) nearly always have the highest IOIs in any single row. Furthermore, Columns 4 and 6 contain virtually the same values, indicating that it is about equally difficult to distinguish the GRM from the PCM as it is to distinguish the GRM from the SM. However, the values in Column 8 are consistently a little higher than those in Column 2. It may also be noted that Columns 1,5 and 9 contain the lowest values for each simulation model, so that it seems more difficult to distinguish between two members from the same family than between two members from different families. With respect to Columns 1,5 and 9, it furthermore appears that Column 9 is usually the highest of these three, indicating that it is easier to distinguish the SM from its own than it is to distinguish the GRM or the PCM from its own.

Using linear interpolation, the values of n needed for the IOI to exceed .95 have been calculated from the numbers in Table 4.1. These values are reported in the first two rows of Table 4.2. It can be seen there that, for L = 8, $\text{IOI}_{PG'}$ and $\text{IOI}_{PS'}$ become higher than .95 for n = 92 and n = 70, respectively. The $\text{IOI}_{PP'}$, on the other hand, will only become higher than .95 for n = 213. Exactly the same pattern holds in the second three columns where the GRM acts as the simulation model: $\text{IOI}_{GP'}$ and $\text{IOI}_{GS'}$ both become larger than .95 for n = 93; whereas $\text{IOI}_{GG'}$ becomes larger than .95 at n = 225 only. In the last three columns, when the SM is the simulation model, the situation is again similar, although all sample sizes are smaller here: $\text{IOI}_{SP'}$ becomes higher than .95 at n = 50; $\text{IOI}_{SG'}$ reaches this significance level at n = 79, and $\text{IOI}_{SS'}$ at n = 167. For L = 4 the patterns are similar, although of course the values for n are higher there.

For the results obtained in Step 12 (the robustness replication), no IOI values are presented as these are very similar to the values in Table 4.1. The sample sizes needed for the IOI to exceed .95 are also very similar in this replication; they are reported, for L = 8 only, in the bottom line of Table 4.2. It is concluded that the results seem to be rather robust with respect to the simulations performed in Step 3.

Apparently, for the 8 item tests investigated, the sample size needed to have the IOI exceed .95 is almost twice as large if the models compared are from the same family, as it is when the models are from a different family.

## 4.4 Deciding between models when the parameter values are unknown

In the previous section, the parameter values were assumed to be known. In the present section it is investigated whether it is possible to distinguish between item response models in a more realistic situation where parameter estimates have to be computed from the data at hand. Consider an $\mathrm{N} \times \mathrm{L}$ matrix $\mathsf{X}$ of scores on polytomous items. The symbol $\mathsf{X}$ is used to distinguish this matrix from the score vector $\boldsymbol{X}$. The question is which family of models in a collection of families can best be used to describe these data. The families considered are again the marginal versions of the PCM, GRM, and SM, defined in Section 4.2, with parameters $\boldsymbol{\delta}, \boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$, and scale factors $\alpha, \beta$ and $\tau$ respectively. For each family under consideration, that member is identified that best fits the data. Let these best fitting members be denoted as $\hat{\mathrm{P}}, \hat{\mathrm{G}}$, and $\hat{\mathrm{S}}$, respectively. These three best-fitting members are the models that will be actually compared. The likelihoods under each of these three models are given by:

$$
\begin{aligned}
L_{\hat{P}} &= L(\hat{\boldsymbol{\delta}}, \hat{\alpha} \mid \mathsf{X}, \mathrm{PCM}) = \Pr(\mathsf{X} \mid \hat{\boldsymbol{\delta}}, \hat{\alpha}, \mathrm{PCM}); \\
L_{\hat{G}} &= L(\hat{\boldsymbol{\gamma}}, \hat{\beta} \mid \mathsf{X}, \mathrm{GRM}) = \Pr(\mathsf{X} \mid \hat{\boldsymbol{\gamma}}, \hat{\beta}, \mathrm{GRM}); \\
L_{\hat{S}} &= L(\hat{\boldsymbol{\sigma}}, \hat{\tau} \mid \mathsf{X}, \mathrm{SM}) = \Pr(\mathsf{X} \mid \hat{\boldsymbol{\sigma}}, \hat{\tau}, \mathrm{SM}),
\end{aligned}
$$

where $\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\sigma}}, \hat{\alpha}, \hat{\beta}$ and $\hat{\tau}$ are maximum likelihood estimates of $\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}, \alpha, \beta$ and $\tau$.

### 4.4.1 Procedure

To decide between the three best-fitting models, a zero/one loss function and a maximum likelihood decision strategy are employed. As can be easily verified, this will result in deciding on the model with the largest likelihood.

The size of the type I error and the power under the alternatives are investigated as follows. Let $\boldsymbol{L}$ be the vector $(L_{\hat{P}}, L_{\hat{G}}, L_{\hat{S}})$. Furthermore, $\mathrm{R}(\boldsymbol{L})$ will be used for the vector $[\mathrm{r}(L_{\hat{P}}), \mathrm{r}(L_{\hat{G}}), \mathrm{r}(L_{\hat{S}})]$, where $\mathrm{r}(L_{\hat{P}})$ is the rank of the likelihood under model $\hat{\mathrm{P}}$, and similarly for $\mathrm{r}(L_{\hat{G}})$ and $\mathrm{r}(L_{\hat{S}})$. Note that the ranks in $\mathrm{R}(\boldsymbol{L})$ are in the order (PCM, GRM, SM). A value $\mathrm{R}(\boldsymbol{L}) = (1, 3, 2)$ therefore means that the GRM has the highest likelihood, and the PCM the smallest. Now suppose, for the sake of presentation, that $\mathrm{r}(L_{\hat{G}}) = 3$, so that $\hat{\mathrm{G}}$, and hence the GRM, is decided on. Then $\Pr[\mathrm{r}(L_{\hat{G}}) \neq 3 \mid \hat{\mathrm{G}}]$ gives the size of the type I error, and $\Pr[\mathrm{r}(L_{\hat{P}}) = 3 \mid \hat{\mathrm{P}}]$ and $\Pr[\mathrm{r}(L_{\hat{S}}) = 3 \mid \hat{\mathrm{S}}]$ give the power under the alternatives $\hat{\mathrm{P}}$ and $\hat{\mathrm{S}}$ against $\hat{\mathrm{G}}$, respectively. As the distributions of these ranks are unknown, they will be approximated by means of simulation.

The algorithm for the entire procedure is as follows:

1. Specify a type I error size $\alpha$.

2. Collect an initial sample $\mathsf{X}$ of size N.

3. For each of the three families of models under consideration, identify the

member fitting best to the data. Denote the estimated members of the PCM, the GRM, and the SM, as $\hat{\mathrm{P}}, \hat{\mathrm{G}}$, and $\hat{\mathrm{S}}$, respectively.

4. Calculate the three likelihoods $L_{\hat{P}}, L_{\hat{G}}$, and $L_{\hat{S}}$.

5. Identify the estimated model having the largest likelihood. Decide that this model best describes the data. For the sake of presentation, assume this is model $\hat{\mathrm{G}}$. The changes necessary below if either $\hat{\mathrm{P}}$ of $\hat{\mathrm{S}}$ has the largest likelihood, are obvious.

6. Simulate a random sample $\mathsf{X}^*$ of size N from $\hat{\mathrm{G}}$.

7. Fit PCM, GRM and SM to the simulated sample $\mathsf{X}^*$, and calculate the three likelihoods $L^*_{\hat{P}}, L^*_{\hat{G}}$, and $L^*_{\hat{S}}$.

8. If $\mathrm{r}(L^*_{\hat{G}}) = 3$, let $Q^* = 0$. Otherwise $Q^* = 1$.

9. Repeat Steps 6 through 8 B times, where B is some large number.

10. If $\sum_{b=1}^{B} Q^*_b > \mathrm{B}\alpha$, conclude that the type I error of the procedure is apparently larger than $\alpha$.

11. Simulate a random sample $\mathsf{X}^*$ of size N from one of the alternatives, say from $\hat{\mathrm{P}}$.

12. Fit PCM, GRM and SM to the simulated sample $\mathsf{X}^*$, and calculate the three likelihoods $L^*_{\hat{P}}, L^*_{\hat{G}}$, and $L^*_{\hat{S}}$.

13. If $\mathrm{r}(L^*_{\hat{P}}) = 3$, let $Z^* = 1$. Otherwise $Z^* = 0$.

14. Repeat Steps 11 through 13 B times.

15. Interpret $\left[\sum_{b=1}^{B} Z^*_b\right]/\mathrm{B}$ as the power under alternative $\hat{\mathrm{P}}$.

16. Repeat steps 11 through 15 for the other alternative model, $\hat{\mathrm{S}}$.

The models fitted in Steps 3, 7, and 12, are the marginal models defined in Section 4.2, and a standard normal ability distribution is assumed. The same estimation programs are used that were mentioned in Section 4.3.5. In the calculation of the likelihoods in Steps 4, 7, and 12, the standard normal distribution

Table 4.3: **Decision procedure: type I error and power. B=100.**

| | | Type I error | | Power | | | |
|---|---|---|---|---|---|---|---|
| Initial sample | R($\boldsymbol{L}$) | Dec. | $\frac{\sum_b Q_b^*}{B}$ | Alt. | $\frac{\sum_b Z_b^*}{B}$ | Alt. | $\frac{\sum_b Z_b^*}{B}$ |
| Artificial data | | | | | | | |
| PCM | (3,2,1) | $\hat{\text{P}}$ | 0 | $\hat{\text{G}}$ | .99 | $\hat{\text{S}}$ | 1.00 |
| GRM | (1,3,2) | $\hat{\text{G}}$ | 0 | $\hat{\text{P}}$ | 1.00 | $\hat{\text{S}}$ | 1.00 |
| SM | (1,2,3) | $\hat{\text{S}}$ | 0 | $\hat{\text{P}}$ | 1.00 | $\hat{\text{G}}$ | 1.00 |
| Empirical data | (1,3,2) | $\hat{\text{G}}$ | .11 | $\hat{\text{P}}$ | .98 | $\hat{\text{S}}$ | .97 |

Dec.: model decided on. Alt.: alternative model. R($\boldsymbol{L}$): vector with ranks of $(L_{\hat{P}}, L_{\hat{G}}, L_{\hat{S}})$.

$\hat{P}, \hat{G}, \hat{S}$: those members of the PCM, GRM, and SM, fitting best to the initial data.

$Q^* = 0$ if the model from which $\mathsf{X}^*$ was drawn, has rank 3; $Q^* = 1$ otherwise.

$Z^* = 1$ if the model from which $\mathsf{X}^*$ was drawn, has rank 3; $Z^* = 0$ otherwise

is also assumed. The likelihoods are evaluated using Gauss-Hermite quadrature with 21 quadrature points. The number B is taken to be 100. Finally, it may be noted that the data simulated in Step 6 constitute in fact a parametric bootstrap sample (Efron, 1982).

### 4.4.2 Results

The procedure was applied to the empirical data on the Dominance scale (see Section 4.3.5). To study the performance of the procedure, it was also applied to some artificial data sets. These artificial data again were generated from the models that were fitted to the empirical Dominance data. All artificial data sets have N = 1000. The Dominance data have N = 362.

Table 4.3 contains the results. In the first column of this table, the initial samples are labelled. In the second column, the vector R($\boldsymbol{L}$) containing the ranks of the three likelihoods for the initial sample is reported. With artificial data, the likelihood for the correct model was always largest. The third column mentions the model decided on. Column 4 contains the fraction of bootstrap samples having $Q^* = 1$. For all three artificial data sets, the type I error appeared to be very small.

The bottom line in the table contains the results of applying the procedure to the Dominance data. These data were classified as GRM data, which corresponds nicely with their being rating scale data. Because of the smaller value for N, the type I error was larger here.

In the last four columns of the table the results of the power investigation are reported. The power under the alternatives was found to be large, both for the artificial data sets and for the empirical data, even though the latter had only N = 362.

Table 4.4: **Relative frequencies of R($\boldsymbol{L}^*$) for the type I error and power investigations. B=100.**

| | Type I error | | | Power | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Initial data | Dec. | R($\boldsymbol{L}^*$) | Freq. | Alt. | R($\boldsymbol{L}^*$) | Freq. | Alt. | R($\boldsymbol{L}^*$) | Freq. |
| Artificial | | | | | | | | | |
| PCM | $\hat{P}$ | (1,2,3): | 1.00 | $\hat{G}$ | (3,2,1): | .01 | $\hat{S}$ | (1,2,3): | 1.00 |
| | | | | | (1,3,2): | .48 | | | |
| | | | | | (2,3,1): | .51 | | | |
| GRM | $\hat{G}$ | (1,3,2): | .61 | $\hat{P}$ | (3,2,1): | 1.00 | $\hat{S}$ | (1,2,3): | 1.00 |
| | | (2,3,1): | .39 | | | | | | |
| SM | $\hat{S}$ | (1,2,3): | 1.00 | $\hat{P}$ | (3,2,1): | 1.00 | $\hat{G}$ | (1,3,2): | .50 |
| | | | | | | | | (2,3,1): | .50 |
| Empirical | $\hat{G}$ | (1,2,3): | .04 | $\hat{P}$ | (1,2,3): | .01 | $\hat{S}$ | (1,2,3): | .90 |
| | | (1,3,2): | .54 | | (2,3,1): | .01 | | (1,3,2): | .01 |
| | | (2,3,1): | .35 | | (3,1,2): | .08 | | (2,1,3): | .07 |
| | | (3,2,1): | .07 | | (3,2,1): | .90 | | (2,3,1): | .01 |
| | | | | | | | | (3,2,1): | .01 |

Dec.: model decided on. Alt.: alternative model. R($\boldsymbol{L}$): vector with ranks of $(L_{\hat{P}}, L_{\hat{G}}, L_{\hat{S}})$.

$\hat{P}, \hat{G}, \hat{S}$: those members of the PCM, GRM, and SM, fitting best to the initial data.

The fractions $\left[\sum_b Q_b^*\right]/\mathrm{B}$ and $\left[\sum_b Z_b^*\right]/\mathrm{B}$, reported in Table 4.3, have been derived from the frequency distributions of $\mathrm{R}(\boldsymbol{L}^*)$. A closer look at these frequency distributions may be instructive. They are reported in Table 4.4. Concentrate on the artificial samples first. Although, with artificial initial data, the likelihood for the 'correct' model nearly always had rank 3, the ranks for the other two models were more variable. In particular, if the data $\mathsf{X}^*$ were drawn from the GRM, then in about half of these simulated samples $L_{\hat{P}}^* > L_{\hat{S}}^*$, and in the other half $L_{\hat{P}}^* < L_{\hat{S}}^*$. This holds for both the type I error investigation and for the power investigation. For data $\mathsf{X}^*$ drawn from the PCM or the SM, the value of $\mathrm{R}(\boldsymbol{L}^*)$ displayed no variation at all.

For the empirical Dominance data the same effect is present, although, as a consequence of the smaller N, to a somewhat lesser extent.

Surprisingly, it seems that if the PCM and the SM are fitted to data generated from the GRM, their likelihoods are, on average, approximately equal. This is confirmed by an inspection of the average likelihoods, which are reported in Table 4.5. The numbers reported here are averages for the loglikelihoods calculated in Steps 7 and 12. They are reported in the order PCM, GRM, SM. In the table it can be seen that indeed, when $\mathsf{X}^*$ is generated from the GRM, the averages of $\log(L_{\hat{P}})$ and $\log(L_{\hat{S}})$ are very close.

## 4.5 Conclusion and discussion

In the first part of this paper a method from signal detection theory has been used to investigate the differences between data generated by different item response models. The criterion variable was the sample size needed for a 95 percent probability of correctly distinguishing between data generated by two models. This sample size was calculated for two models belonging to the same family, and also for two models belonging to different families. For the models investigated, the former sample size came out about twice as large as the latter one. The yield of the present study therefore is that it can be concluded that the small but signif-

Table 4.5: **Means and standard deviations of the loglikelihoods for the models fitted to $\mathsf{X}^*$. $\mathbf{B} = 100$.**

| | Type I error | | Power | | | |
|---|---|---|---|---|---|---|
| Initial data | Dec. | Mean (Sd) | Alt. | Mean (Sd) | Alt. | Mean (Sd) |
| Artificial | | | | | | |
| PCM | $\hat{P}$ | $\dot{P}$ -9841 (62) | $\hat{G}$ | $\dot{P}$ -9911 (65) | $\hat{S}$ | $\dot{P}$ -9909 (63) |
| | | $\dot{G}$ -9858 (62) | | $\dot{G}$ -9893 (64) | | $\dot{G}$ -9890 (63) |
| | | $\dot{S}$ -9873 (63) | | $\dot{S}$ -9911 (64) | | $\dot{S}$ -9874 (62) |
| GRM | $\hat{G}$ | $\dot{P}$ -9798 (53) | $\hat{P}$ | $\dot{P}$ -9789 (58) | $\hat{S}$ | $\dot{P}$ -9804 (63) |
| | | $\dot{G}$ -9776 (53) | | $\dot{G}$ -9809 (59) | | $\dot{G}$ -9786 (62) |
| | | $\dot{S}$ -9796 (53) | | $\dot{S}$ -9825 (59) | | $\dot{S}$ -9769 (62) |
| SM | $\hat{S}$ | $\dot{P}$ -9841 (62) | $\hat{P}$ | $\dot{P}$ -9820 (56) | $\hat{G}$ | $\dot{P}$ -9825 (69) |
| | | $\dot{G}$ -9821 (62) | | $\dot{G}$ -9838 (57) | | $\dot{G}$ -9806 (70) |
| | | $\dot{S}$ -9805 (61) | | $\dot{S}$ -9852 (58) | | $\dot{S}$ -9825 (71) |
| Empirical | $\hat{G}$ | $\dot{P}$ -3540 (36) | $\hat{P}$ | $\dot{P}$ -3558 (35) | $\hat{S}$ | $\dot{P}$ -3547 (36) |
| | | $\dot{G}$ -3532 (37) | | $\dot{G}$ -3564 (34) | | $\dot{G}$ -3539 (36) |
| | | $\dot{S}$ -3539 (37) | | $\dot{S}$ -3569 (35) | | $\dot{S}$ -3534 (36) |

Dec.: model decided on. Alt.: alternative model. $\hat{P}, \hat{G}, \hat{S}$: models fitted to the initial samples.

$\dot{P}, \dot{G}, \dot{S}$ : models fitted to the simulated samples $\mathsf{X}^*$.

icant differences found by Maydeu-Olivares et al. (1994), are large enough to be of practical relevance.

The models selected for the comparison were those members from each family under consideration that fitted best to an initial data set. A baseline procedure was used to interpret the values of the observed IOIs. It would of course be possible to select models for the comparison by means of another procedure. For instance, one might start with a member from one of the families, and then select members from the other families by means of minimizing the area between the two sets of curves (see e.g. De Vries, 1988). In this case no baseline procedure could be used.

In the second part of the paper a procedure was investigated for deciding upon a family of item response models, when the parameters still have to be estimated from the data. At first sight, the procedure seems to perform well, but it should certainly be examined in some more detail before definite conclusions are drawn. It would be wise to investigate what happens with altogether different test specifications, that is, with other artificial data sets, and with other empirical data as well. The effects of test length and of sample size could be investigated, and also the effect of including a discrimination parameter in the models.

A drawback of the procedure is that it is implicitly assumed that the 'correct' family of models is among the families considered. An example may clarify this. Suppose the 'correct' model for a data set is the SM, but, not knowing this, one wants to decide between the PCM and the GRM. Then certainly one of the two likelihoods will be larger, and upon investigating the type I error and the power, one may well be led to have faith in the decision. Even though the decision is wrong. So it would be interesting to try and construct a statistical test for model choice. Suppose one hypothesizes that the GRM would be most suited for a certain data set. This hypothesis might be testable against e.g. the PCM or the SM by means of a likelihood ratio statistic. The function of the null and alternative hypothesis in this context, however, would call for a careful consideration. But if these problems were solved, one would have a much more versatile procedure than the crude method proposed in this chapter. When the 'correct' model is not included in the investigation, a likelihood ratio test might well be better at detecting that something is wrong than the decision procedure investigated in this chapter.

Finally, it was found that for data generated according to the GRM, the likelihoods under the best-fitting members of the PCM and the SM were, on average, approximately equal. It would be interesting to find out whether this phenomenon occurs with other GRM data as well, and if so, to find an explanation for it.

# Chapter 5

# Monte Carlo estimation of the conditional Rasch model[1]

## Abstract

In order to obtain conditional maximum likelihood estimates, the conditioning constants are needed. Geyer and Thompson (1992) proposed a Markov chain Monte Carlo method that can be used to approximate these constants when they are difficult to calculate exactly. In the present paper, their method is applied to the conditional estimation of person parameters in the Rasch model. The results obtained with the Monte Carlo method can be very accurate, but in that case the method is rather slow. However, for only slightly less precise results the Monte Carlo method can be faster than the exact calculations. For the estimation of the ability parameters in a 5 item test taken by 1000 persons the Monte Carlo method took about half the time needed for the exact calculations; and still the difference between two corresponding estimates was less than 1 percent of the associated standard error in all cases.

Key words: Conditional maximum likelihood estimation, Markov chain Monte Carlo methods, Rasch model, item response theory.

[1] This paper has been published in *Computational Statistics, 13*, 1998, p.185-211. The author wishes to thank dr. Alun Thomas from the University of Bath (presently at Myriad Genetics Inc., Salt Lake City) for his stimulating comments and advice. Financial support for this paper was granted by the Netherlands Organization for Scientific Research (NWO) and by the Dr. Catharina van Tussenbroekfonds.

## 5.1 Introduction

In this paper, an estimation method developed by Geyer and Thompson (1992) is applied to the Rasch model (Rasch, 1960). This method provides a Monte Carlo alternative to exact calculation of the normalizing denominators in a likelihood. Geyer and Thompson showed that in the exponential family a quantity which is proportional to the normalizing denominator can be expressed as an expectation with respect to a certain distribution. Upon simulating from this distribution, the observed sample mean can serve as an estimate of the proportional quantity. Inserting this estimate into the likelihood then allows one to maximize the approximate likelihood, as the proportionality constant does not depend upon the parameters to be estimated. The method will be described in detail in the next section.

This procedure can be directly applied to conditional likelihoods as well. In fact, Geyer and Thompson observe an interesting parallel between maximum likelihood and computation of posterior likelihoods: they note that "almost any Bayesian computation can be carried out via Gibbs sampling, and almost any maximum likelihood computation can be done by some Markov chain Monte Carlo scheme", such as will be described in section 5.2. In this paper their scheme will be applied to the conditional estimation of person parameters in the Rasch model (Rasch, 1960), which is a probabilistic model for intelligence and attainment tests. The purpose of administering a test is to infer something about the test-taker's value or location on some latent trait, such as intelligence. Let the latent trait of interest be denoted by the symbol $\theta$; it will be referred to as ability. Assume the test is taken by N persons, and it consists of M dichotomous items, scored 1 if answered correctly and 0 otherwise.

In the Rasch model the probability of a correct response of person $i$ with ability $\theta_i$ on an item with difficulty level $\delta$ is assumed to be given by

$$\Pr(X_i = 1; \theta_i, \delta) = \frac{\exp(\theta_i \Leftrightarrow \delta)}{1 + \exp(\theta_i \Leftrightarrow \delta)}, \tag{5.1}$$

where $\delta$ is an item parameter expressing the item's location on the same scale as $\theta_i$, and the variable $X_i = x_i$, is used to indicate the response of the $i$'th person.

Evidently the probability of a correct response is assumed to increase with ability. Note that the model is not identified without a restriction on the parameters: if $\delta^* = \delta + k$ and $\theta_i^* = \theta_i + k$ for all $i$, then $\Pr(X_i = 1; \theta_i^*, \delta^*) = \Pr(X_i = 1; \theta_i, \delta)$. Usually one of the parameters, or their average, is constrained to be 0.

Under the Rasch model there are two sets of parameters: the person abilities and the item difficulties. These two sets of parameters cannot be simultaneously estimated. Usually first the deltas are estimated, and these are then inserted into the likelihood and treated as known in order to obtain maximum likelihood (ML) estimates of the thetas. One way of estimating the item parameters is marginal maximum likelihood estimation or MML (see e.g. Bock and Aitkin, 1981). Here $\boldsymbol{\theta} = (\theta_1 \ldots \theta_N)$ is considered as a random variable, and it is integrated out of the likelihood. This will result in a likelihood which is independent of $\boldsymbol{\theta}$, and which can therefore be maximized w.r.t. the $\delta$'s. The alternative method is CML or conditional maximum likelihood estimation: let $T_i$ be the total score $\sum_{j=1}^{M} X_{ji}$, where the index $j$ refers to the M items in the test. Then $T_i$ is sufficient for $\theta_i$, and conditioning on the $T_i$'s will also result in a likelihood which is independent of the abilities. Andersen (1970; 1972) proved that the CML estimator is consistent. Again the estimated $\delta$'s are then treated as known in order to obtain ML estimates of $\boldsymbol{\theta}$.

When there is an interest in obtaining estimates of the $\delta$'s (e.g. for future use of the test items) this is a reasonable way of proceeding. However, when the ultimate goal is to estimate $\boldsymbol{\theta}$, the current practice seems rather like a detour: in this case really the $\delta$'s should be considered as nuisance parameters instead of $\boldsymbol{\theta}$, and preferably the likelihood should be maximized w.r.t. $\boldsymbol{\theta}$ in the first place. Two more topics can be raised in this context. First, the consequences of substituting $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$ in the ML procedure are not exactly known. It is conceivable that there could be at least an influence on the (asymptotic) variance of $\hat{\boldsymbol{\theta}}$. The terms $\hat{\delta}$-based MLE and $\delta$-based MLE will be used to refer to the maximum likelihood estimators of $\boldsymbol{\theta}$, using estimated and true values of the item parameters, respectively. The second point to be mentioned is that although the $\delta$-based MLE of $\boldsymbol{\theta}$ is consistent, for finite M it is known to be biased away from 0 (Lord, 1983; Warm, 1989). If $\hat{\boldsymbol{\delta}}$ is biased this might possibly propagate extra bias into $\hat{\boldsymbol{\theta}}$. These topics will not be extensively pursued in the present paper, but they are raised in order to underline the need for an estimation procedure that is independent of $\boldsymbol{\delta}$ and $\hat{\boldsymbol{\delta}}$.

Although the Rasch model is symmetric with respect to $\theta$ and $\delta$, it is not easy just to reverse the marginal or conditional estimation procedures and maximize the likelihood directly with respect to $\boldsymbol{\theta}$. Integrating out item difficulty would more or less imply that the item parameters are considered as random. This is an assumption few researchers are willing to make (see, however, Rigdon and Tsutakawa, 1986). On the other hand, the item total $S = \sum_{i=1}^{N} X_i$ is sufficient for $\delta$. So at least theoretically it is possible to conditionally estimate the $\theta$'s. However, in practice N >> M, that is, usually there are far more persons than items. It may then become time-consuming to calculate the denominators of the conditional likelihood.

In the present paper, a Monte Carlo method for the calculation of the denominators will be investigated. The mathematical form of these denominators in the conditional Rasch model will be presented in section 5.1.1. Several algorithms for their calculation will be described in section 5.1.2. All of these algorithms employ recursive relations. Applied to large numbers of parameters, they may therefore become either time-consuming or inaccurate. So it will be interesting to apply the Geyer and Thompson method to the CML estimation of person parameters in the Rasch model, and to compare the results to some of the algorithms for exact estimation. Section 5.2 gives a description of the Monte Carlo processes operating in the Geyer and Thompson procedure. The estimation equations will be examined more closely in section 5.3. Section 5.4 contains the results of some tests on the performance of the Monte Carlo method. In section 5.5 some parameter estimates will be presented and these will be compared to exact CML estimates, both with respect to accuracy and speed of the calculations.

### 5.1.1 The Conditional Rasch Model

The derivations in this section are for the conditional estimation of the person parameters in the Rasch model; making the appropriate changes, the formulae for the conditional estimation of the item parameters can be analogously obtained. Consider a test consisting of M items, taken by N persons, the objective being the estimation of the person parameters. The data matrix then is of size M × N, i.e. it is wider than long. We will be interested in the vector-valued variable $\boldsymbol{X} = \boldsymbol{x}$, where $\boldsymbol{X} = (X_1 \ldots X_i \ldots X_N)$ is the N-vector of the responses of all persons on a single item. The variable $\boldsymbol{X}$ will be denoted as the item score pattern. Then $X_i = x_i$, with $x_i = 0, 1$ is the variable denoting the response given by person $i$, and the item total $S$ was defined as $\sum_i^N X_i$. Furthermore, $\boldsymbol{\theta}$ is the vector of abilities $(\theta_1 \ldots \theta_i \ldots \theta_N)$, and the parameter $\delta$ indicates the unknown item difficulty.

Assuming independence between responses given by different persons, in the Rasch model the probability of observing a vector $\boldsymbol{X}$ is a function of the item difficulty $\delta$; it can be derived from (5.1) as

$$\begin{aligned} \Pr(\boldsymbol{X} = \boldsymbol{x}; \delta, \boldsymbol{\theta}) &= \prod_{i=1}^{N} \frac{[\exp(\theta_i \Leftrightarrow \delta)]^{X_i}}{1 + \exp(\theta_i \Leftrightarrow \delta)} \\ &= \frac{\exp(\sum_{i=1}^N \theta_i X_i \Leftrightarrow S\delta)}{\prod_{i=1}^N [1 + \exp(\theta_i \Leftrightarrow \delta)]}. \end{aligned} \tag{5.2}$$

The probability of observing an item total $S = s$ is equal to

$$\begin{aligned} \Pr(S = s; \delta, \boldsymbol{\theta}) &= \sum_{\mathbf{X}:S(\mathbf{X})=s} \Pr(\boldsymbol{X} = \boldsymbol{x}; \delta, \boldsymbol{\theta}) \\ &= \sum_{\mathbf{X}:S(\mathbf{X})=s} \frac{\exp(\sum_{i=1}^N X_i \theta_i \Leftrightarrow S\delta)}{\prod_{i=1}^N [1 + \exp(\theta_i \Leftrightarrow \delta)]}, \end{aligned} \tag{5.3}$$

where the summation is over all item score patterns leading to the same total $S = s$. The conditional probability of one score pattern $\boldsymbol{X}$ given $S$ is therefore equal to the division of (5.2) by (5.3):

$$\Pr(\boldsymbol{X} \mid S; \delta, \boldsymbol{\theta}) = \frac{\exp\left(\sum_{i=1}^N X_i \theta_i\right)}{\sum_{\mathbf{X}:S(\mathbf{X})=s} \exp\left(\sum_{i=1}^N X_i \theta_i\right)}. \tag{5.4}$$

Note that indeed this likelihood is independent of $\delta$ and can therefore be written as $\Pr(\boldsymbol{X} \mid S; \boldsymbol{\theta})$. The denominators in (5.4), there is one for each value of $S$, are the well-known elementary symmetric functions and in item response theory these are usually denoted by $\gamma_s(\boldsymbol{\theta})$:

$$\gamma_s(\boldsymbol{\theta}) \quad \overset{d}{=} \quad \sum_{\mathbf{X}:S(\mathbf{X})=s} \exp\left(\textstyle\sum_{i=1}^N X_i\theta_i\right). \tag{5.5}$$

With this notation the conditional probability in (5.4) can be rewritten as

$$\Pr(\boldsymbol{X} \mid S; \boldsymbol{\theta}) = \frac{\exp\left(\sum_{i=1}^N X_i\theta_i\right)}{\gamma_s(\boldsymbol{\theta})}, \tag{5.6}$$

and the corresponding conditional log likelihood is then given by

$$\log L(\boldsymbol{\theta} \mid S; \boldsymbol{X}) = \textstyle\sum_{i=1}^N X_i\theta_i \Leftrightarrow \log \gamma_s(\boldsymbol{\theta}).$$

Let $\mathrm{M}_s$ be the number of items in the sample having the value $s$ on the statistic $S$, and note that all items with the same total score have the same value for $\gamma_s(\boldsymbol{\theta})$. Also, given $\theta_i$, independence of responses on different items is assumed; therefore the log-likelihood for the whole sample becomes

$$\begin{aligned}\log L(\boldsymbol{\theta} \mid \boldsymbol{S}; \mathsf{X}) \quad &= \quad \sum_{j=1}^{M}\sum_{i=1}^{N} \theta_i X_{ji} \Leftrightarrow \sum_s \mathrm{M}_s \log \gamma_s(\boldsymbol{\theta}) \\ \\ &= \quad \textstyle\sum_{i=1}^N \theta_i T_i \Leftrightarrow \displaystyle\sum_s \mathrm{M}_s \log \gamma_s(\boldsymbol{\theta}),\end{aligned} \tag{5.7}$$

where the index $j$ refers to the M items in the sample, $T_i$ is the total score $\sum_{j=1}^M X_{ji}$, $\boldsymbol{S}$ is a column vector of observed item totals and $\mathsf{X}$ denotes the entire data matrix. The symbol $\mathsf{X}$ is used to distinguish this matrix from the score vector $\boldsymbol{X}$. Note that (5.7) still depends on the data through $T_i$. Zeroing the partial derivatives of (5.7) with respect to $\boldsymbol{\theta}$ will yield the CML solution equations. These partial derivatives are given by

$$\frac{\partial \log L(\boldsymbol{\theta} \mid \boldsymbol{S}; \mathsf{X})}{\partial \theta_k} = T_k \Leftrightarrow \sum_s \mathrm{M}_s \frac{\gamma_{s-1}^{(k)}(\boldsymbol{\theta}) \exp(\theta_k)}{\gamma_s(\boldsymbol{\theta})}, \tag{5.8}$$

in which the numerator is equal to $\partial\gamma_s(\boldsymbol{\theta})/\partial\theta_k$, with $\gamma_{s-1}^{(k)}(\boldsymbol{\theta})$ a symmetric function for the set of persons not containing person $k$. For example, if there were 3 persons, some of the gamma functions would be

$$\begin{aligned}\gamma_2(\boldsymbol{\theta}) &= \exp(\theta_1+\theta_2)+\exp(\theta_1+\theta_3)+\exp(\theta_2+\theta_3),\\ \gamma_1^{(3)}(\boldsymbol{\theta}) &= \exp(\theta_1)+\exp(\theta_2),\end{aligned}$$

so that indeed $\partial\gamma_2(\boldsymbol{\theta})/\partial\theta_3 = \gamma_1^{(3)}(\boldsymbol{\theta})\exp(\theta_3)$.

### 5.1.2 Algorithms for the exact calculation of $\gamma_s(\boldsymbol{\theta})$

In order to be able to compare the Monte Carlo algorithm to the algorithms for exact calculation, some of these latter algorithms will now be described. For an efficient description let, in this section only,

$$\xi_i = \exp(\theta_i) \qquad \text{for} \quad i = 1, 2, \ldots \mathrm{N}.$$

From (5.5) the gamma functions can be adapted to the transformed metric:

$$\gamma_s(\boldsymbol{\xi}) = \sum_{\mathbf{X}:S(\mathbf{X})=s} \prod_i \xi_i^{X_i},$$

where $\boldsymbol{\xi} = (\xi_1 \ldots \xi_i \ldots \xi_N)$. The order of the gamma function $\gamma_s(\boldsymbol{\xi})$ is defined as the length of the vector $\boldsymbol{\xi}$.

The sum and the difference algorithms are clearly described in Fischer (1974). The sum algorithm starts from $\gamma_0(\xi_1, \xi_2, \ldots, \xi_k) = 1$ for all $k$, and $\gamma_1(\xi_1) = \xi_1$; after that it uses

$$\gamma_r(\xi_1 \ldots \xi_p) = \xi_p \gamma_{r-1}(\xi_1 \ldots \xi_{p-1}) + \gamma_r(\xi_1 \ldots \xi_{p-1}) \tag{5.9}$$

for $p = 2, 3, \ldots, \mathrm{N}$ and $r = 1, 2, \ldots, p$, to build up the set of gamma functions of order N via all the sets of order $2, 3, \ldots, \mathrm{N} \Leftrightarrow 1$. Let $S_{max}$ be the maximum value observed for $S$; then of course, if $S_{max} < \mathrm{N}$, there is no need to build up the gamma function in (5.9) for $p > S_{max}$. The sum algorithm is numerically accurate and stable, and it can be used for the estimation of $\boldsymbol{\theta}$, but it is time-consuming. Presently calculating a gamma function of order 700 will take about 1 or 2 seconds on a Dos Pentium machine. This may seem little, but note that in order to evaluate (5.8), also the gamma functions $\gamma_{s-1}^{(k)}(\boldsymbol{\xi})$ are needed $k = 1, 2, \ldots, \mathrm{N}$; in practice however (see also Section 5.3.3) only $\mathrm{N}'$ of these functions needed, where $\mathrm{N}'$ is the number of different values in $\boldsymbol{T}$, not counting 0 and M. As an example, for a test consisting of 5 items there are 4 different values possible for $T_i$, so in all, with the sum algorithm, 5 sets of gamma functions have to be evaluated, irrespective of the value of N. These 5 sets of gamma functions have to be evaluated in each iteration of the maximization routine. Supposing some 7 iterations are necessary and each evaluation takes 1.5 seconds, the entire estimation process will take about 50 seconds. So indeed a reduction of this time would be interesting.

The difference algorithm is faster than the sum algorithm. However, Verhelst et al. (1984) showed that this algorithm is not accurate enough for the conditional estimation of $\boldsymbol{\theta}$ and therefore it will not be further examined in the present paper. These latter authors proved a 'union property', which is fully exploited in an extended algorithm proposed by Liou (1994). The extended algorithm is an extension of the difference algorithm and it computes $\gamma_s(\boldsymbol{\xi})$ via the sum algorithm; the functions $\gamma_{s-1}^{(k)}(\boldsymbol{\xi})$ however are then obtained from $\gamma_s(\boldsymbol{\xi})$ upon starting from $\gamma_0^{(k)}(\boldsymbol{\xi}) = 1$ and using

$$\gamma_r^{(k)}(\boldsymbol{\xi}) = \gamma_r(\boldsymbol{\xi}) \Leftrightarrow \xi_k \gamma_{r-1}^{(k)}(\boldsymbol{\xi})$$

for $r = 1, 2, \ldots, \mathrm{N} \Leftrightarrow 1$ and $k = 1, 2, \ldots, \mathrm{N}$. Again, in practice only $\mathrm{N}'$ instead of N functions $\gamma_r^{(k)}(\boldsymbol{\xi})$ will have to be calculated. In order to minimize rounding errors, a 'backward' equation, starting from $\gamma_N(\boldsymbol{\theta})$ can be used in the second half of the calculations. Details can be found in Liou (1994), who investigated the accuracy of the extended algorithm for tests consisting of up to 60 items. For tests of this length it proved accurate, and about M times faster than the sum algorithm, where M is the number of parameters to be estimated. Hence, for CML estimation of the item parameters in tests of this length the extended algorithm is undoubtedly the best choice. However, the behavior of this algorithm in the case of estimating $\boldsymbol{\theta}$, i.e. in estimating a large number of parameters, has not yet been investigated. Some results for this case will be presented below in Section 5.5.2.

## 5.2 The Monte Carlo estimation method

Applying the theory developed by Geyer and Thompson (1992), it will in this section be shown that Markov chain Monte Carlo (MCMC) methods can be used to approximate a quantity that is proportional to the gamma functions. Let the probability density function for one observation in the conditional formulation of the Rasch model be known as $\mathrm{f}_s(\boldsymbol{X};\boldsymbol{\theta})$; then the corresponding conditional likelihood of $\boldsymbol{\theta}$ as a function of $\boldsymbol{X}$ can be written as $\mathrm{f}_s(\boldsymbol{\theta};\boldsymbol{X})$. In this notation the conditioning variable $S$ has moved from behind the bar to a subscript on f. We therefore have

$$\mathrm{f}_s(\boldsymbol{\theta};\boldsymbol{X}) \quad \overset{d}{=} \quad L(\boldsymbol{\theta} \mid S;\boldsymbol{X}) = \frac{\exp\left(\sum_{i=1}^{N} X_i\theta_i\right)}{\gamma_s(\boldsymbol{\theta})}.$$

If $\boldsymbol{\psi}$ were another set of parameters, then trivially, using the definition of $\gamma_s(\boldsymbol{\theta})$ given in (5.5),

$$\gamma_s(\boldsymbol{\theta}) = \sum_{\mathbf{X}:S(\mathbf{X})=s} \exp\left(\sum_{i=1}^{N} X_i\theta_i\right)\frac{\gamma_s(\boldsymbol{\psi})}{\exp\left(\sum_i X_i\psi_i\right)}\mathrm{f}_s(\boldsymbol{\psi};\boldsymbol{X}).$$

This is the formula for importance sampling (see e.g. Ripley, 1987); however, the purpose is not to estimate $\gamma_s(\boldsymbol{\theta})$ but the parameters $\boldsymbol{\theta}$, i.e. to maximize the loglikelihood function given in (5.7). Theoretically, one could do this by obtaining an estimate $\gamma_s(\hat{\boldsymbol{\theta}})$ of $\gamma_s(\boldsymbol{\theta})$ by means of importance sampling, inserting the estimate into the loglikelihood function (5.7), and then maximizing that function. In that case, however, one would still have to calculate $\gamma_s(\boldsymbol{\psi})$ which is exactly what we are trying to avoid. It will therefore prove fruitful to move $\gamma_s(\boldsymbol{\psi})$ to the left hand side and obtain

$$\begin{aligned}\frac{\gamma_s(\boldsymbol{\theta})}{\gamma_s(\boldsymbol{\psi})} &= \sum_{\mathbf{X}:S(\mathbf{X})=s} \left\{\exp\left[\sum_{i=1}^{N} X_i(\theta_i \Leftrightarrow \psi_i)\right]\right\}\ \mathrm{f}_s(\boldsymbol{\psi};\boldsymbol{X}) \\ &= E_{\boldsymbol{\psi};S=s}\left\{\exp\left[\sum_{i=1}^{N} X_i(\theta_i \Leftrightarrow \psi_i)\right]\right\}.\end{aligned}$$

In other words, if

$$\mathrm{q}_s(\boldsymbol{\theta}) \quad \overset{d}{=} \quad \frac{\gamma_s(\boldsymbol{\theta})}{\gamma_s(\boldsymbol{\psi})}, \quad \text{for } s = 0, 1, \ldots, \mathrm{M},$$

then, upon simulating a random sample of size B from $\mathrm{f}_s(\boldsymbol{\psi};\boldsymbol{X}) = \mathrm{f}_s(\boldsymbol{X};\boldsymbol{\psi})$, all $\mathrm{q}_s(\boldsymbol{\theta})$'s could be estimated by the sample means:

$$\widehat{\mathrm{q}_s}(\boldsymbol{\theta}) = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\exp\left[\textstyle\sum_{i=1}^{N} X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right] \tag{5.10}$$

for any value of $\boldsymbol{\theta}$. Note that the first subscript on $X$, the $s$, is there to indicate that every simulated item score vector $\boldsymbol{X}$ belongs to a set having common total $S = s$. Defining $\log L^*$ to be

$$\begin{aligned}
\log L^*(\boldsymbol{\theta} \mid \boldsymbol{S};\mathsf{X},\boldsymbol{\psi}) &\stackrel{d}{=} \textstyle\sum_{i=1}^{N}\theta_i T_i \Leftrightarrow \displaystyle\sum_s \mathrm{M}_s \log \mathrm{q}_s(\boldsymbol{\theta}) && (5.11)\\
&= \textstyle\sum_{i=1}^{N}\theta_i T_i \Leftrightarrow \displaystyle\sum_s \mathrm{M}_s \log\gamma_s(\boldsymbol{\theta}) + \sum_s \mathrm{M}_s \log\gamma_s(\boldsymbol{\psi})\\
&= \log L(\boldsymbol{\theta} \mid \boldsymbol{S};\mathsf{X}) + \sum_s \mathrm{M}_s \log\gamma_s(\boldsymbol{\psi}),
\end{aligned}$$

note that $\log L^*(\boldsymbol{\theta} \mid \boldsymbol{S};\mathsf{X},\boldsymbol{\psi})$ attains its maximum for the same value of $\boldsymbol{\theta}$ as does $\log L(\boldsymbol{\theta} \mid \boldsymbol{S};\mathsf{X})$. So it is now possible to substitute $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ for $\mathrm{q}_s(\boldsymbol{\theta})$ in (5.11) and maximize the resulting expression

$$\begin{aligned}
\log L^*(\boldsymbol{\theta} \mid \boldsymbol{S};\mathsf{X},\boldsymbol{\psi}) &\approx \textstyle\sum_{i=1}^{N}\theta_i T_i \Leftrightarrow \displaystyle\sum_s \mathrm{M}_s \log \widehat{\mathrm{q}_s}(\boldsymbol{\theta})\\
&= \textstyle\sum_{i=1}^{N}\theta_i T_i \Leftrightarrow \displaystyle\sum_s \mathrm{M}_s \log\left\{\tfrac{1}{\mathrm{B}}\textstyle\sum_{b=1}^{B}\exp\left[\sum_{i=1}^{N} X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]\right\} && (5.12)
\end{aligned}$$

with respect to $\boldsymbol{\theta}$ to get an approximate solution to the original likelihood equations. The partial derivatives of this function are given by

$$\frac{\partial \log L^*}{\partial \theta_k} \approx T_k \Leftrightarrow \textstyle\sum_s \mathrm{M}_s \displaystyle\frac{\sum_b X_{sbk}\exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]}{\sum_b \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]}. \tag{5.13}$$

Recapitulating: the purpose is to estimate the parameters $\boldsymbol{\theta}$. Therefore, the conditional loglikelihood function in (5.7) has to be maximized, but the gamma functions figuring in its derivative are difficult to calculate. In (5.11) another function has been found that attains its maximum for the same value of $\boldsymbol{\theta}$ as does (5.7), so instead of maximizing (5.7), one could maximize (5.11). In (5.11), the incalculable quantity $\gamma_s(\boldsymbol{\theta})$ has been replaced by the quantity $\mathrm{q}_s(\boldsymbol{\theta})$, which may be obtained by means of simulation (see equation 5.10).

However, the simulations needed in order to obtain an estimate of $\mathrm{q}_s(\boldsymbol{\theta})$ have to be drawn from $\mathrm{f}_s(\boldsymbol{X};\boldsymbol{\psi})$. The denominator of this function is $\gamma_s(\boldsymbol{\psi})$, which is another gamma function. Now theoretically, one is at liberty to choose a convenient value for $\boldsymbol{\psi}$. Hence one might choose it such that $\gamma_s(\boldsymbol{\psi})$ were easy to calculate. Unfortunately, in practice this liberty is but limited; this topic will be taken up in sections 5.3.1 and 5.4.3. It will therefore be assumed that $\gamma_s(\boldsymbol{\psi})$ is just as difficult to evaluate as is $\gamma_s(\boldsymbol{\theta})$; and hence, that it is difficult to simulate from $\mathrm{f}_s(\boldsymbol{X};\boldsymbol{\psi})$ directly. Below the Metropolis (1953) algorithm will be described, which can be used to obtain the simulations in an indirect way.

### 5.2.1 Simulation of response patterns

A Markov chain is a sequence of realizations of a random variable $Z$ with the property that

$$\Pr(Z_k = z_k \mid Z_1 = z_1, \ldots, Z_{k-1} = z_{k-1}) = \Pr(Z_k = z_k \mid Z_{k-1} = z_{k-1}),$$

where the subscript $k$ denotes the ordering of the sequence in time, and $Z$ may be vector valued. The probabilities of going from one state to another in a Markov chain can be represented in a matrix P, having as entries $\mathrm{p}_{ij} = \Pr(Z_k = i \mid Z_{k-1} = j)$. The Markov chain is irreducible if it is possible to get from any state to any other state in a finite number of transitions. The states of irreducible Markov chains on finite sets of values follow a unique limiting or ergodic distribution; denoting this (discrete) distribution by $\pi$, it is given as the solution to

$$\pi \mathrm{P} = \pi.$$

(see e.g. Proth and Hillion, 1990). In words: if the transitions are made according to P, then for large N, $\Pr(Z_{k+N} = i \mid Z_k = j) \approx \pi_i$, independent of the value of $Z_k$.

In order to simulate from $\pi$, one could select a Markov chain with transition matrix P satisfying $\pi\mathrm{P} = \pi$ and run the chain until it appears to have reached its equilibrium. If this Markov chain Monte Carlo sampling scheme is used to estimate the expectation of a function g of $Z$, say $\mathrm{I} = E\mathrm{g}(Z)$, then $\hat{\mathrm{I}}$ is asymptotically normally distributed and approaches I in mean square as the number of Monte Carlo samples $\mathrm{B} \to \infty$ (Hastings, 1970). Of course, the problem is to find a P satisfying $\pi\mathrm{P} = \pi$. Metropolis et al. (1953) found a way to sample from $\pi$ without actually knowing P. Let $Z$ be the present state of the sequence, and $Z'$ an alternative state. Then the Metropolis algorithm is as follows: (a) define a convenient, irreducible, symmetric transition matrix Q; (b) propose a new state, say $Z'$, for the variable $Z$, according to the probabilities in the relevant row of Q; and (c) accept the proposed state with probability $\alpha(Z', Z) = \min\{1, \pi(Z')/\pi(Z)\}$. The algorithm can be proved to work by the detailed balance lemma (see e.g. Ripley, 1987). There may be need for a 'burn-in' period in the very beginning, in order to allow the algorithm to move away from a possibly badly chosen starting value $Z$. In the present case, $Z$ would be the item score pattern $\boldsymbol{X}$, and $\pi(Z)$ would be $\mathrm{f}_s(\boldsymbol{X}; \boldsymbol{\psi})$. The cleverness of the algorithm lies in the fact that in calculating $\pi(Z')/\pi(Z)$ there is no need to calculate the denominators $\gamma_s(\boldsymbol{\psi})$, as these will cancel.

If $Z$ is vector valued, and if many elements of $Z$ are independent of each other, another computational simplification is achieved by proposing a new state $Z'$ not in one draw, but stepwise. If only a single element of $Z$ is updated at a time, many additional factors in $\pi(Z')/\pi(Z)$ may cancel. In this case *one* new simulated vector $Z$ is obtained upon having consecutively considered a new state for every single element of $Z$ in turn.

The vectors simulated in this way will not be independent. The autocorrelation could be reduced by inserting 2 or more scans between successive Monte Carlo simulations. However, its only influence will be on the variance of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$; therefore it is equally well possible to use all generated response vectors and to go on generating them until the variance has become acceptable.

In the conditional Rasch model the simulations are from a distribution conditional on total score. Hence it is impossible to change the value of only one variable $X_i$ at a time: a new proposal state $\boldsymbol{X}'$ has to be obtained by interchanging the position of two different values in the item score vector. Under the Rasch model it is easy to obtain a good starting configuration: the most likely response vector is the one with $X_i = 1$ for the $s$ cleverest persons, i.e. the ones with the highest scores $T_i$. Therefore there is hardly any need for a burn-in period. Having obtained a fairly large number, say B, of simulated vectors it is possible to start estimating the parameters $\boldsymbol{\theta}$, that is, equation (5.12) can then be maximized by equating to zero its partial derivatives given in (5.13). In the next section these estimation equations will be investigated more closely.

## 5.3 Estimation equations

Several aspects of the estimation equations will now be commented upon. In particular the equations for the Monte Carlo CML estimation will be compared to those for exact CML estimation.

### 5.3.1 Starting values

In theory there is no need for iterating the MCMC procedure, and therefore the easiest choice for $\boldsymbol{\psi}$ would undoubtedly be $\psi_i = c$ for all $i$, where $c$ would be some constant, for example $c = 0$. In that case the function $\mathrm{f}_s(\boldsymbol{X}; \boldsymbol{\psi})$ would assign equal probability to all $\binom{N}{s}$ item score patterns. However, from the theory of importance sampling it may be inferred that the approximation of $\log L(\boldsymbol{\theta} \mid \boldsymbol{S}; \mathsf{X})$ by $\log L^*(\boldsymbol{\theta} \mid \boldsymbol{S}; \mathsf{X}, \boldsymbol{\psi})$ will be good in the neighborhood of $\boldsymbol{\theta} = \boldsymbol{\psi}$, but at $\boldsymbol{\theta}$ far from $\boldsymbol{\psi}$ the approximation may be bad. Therefore Geyer and Thompson suggest, firstly, that it may be wise to use a few small initial Monte Carlo cycles in order to make sure that one has arrived in the neighborhood of the final $\hat{\boldsymbol{\theta}}$. Then a truly large sample can used for the actual estimation. And, second, for the same reason they consider it useful to employ a restriction on the maximum step length per Monte Carlo cycle. So although one could use $\boldsymbol{\psi} = \mathbf{0}$ in the first cycle, if better starting values are available one might as well use these from the beginning. In the present paper the well known Gustafsson starting values (Gustafsson, 1979;

Martin-Löf, 1973) will be adapted for use with $\theta$:

$$\psi_i^{(0)} = \frac{T_i \Leftrightarrow \bar{T}_i}{\sum_{s=1}^{N-1} \mathrm{M}_s \frac{s(\mathrm{N}-s)}{\mathrm{N}(\mathrm{N}-1)}},$$

in which N is the number of persons in the sample, $\mathrm{M}_s$ the number of items with total score equal to $s$, and $\bar{T}_i = \sum_i^N T_i/\mathrm{N}$.

### 5.3.2 Identifiability

Recall from the introduction that the constraint $\theta_1 = 0$ or $\sum_i \theta_i = 0$ has to be imposed to make the model identified. Next, it is well known (Ford, 1957) that in order for a solution to the exact CML equations to exist, all persons and all items with perfect scores ($S = 0$ or $S = \mathrm{N}$, and $T_i = 0$ or $T_i = \mathrm{M}$) have to be deleted from the sample. Meaningful estimates for such persons and items cannot be obtained using maximum likelihood, and deleting them will not influence the estimation of the remaining parameters in exact CML estimation.

For the present estimation procedure however this latter assertion remains yet to be investigated. First the influence of perfect items on person parameter estimation will be considered. Repeating (5.12), we have

$$\log L^* \approx \textstyle\sum_{i=1}^N \theta_i T_i \Leftrightarrow \sum_s \mathrm{M}_s \log \left\{ \frac{1}{\mathrm{B}} \sum_{b=1}^B \exp \left[ \sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i) \right] \right\}.$$

It is evident that the term with $S = 0$ does not contribute to the value of $\log L^*$, as all the $X_{sbi}$'s are equal to zero, so $\log(1/\mathrm{B} \times \mathrm{B}) = 0$ too. Likewise, the term with $S = \mathrm{N}$ would have no contribution: let there be k items with total score $S = \mathrm{N}$, then if these items were deleted the approximated $\log L^*$ would become equal to

$$\sum_i \theta_i(T_i \Leftrightarrow \mathrm{k}) \Leftrightarrow \sum_s \mathrm{M}_s \log \left[ \frac{1}{\mathrm{B}} \sum_{b=1}^B \exp \left\{ \sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i) \right\} \right] + \mathrm{k} \sum_i (\theta_i \Leftrightarrow \psi_i).$$

The terms with $\mathrm{k} \sum_i \theta_i$ cancel, and although the resulting formula is not equal to the approximated $\log L^*$ itself, the difference does not depend on the parameters to be estimated, so the resulting estimates will be the same. For these reasons, items with perfect score patterns can be safely omitted from the estimation of person parameters.

Next, the influence of perfect persons on the estimation of other person parameters will be examined. It seems plausible to use $\psi_i \rightarrow \Leftrightarrow\infty$ or $\psi_i \rightarrow \infty$ for persons with $T_i = 0$ or $T_i = \mathrm{M}$ respectively. Thus they will generate only perfect response patterns. It is easy to show that deleting those persons from the approximate loglikelihood will not influence the estimation equations for the other persons.

### 5.3.3 Rank of the system of equations

One topic clearly needs some more attention. As $T_i$ is sufficient for $\theta_i$, in the context of exact CML there will only be as many estimation equations as there are different values observed for $T_i$. The situation is different for Monte Carlo CML estimation. Here, remembering (5.13),

$$\frac{\partial \log L^*}{\partial \theta_k} \approx T_k \Leftrightarrow \textstyle\sum_s \mathrm{M}_s \frac{\sum_b X_{sbk} \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]}{\sum_b \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]},$$

and because of the Monte Carlo processes there is no guarantee that if $T_k = T_l$, then also $X_{sbk}$ will be equal to $X_{sbl}$, not even if $\psi_k$ were chosen equal to $\psi_l$. So without taking precautions one would in this case end up with different estimates for persons with the same value of the sufficient statistic. This is undesirable. One way out, retaining only one of the persons with equal values on $T$ in the analysis, would cause problems for the conditional Monte Carlo sampling scheme. Instead, therefore, it was decided to average the estimation equations for persons with equal values on $T$. This has implications for the way the equations can be written. If $T_k = T_l$, then $\hat{\theta}_k$ will have to be equal to $\hat{\theta}_l$. To begin with, therefore, take $\psi_k = \psi_l$. Now, with $T_k = T_l$, at the maximum of the approximate $\log L^*$ it holds that

$$T_k = \textstyle\sum_s \mathrm{M}_s \frac{\sum_b X_{sbk} \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]}{\sum_b \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]},$$

and

$$T_l = \textstyle\sum_s \mathrm{M}_s \frac{\sum_b X_{sbl} \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]}{\sum_b \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]},$$

so that $T_k + T_l$ is equal to

$$\sum_s \mathrm{M}_s \frac{\sum_b (X_{sbk}+X_{sbl}) \exp\left[X_{sbk}(\theta_k-\psi_k)+X_{sbl}(\theta_l-\psi_l)+\sum_{i\neq k,l} X_{sbi}(\theta_i-\psi_i)\right]}{\sum_b \exp\left[X_{sbk}(\theta_k-\psi_k)+X_{sbl}(\theta_l-\psi_l)+\sum_{i,i\neq k,i\neq l} X_{sbi}(\theta_i-\psi_i)\right]}.$$

If $\psi_k=\psi_l$, and if we let $X^*_{sbh}$ represent $X_{sbk}+X_{sbl}$, the above equation simplifies considerably. The same can be done for all sets of persons with equal values on $T_i$, so that a different vector is obtained, say $\boldsymbol{X}^* = (X^*_1 \ldots X^*_r \ldots X^*_{N'})$, where $\mathrm{N}' \leq \mathrm{N}$, and even $\mathrm{N}' \leq \mathrm{M}$, is the number of different values actually appearing in $\boldsymbol{T}$, and each $X^*_r$ is equal to a sum over several (possibly only 1) $X_i$'s. In the same way, let $\boldsymbol{T}^* = (T^*_1 \ldots T^*_r \ldots T^*_{N'})$ be a vector containing only the different values occurring in $\boldsymbol{T}$, and let $\boldsymbol{\theta}^* = (\theta^*_1 \ldots \theta^*_r \ldots \theta^*_{N'})$ be a vector containing the $\mathrm{N}'$ different $\theta$'s corresponding to the $\mathrm{N}'$ different values in $\boldsymbol{T}^*$; and similarly for $\boldsymbol{\psi}^*$. Then $X^*_{sbr}$ can be found from

$$X^*_{sbr} = \sum_{i:T_i=t^*_r} X_{sbi}, \qquad r=1,2,\ldots,\mathrm{N}'. \tag{5.14}$$

This will give the following set of alternative estimation equations:

$$T^*_h \mathrm{N}_h = \sum_s \mathrm{M}_s \frac{\sum_b X^*_{sbh} \exp\left[\sum_{r=1}^{N'} X^*_{sbr}(\theta^*_r \Leftrightarrow \psi^*_r)\right]}{\sum_b \exp\left[\sum_{r=1}^{N'} X^*_{sbr}(\theta^*_r \Leftrightarrow \psi^*_r)\right]}, \qquad h=1\ldots\mathrm{N}',$$

where $\mathrm{N}_h$ is the number of persons with total score equal to $T^*_h$. The above is particularly relevant in the case of estimating abilities, because usually the number of persons will be much larger than the number of items. As only $\mathrm{M} \Leftrightarrow 1$ different nonperfect values are possible for $T$, maximally $\mathrm{M} \Leftrightarrow 1$ equations will then result instead of N.

## 5.4 Testing the algorithm

This section gives some results obtained in testing the algorithm used for the Monte Carlo method. All pseudo-random numbers were obtained using the generator proposed by Wichmann and Hill (1982).

### 5.4.1 The data

In order to empirically test the algorithm some realistic values for $\boldsymbol{\psi}$ were needed. It was decided to use the starting values for some data sets provided by Thissen (1982). Thissen reports the results of CML estimation on a 10 item memory test. This test was taken by 40 persons, 5 of which had a zero score, so 35 of them were left for estimation. In addition Thissen reanalyses two 5-item sections of the Law School Admissions Test. These two subtests, which will be denoted as LSAT6 and LSAT7, were analyzed earlier by Andersen and Madsen (1977) and by Bock and Lieberman (1970). The data represent responses of 1000 subjects drawn from a larger sample of students applying for admission to law schools at various universities in the United States. After omitting persons with perfect scores, 699 and 680 respectively remained for analysis.

### 5.4.2 Generation of score patterns

In the first test all possible score patterns will be considered, so preferably there should not be too many of them. Therefore it was decided to put this test not to the generation of item score patterns, which are of length N, but to the generation of person response patterns, which are only of length M. Person response patterns figure in the conditional estimation of item parameters in the same way as item score patterns figure in the conditional estimation of the person parameters. For the CML estimation of item parameters the data matrix is transposed so that it is of size N × M instead of M × N; the conditional distribution of the person response patterns given the total score $T = t$ can be denoted by $\mathrm{f}_t(\boldsymbol{X};\boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\varphi_1 \ldots \varphi_M)$ is the analog of $\boldsymbol{\psi} = (\psi_1 \ldots \psi_N)$. Data were generated for $\mathrm{M} = 5$, and $\boldsymbol{\phi}$ was taken equal to the starting values of the item parameters in the LSAT6 data. For all 4 non-perfect values of the total score 500 response vectors were generated from $\mathrm{f}_t(\boldsymbol{X};\boldsymbol{\phi})$ using the Metropolis algorithm described in Section 5.2.1. The number of different response patterns is quite small in this case ($2^5 \Leftrightarrow 2 = 30$). Therefore it was possible to calculate the theoretical conditional probabilities, i.e. the value of $\mathrm{f}_t(\boldsymbol{X};\boldsymbol{\phi}) = \exp(\sum_{j=1}^{M} X_j\varphi_j)/\gamma_t(\boldsymbol{\varphi})$, for

Table 5.1: **Chi-square goodness of fit values for the distribution of generated score patterns.**

| | | $\chi^2$ values for | | |
|---|---|---|---|---|
| T | df | B=500 | B=1000 | B=2000 |
| 1 | 4 | 7.8 | 1.8 | 2.4 |
| 2 | 9 | 5.2 | 9.7 | 13.9 |
| 3 | 9 | 4.8 | 8.5 | 5.3 |
| 4 | 4 | 3.2 | 1.6 | 3.1 |

T=total score; df=degrees of freedom.

B=number of generated score patterns.

each pattern and to compare these to the observed frequencies for the generated score patterns. Next, a chi-square statistic was calculated for each conditional distribution $\mathrm{f}_t(\boldsymbol{X};\boldsymbol{\phi})$, for $t = 1,\ldots,4$. This process was repeated with Monte Carlo sample sizes B equal to 1000 and 2000. The results are given in Table 5.1.

None of the values in this table is significant, but two remarks apply. First, it is probably not really justifiable to perform a $\chi^2$ goodness of fit test, because the generated score patterns are not independent, being subsequent realizations under the Metropolis algorithm. So these values should be interpreted with some care. And second, as a result of the randomness in the data simulation process, the generation of tables like Table 5.1 is in this case itself a random process. Therefore, the process of generating simulated data and calculating $\chi^2$ was repeated several times, and largely the pattern was as above. Unfortunately, no such test could be done for testing the generation of item score vectors for the estimation of $\boldsymbol{\theta}$ in a large sample. The, say, $2^{699} \Leftrightarrow 2$ possible score patterns simply are too many. For the time being therefore the conclusion will be that the score pattern generator works to satisfaction.

### 5.4.3 Empirical convergence of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ to $\mathrm{q}_s(\boldsymbol{\theta})$

Next the performance of the estimator $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ was examined. Note that the focus is now again on the estimation of $\boldsymbol{\theta}$ instead of $\boldsymbol{\delta}$. In this section its convergence to $\mathrm{q}_s(\boldsymbol{\theta})$ will be empirically investigated; in the next section its accuracy will be considered.

There are several factors that will influence $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$, which is the value for $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ as calculated from the first $b$ (out of B) Monte Carlo simulations. To start with there is the number of simulations $b$ upon which it is based. Hopefully, with increasing $b$, the estimate will become stable, i.e. converge to a certain value. In other words, $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b+j)} \Leftrightarrow \widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$ should go to zero for large $b$ and any value of $j$. Next, it seems likely that the convergence will be influenced by the shape of the distribution $\mathrm{f}_s(\boldsymbol{X};\boldsymbol{\psi})$, as the empirical pmf of a sample from a regularly shaped distribution in general will more closely resemble the shape of its parent than a sample of the same size from an irregularly shaped distribution. Third, recall that $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ estimates $\gamma_s(\boldsymbol{\theta})/\gamma_s(\boldsymbol{\psi})$, and that the estimate will probably be better for $\boldsymbol{\theta}$ close to $\boldsymbol{\psi}$ than for $\boldsymbol{\theta}$ a large distance from $\boldsymbol{\psi}$. So the distance from $\boldsymbol{\theta}$ to $\boldsymbol{\psi}$ is a third factor that might influence the goodness of the estimate.

In view of this last point, all subsequent investigations were performed for several distances $\boldsymbol{\theta} \Leftrightarrow \boldsymbol{\psi}$. In the following, $d$ will denote a normed Euclidean distance, i.e. a 'distance per parameter' such that the distance between $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ will be ascertained by having $\sum_r^{N'}(\theta_r^* \Leftrightarrow \psi_r^*)^2 = \mathrm{N}'d^2$, for a particular value of $d$; with $\mathrm{N}'$ the number of $\theta$'s to be estimated, i.e. the number of different values appearing in $\boldsymbol{T}$. For the LSAT data a large distance will then be one where $d = .15$, an intermediate distance will have $d = 0.10$ and a small distance will mean $d = .05$. Finally a very small distance will have $d = .01$. These numbers may seem small, but note the following. In the LSAT6 data 4 person parameters have to be estimated so $\mathrm{N}'d^2$ for the large distance would be .09; and in practice the Gustafsson starting values appeared to be well within this range of the final estimates. Furthermore, $\mathrm{N}'d^2$ for the small distance would be .01; and in all estimation runs it was observed that the final estimates were well within this distance of the previous ones. Actually, in the final run the square of the normed Euclidean distance usually was about .0001, which is the reason why also the very small distance of $d = .01$ will be investigated. So these distances do seem realistic. For other data sets, however, realistic distances might be different.

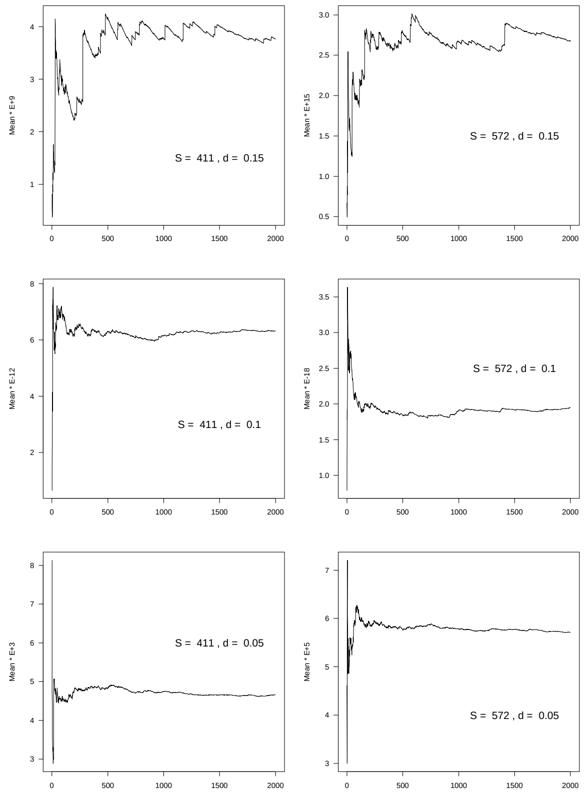

Figure 5.1: LSAT6 data: Sequential estimates $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$ for B = 2000 and some values of $S$ and $d$, where $d^2 = \sum_i^{N'}(\theta_i \Leftrightarrow \psi_i)^2/\mathrm{N}'$.

In order to study the empirical convergence, plots were constructed depicting the relationship between $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$ and $b$. Having generated Monte Carlo data at $\boldsymbol{\psi}$ and having found an arbitrary value for $\boldsymbol{\theta}$ at the required distance, the sequential estimates $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$ were then calculated for $b$ ranging from 1 to 2000.

Unfortunately, the plots arising in the context of estimating $\boldsymbol{\theta}$ from a 5 item test for N = 699 are not very regular for the larger distances. Figure 5.1 displays an average result. The sequential means of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ are depicted for $S = 411$ and for $S = 572$, for distances $d = .15$, .10 and .05. The values $S = 411$ and 572 were actually observed in the LSAT6 data; and again the starting values for these data were used to obtain $\boldsymbol{\psi}$. In both cases the plot for $d = .15$ is rather jagged, and it does not seem to have settled down even after 2000 replications. The plot for $d = .10$ seems less jagged; and the plot for $d = .05$ is markedly smoother. The plots for $d = .01$ (not presented) are still more regular than the ones with .05.

So the distance $\boldsymbol{\theta} \Leftrightarrow \boldsymbol{\psi}$ indeed does seem to play a critical role in the behavior of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$. For $\boldsymbol{\theta}$ far from $\boldsymbol{\psi}$ the estimate $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ may be unreliable, even after B as large as 2000.

In order to investigate this matter further, sequential plots were also constructed for the estimation of the *item* parameters in the 5 item LSAT test. These plots came out very smooth indeed; it seems a safe conclusion to say, without presenting the plots, that the equilibrium is reached after 1000 or even 500 replications. Even for a distance as large as $d = .50$ the plots are still smooth in this case. For the estimation of the item parameters in a 10 item test, using as $\boldsymbol{\phi}$ the starting values for the memory test, the plots are only slightly more irregular than for a 5 item test.

A closer inspection revealed that the unexpected bad results for $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$, as compared to $\hat{\mathrm{q}}_t(\boldsymbol{\delta})$, can happen because in the values of $\exp[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)]$ extreme differences may occur. In calculating $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$ two processes are involved: each new score pattern is first *generated* from $\mathrm{f}_s(\boldsymbol{X};\boldsymbol{\psi})$, and then it is *added to* $\sum_b \exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]$; this means each score pattern has a probability of occurring, depending only on $\mathrm{f}_s(\boldsymbol{X};\boldsymbol{\psi})$, and it has a particular value for $\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)$, depending on both the values $X_{sbi}$ and on the differences $\theta_i \Leftrightarrow \psi_i$. Now problems may occur if there is a (or a few) score patterns with a very small probability of occurrence, and at the same time a comparatively large value for $\exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]$; large, that is, compared to the value of $\exp\left[\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)\right]$ for other $\boldsymbol{X}$.

Recall from (5.14) that $\sum_i X_{sbi}(\theta_i \Leftrightarrow \psi_i)$ can be written as $\sum_{r=1}^{N'} X^*_{sbr}(\theta^*_r \Leftrightarrow \psi^*_r)$, with $X^*_{sbr} = \sum_{i:T_i = t_r*} X_{sbi}$, for $r = 1 \ldots \mathrm{N}'$ and $\mathrm{N}'$ the number of different values that have to be estimated for $\boldsymbol{\theta}$, which in the case of the LSAT6 data is equal to 4. In estimating $\boldsymbol{\delta}$ the variables $X^*_{tbr}$ will usually be equal to either 0 or 1, as there will not be many items with the same item total. In estimating $\boldsymbol{\theta}$ the situation is different. Still considering the LSAT6 example it may be noted first that the possibility of a rare score pattern is larger here, there being for example $\binom{699}{255}$ different patterns leading to an item total $S = 255$. Second, there are 699 persons and only 4 different total scores to be obtained. Therefore the 255 correct responses are to be split up over only 4 different $X^*$-values. These $X^*$-values will then be combined with 4 different differences $\theta^*_r \Leftrightarrow \psi^*_r$. Because of the exponentiation here is a possibility for serious trouble. The only thing one could do to avoid this, is to make sure one has good starting values. This then finally is an empirical reason prohibiting the use of starting values $\boldsymbol{\psi} = \mathbf{0}$ in estimating $\boldsymbol{\theta}$. In practice, $\boldsymbol{\psi} = \mathbf{0}$ even caused the estimation procedure to break down. On the other hand, in estimating $\boldsymbol{\delta}$ the use of $\boldsymbol{\phi} = \mathbf{0}$ posed no problems at all.

So the conclusion might be that when the ratio of the number of persons to the number of items, i.e. the ratio of the number of structural to the number of incidental parameters, is large, this may cause irregularity in the sequential plots for large distances $\boldsymbol{\theta} \Leftrightarrow \boldsymbol{\psi}$. For smaller values of that ratio the plots may well be smoother. Having understood a possible source of erratic behavior of the sequential plots, the question becomes: is there need to worry about it? To answer this question a simulation study was conducted which will be described in the next section.

### 5.4.4 Accuracy of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$

Sequential plots, as drawn in the previous section, can be enlightening and instructive, especially when one happens to come across one that displays unexpected or unwanted behavior. But for finding out something about the average behavior of the estimator alternative means are needed. A simulation study was conducted which is algorithmically given by:

1. Take a specific value for $\boldsymbol{\psi}$
2. Specify a distance, say $d = .15$
3. Choose a value for $\boldsymbol{\theta}$ at the required distance from $\boldsymbol{\psi}$
4. Calculate $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ for B = 50
5. Repeat step 4 1000 times and calculate the average and standard deviation for $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$. Compare this with the expected value

6. Repeat steps 3 - 5 for B = 200, B = 800 and B = 2000

7. Repeat steps 2 - 6 for distances of .10, .05 and .01.

Of course, the trends that are so nicely visible in the sequential plots will not appear here: looking only at $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ for B = 50 one sees a fixed point in the plot only. Moreover, calculating the mean and variance of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ for 1000 replications might not be very instructive in itself, as the values of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ should in the first place be compared to $\gamma_s(\boldsymbol{\theta})/\gamma_s(\boldsymbol{\psi})$, which are different in each of the 1000 replications because they each have a different value for $\boldsymbol{\theta}$. Therefore, for each of the 1000 replications, the relative difference $\{\widehat{\mathrm{q}_s}(\boldsymbol{\theta}) \Leftrightarrow \mathrm{q}_s(\boldsymbol{\theta})\}/\mathrm{q}_s(\boldsymbol{\theta})$ was calculated. This relative difference was the variable of interest in the present investigation; its mean and standard deviation are displayed in Table 5.2 for B = 50, 200, 800 and 2000, and for distances .01, .05, .10 and .15.

Again, the values for $S$ are the ones observed in the LSAT6 data; and the Gustafsson starting values for this data set were used to obtain a realistic point $\boldsymbol{\psi}$. Turning to the bottom part of the table first, it can be seen that for $d$ = .15, even with B = 2000 the estimates will on average be about 2 to 3 percent wrong; for $d$ = .10 and $d$ = .05 they become increasingly better, and for $d$ = .01 they are very good, that is, here they all are within less than .1 percent of the correct value, on average. The standard errors are also very acceptable for this distance. The numbers for B = 800 and B = 200 display the same pattern, and the averages are only slightly farther away form 0 here. The main difference is in the standard deviations: these are larger for smaller B. But note that even with B = 200 the standard deviation for $d$ = .01 is still less than .01 in all cases.

So in the final Monte Carlo cycle good values for $\boldsymbol{\psi}$ are necessary; and in this respect the values in the top part of the table are reassuring. Suppose one starts with starting values having $d \approx .15$ and a preliminary Monte Carlo sample size of B = 50. Then, although occasionally one might be way off, on average one could expect to come to within some 10 or 11 percent of the correct value $\mathrm{q}_s(\boldsymbol{\theta})$. A couple of preliminary cycles with B = 50 would therefore almost certainly bring one to within a distance of .05 or .01 of the true $\boldsymbol{\theta}$, and a final Monte Carlo cycle with larger B can then be performed. B = 2000 would certainly be large enough,

Table 5.2: **Relative accuracy of the estimator $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$. Values of $\{\widehat{\mathrm{q}_s}(\boldsymbol{\theta}) \Leftrightarrow \mathrm{q}_s(\boldsymbol{\theta})\}/\mathrm{q}_s(\boldsymbol{\theta})$ for 1000 replications.**

| B | S | mean | sd | mean | sd | mean | sd | mean | sd |
|---|---|---|---|---|---|---|---|---|---|
| | | d=.01 | | d=.05 | | d=.10 | | d=.15 | |
| 50 | 255 | .0001 | .013 | -.0073 | .024 | -.0938 | .127 | .1182 | .351 |
| | 411 | -.0009 | .015 | .0014 | .030 | -.0563 | .127 | -.1195 | .151 |
| | 465 | .0000 | .008 | .0167 | .103 | -.0622 | .118 | -.1081 | .164 |
| | 572 | .0001 | .008 | .0075 | .034 | .0368 | .109 | .0354 | .215 |
| | 626 | .0004 | .012 | -.0061 | .045 | .0161 | .107 | -.0411 | .122 |
| 200 | 255 | .0002 | .007 | .0108 | .021 | -.0179 | .029 | .0530 | .159 |
| | 411 | -.0001 | .004 | .0013 | .035 | -.0193 | .040 | .1208 | .238 |
| | 465 | .0002 | .005 | .0180 | .034 | -.0266 | .075 | .0503 | .155 |
| | 572 | -.0000 | .006 | .0026 | .028 | .0198 | .031 | -.0012 | .072 |
| | 626 | -.0000 | .003 | -.0055 | .014 | -.0015 | .073 | .0269 | .081 |
| 800 | 255 | .0002 | .003 | .0050 | .012 | .0187 | .028 | .0128 | .081 |
| | 411 | .0000 | .004 | -.0044 | .013 | .0151 | .020 | -.0105 | .069 |
| | 465 | .0002 | .002 | -.0024 | .007 | -.0131 | .045 | -.0581 | .057 |
| | 572 | -.0001 | .002 | -.0006 | .014 | .0043 | .045 | -.0340 | .062 |
| | 626 | -.0000 | .001 | .0011 | .005 | -.0137 | .017 | -.0080 | .054 |
| 2000 | 255 | -.0001 | .002 | -.0016 | .002 | .0114 | .031 | -.0334 | .079 |
| | 411 | -.0000 | .002 | .0026 | .014 | .0129 | .032 | .0265 | .074 |
| | 465 | -.0002 | .001 | .0001 | .008 | .0086 | .021 | .0140 | .055 |
| | 572 | .0000 | .002 | .0013 | .002 | -.0096 | .026 | -.0121 | .032 |
| | 626 | -.0000 | .001 | -.0019 | .006 | -.0083 | .014 | .0214 | .026 |

B:Number of simulations; S=item total.

d: distance from $\boldsymbol{\psi}$ to $\boldsymbol{\theta}$; a distance of $d$ means $\sum_r^{N'}(\theta_r^* - \psi_r^*)^2 = \mathrm{N}'d^2$

mean and sd: average and standard error of the relative error of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$.

but in a very close neighborhood of $\boldsymbol{\theta}$ also a Monte Carlo sample size of B = 200 could be adequate.

## 5.5 Results

In this section results will be presented for the Monte Carlo estimation of ability parameters for the LSAT and memory data. The results will be compared to exact estimates with respect to both accuracy and necessary computing times.

### 5.5.1 Parameter estimates

Table 5.3 contains the ability estimates for the memory test. This is a 10 item test, so there are 9 nonperfect values for $T$, and hence maximally 9 different estimates for $\theta$ are possible. The estimates reported in this table were obtained using an initial Monte Carlo sample size of B = 500 and a final one of B = 2000. In column 3 the Monte Carlo CML estimates are reported; column 4 contains the exact CML estimates. Recall from the introduction that in practice exact CML estimates for $\boldsymbol{\theta}$ are never calculated: usually the item parameters are estimated first, and the abilities are then obtained by ordinary ML, treating the difficulties as known. These ML estimates of $\boldsymbol{\theta}$ are reported in column 5. It may be noted that there were no subjects with total scores larger than 7. Therefore both MC and exact CML estimates for $\boldsymbol{\theta}$ were obtained only for $T = 1$ up to $T = 7$. This is in contrast to the results in column 5: if it is assumed that the item parameters are known, it is no problem to calculate ML estimates for ability for any value of $T$, whether this value actually appears in the data or not. In order to be able to perform a proper comparison, the ML estimates have been rescaled to a mean of 0 for the first 7 estimates. The differences between the Monte Carlo and the exact CML estimates are small; they are reported in the column labelled 'diff1'. The last column contains the differences between the ML and the exact CML estimates. It would be tempting to compare diff1 with diff2. However, although

Table 5.3: **Ability estimates for the memory data.**

| $T$ | $\mathbf{N}_t$ | MC | Exact | ML | se | diff1 | diff2 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | -2.203 | -2.187 | -2.123 | 1.157 | -.016 | .064 |
| 2 | 9 | -1.129 | -1.133 | -1.109 | .896 | .004 | .024 |
| 3 | 9 | -.399 | -.406 | -.412 | .785 | .007 | -.006 |
| 4 | 9 | .180 | .175 | .156 | .728 | .005 | -.019 |
| 5 | 4 | .696 | .689 | .664 | .702 | .007 | -.025 |
| 6 | 2 | 1.184 | 1.179 | 1.155 | .704 | .005 | -.024 |
| 7 | 1 | 1.670 | 1.683 | 1.672 | .739 | -.013 | -.011 |
| 8 | 0 | - | - | 2.276 | .829 | - | - |
| 9 | 0 | - | - | 3.147 | 1.081 | - | - |

T: total score; $\mathrm{N}_t$: number of persons with total score $T = t$.

MC: Monte Carlo CML estimate; Exact: Exact CML estimate

ML and se: estimate and standard error obtained treating $\delta$ as known.

diff1: MC - Exact; diff2: ML - Exact.

both the MLE and the CMLE are consistent estimators of $\boldsymbol{\theta}$, there is no reason why their small-sample estimates should be equal, or even similar; not even when the MLE would be based on the true $\boldsymbol{\delta}$ instead of on $\hat{\boldsymbol{\delta}}$.

Table 5.4 contains the estimates for the two LSAT data sets. In both data sets there were nearly 700 persons with a nonperfect total score. As there are 5 items, 4 different nonperfect total scores could be obtained and hence 4 different estimates result. In order to give an impression of the difference in accuracy resulting from different Monte Carlo sample sizes, the estimates reported for the LSAT6 data have been obtained with initial and final Monte Carlo sample sizes of B = 50 and 200, respectively; and the estimates for the LSAT7 data have been obtained with B = 500 and B = 2000. These latter estimates are more accurate, diff1 being smaller for the LSAT7 than for the LSAT6 data, but for the LSAT6 data too, i.e. with B = 50/200, the differences occurred in the third decimal place only.

It may be remarked here that one initial Monte Carlo cycle with B = 500 always resulted in estimates very close to the final ones; with an initial Monte Carlo

Table 5.4: **Ability estimates for the LSAT data.**

| | $T$ | $\mathbf{N}_t$ | MC | Exact | ML | se | diff1 | diff2 |
|---|---|---|---|---|---|---|---|---|
| LSAT6 | 1 | 20 | -1.721 | -1.717 | -1.602 | 1.181 | -.004 | .115 |
| B=50/200 | 2 | 85 | -.512 | -.519 | -.474 | .990 | .007 | .045 |
| | 3 | 237 | .513 | .516 | .481 | .987 | -.003 | -.035 |
| | 4 | 357 | 1.719 | 1.720 | 1.600 | 1.177 | -.001 | -.120 |
| LSAT7 | 1 | 40 | -1.535 | -1.539 | -1.486 | 1.144 | .004 | .053 |
| B=500/2000 | 2 | 114 | -.463 | -.463 | -.443 | .948 | .000 | .020 |
| | 3 | 205 | .453 | .455 | .436 | .950 | -.002 | -.019 |
| | 4 | 321 | 1.545 | 1.547 | 1.487 | 1.149 | -.002 | -.060 |

T: total score; N$_t$: number of persons with total score $T = t$.

MC: Monte Carlo CML estimate; Exact: Exact CML estimate

ML and se: estimate and standard error obtained treating $\delta$ as known.

diff1: MC - Exact;     diff2: ML - Exact.

sample size of B = 50 two initial cycles were always sufficient and often even only one would have done the job.

The values of diff2 are larger here than in the previous table, which seems to suggest that indeed the ($\hat{\delta}$-based) MLE and the CMLE become closer with increasing sample size, i.e. with increasing value of the number of items M. This suggestion was confirmed by some analyses on a simulated data set with N = 500 and M = 30, where the largest value of diff2, occurring at $\hat{\theta} = 3.114$, was equal to ⇔.035.

### 5.5.2 Comparison to algorithms for exact calculation

In this section the performance of the Monte Carlo method will be compared to the algorithms for exact computation. The sum algorithm never caused any problems. The extended algorithm, however, proved to be too inaccurate for the present problem. To investigate the behavior of the extended algorithm, it was used to calculate the gamma functions $\gamma^{(1)}_{s-1}(\boldsymbol{\theta})$, for $\boldsymbol{\theta}$ being the vector of

Gustafsson starting values for the LSAT6 data. As an example, $\gamma^{(1)}_{410}(\boldsymbol{\theta})$ was equal to 3.774E + 325 when calculated with the extended algorithm, whereas with the sum algorithm it was calculated as 5.699E+229. Furthermore, with the extended algorithm there also come a point where $\gamma^{(k)}_{s-1}(\boldsymbol{\theta})$ would become negative; in the above example the extended algorithm calculated $\gamma^{(1)}_{411}$ as $\Leftrightarrow$3.160E+324, whereas the sum algorithm calculated it as 3.930E + 229. Hence the extended algorithm is not accurate enough for the conditional estimation of person parameters, and the speed of the Monte Carlo algorithm will therefore be compared to the speed of the sum algorithm only.

Using the sum algorithm, the CML estimation of the person parameters for the LSAT6 data took 63 seconds; for the LSAT7 data this was accomplished in 53 seconds. The function maximization for the exact algorithms was carried out with a quasi-Newton routine, using 9 and 8 iterations to converge respectively, starting from the Gustafsson starting values. This means 9 or 8 evaluations of the gamma functions were necessary, and hence each iteration took about 7 seconds. Considering the fact that there are 4 parameters per problem, it may be recalled from section 5.1.2 that there are therefore 5 sets of gamma functions to be evaluated. As $7/5 \approx 1.5$, this nicely coincides with the earlier observation that the calculation of one set of gamma functions would cost 1 or 2 seconds.

Recall that for the Monte Carlo CML estimation of the person parameters Monte Carlo sample sizes of 50 and 200 appeared to be adequate. One cycle with B = 200 took about 25 seconds. Most of this time (i.e. about 22 seconds) was spent in generating the score patterns. The score pattern generation takes this long time because each simulated vector is of length N, i.e. nearly 700; and 200 vectors have to be simulated for all observed values of $S$. The remaining time was spent in the maximization of the loglikelihood function. The Monte Carlo equations were maximized using the Fletcher-Reeves algorithm (Fletcher and Reeves, 1964), in a a slightly modified form proposed by Polak and Ribiere, details of which can be found in Press et al. (1992). The results were slightly more accurate with this routine than with the quasi-Newton one, but it needed more function evaluations which made it less suitable for use with the exact algorithms.

The entire Monte Carlo estimation procedure using B = 50 and B = 200 took 26 to 33 seconds, depending on whether 1 or 2 preliminary cycles with B = 50 were necessary. With initial and final Monte Carlo sample sizes of B = 200 and 800 the entire estimation procedure would take 122 seconds; with B = 500 and 2000 the time needed was 319 seconds.

## 5.6 Conclusion and Discussion

The Monte Carlo estimation method proposed by Geyer and Thompson (1992) has been compared to two algorithms for exact calculation of the person parameter estimates in the conditional Rasch model. It was remarked in the introduction that for small numbers of parameters, say up to 60, the extended algorithm formulated by Liou (1994) is at present the best choice. For large numbers of parameters however, say 700, the Monte Carlo method is definitely superior to the extended algorithm. The divisions and subtractions necessary in the latter algorithm would cause the rounding errors to accumulate to such an extent that some of the functions $\gamma_{s-1}^{(k)}(\boldsymbol{\theta})$ would take on negative values. Which they should not, being sums of products of exponentials.

As to the sum algorithm, the conclusion is less unambiguous. Summarizing the results from section 5.5 it can be stated that if one wants to be absolutely sure to have very accurate estimates, the Monte Carlo method is slower than exact estimation. However, if one is willing to accept results such as the top part of Table 5.4, where the differences occur in the third decimal place only, the Monte Carlo method is faster than the sum algorithm. The gain in speed however is not enormous and someone insisting on CML estimates may well be prepared to wait the additional 30 seconds for exact values.

As to the Monte Carlo method itself, many topics could still be further investigated. First, in estimating $\boldsymbol{\delta}$ the method is much slower than the sum algorithm; in estimating $\boldsymbol{\theta}$ it seems to be of about the same speed. Therefore the method appears to become more interesting with increasing number of parameters. On the other hand, it seems that also an increasing precision is then required in the starting values. Recall that for estimating $\boldsymbol{\delta}$ starting values $\boldsymbol{\phi} = \mathbf{0}$ posed no problem; whereas in estimating $\boldsymbol{\theta}$ these values were useless. So there might come a point, i.e. a number of parameters, where the starting values would have to be so accurate that they could serve as proper and acceptable estimates themselves.

Next, recall that the estimates presented in Tables 5.3 and 5.4 were for one Monte Carlo run only. The estimation procedure could be reproduced a large number of times in order to get some additional insight into the average and variance of the Monte Carlo estimates.

Also note that, using Monte Carlo estimation methods, the covariance matrix of the parameter estimates usually is not a by-product of the estimation procedure. If it is needed, it will have to be obtained by means of another procedure. One idea might be to approximate the Hessian by the same methods as the derivative of the loglikelihood. However, this would increase the CPU time needed. Another possibility could be to numerically differentiate the derivatives of the loglikelihood (Meilijson, 1989).

The variance of the estimator $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(B)}$ needs some more attention too. According to Hastings (1970) this variance is equal to

$$\frac{\sigma^2}{\mathrm{B}} \sum_{j=-B+1}^{B-1} \left(1 \Leftrightarrow \frac{|\, j \,|}{\mathrm{B}}\right) \rho_j$$

where $\rho_j$ is the autocorrelation for lag $j$. As pointed out before, the autocorrelation cannot a priori be assumed to be zero. Two methods for investigating this are, first, similar sequential plots for the estimated variance of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$ as the ones for $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(b)}$ itself. The second method would be to calculate the variance of $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})^{(B)}$ for different numbers of scans between two successive simulations; this should obviously reduce the autocorrelation.

The conditions under which the estimators $\widehat{\mathrm{q}_s}(\boldsymbol{\theta})$ and $\hat{\mathrm{q}}_t(\boldsymbol{\delta})$ perform well could be examined more closely. Especially the influence of the distance from $\boldsymbol{\theta}$ to $\boldsymbol{\psi}$ (see section 5.4.3), or from $\boldsymbol{\delta}$ to $\boldsymbol{\phi}$, might be investigated further.

Next, the storage required for the Monte Carlo estimation procedure is considerable, especially if B = 2000 is needed. A tensor of approximately size M $\times$ M $\times$ B has to be stored, in which B is the number of Monte Carlo samples, and M is the number of items. Storage is necessary because the maximization of every approximate $\log L^*$ needs several iterations, in each of which the Monte Carlo data appear, together with different values for the parameters (see equation 5.12).

The final remark concerns not the Monte Carlo approximation of the CMLE of $\boldsymbol{\theta}$, but this latter estimator itself. The small sample $\delta$-based MLE of $\boldsymbol{\theta}$ is known to be biased outward (Lord, 1983; Warm, 1989), i.e. its expectation is smaller than the true value for negative $\theta$ and larger than the true value for positive $\theta$. In Tables 5.3 and 5.4 however it appears that, for negative $\theta$, the exact CML estimate is smaller than the $\hat{\delta}$-based ML estimate, and for positive $\theta$ it is larger. Hence, it would be interesting to investigate (a) whether the $\hat{\delta}$-based MLE has bias comparable to the $\delta$-based MLE; and (b) whether the CMLE has larger small sample bias than either the $\delta$-based or the $\hat{\delta}$-based MLE.

# Epilogue

With the exception of Chapter 5, this thesis has dealt with models for polytomous item responses. Chapter 5 was concerned with an estimation method for person parameters in the Rasch model for binary items. In particular, the performance of a Monte Carlo Markov chain method for CML estimation of $\theta$ was investigated. For large numbers of parameters the speed of this procedure appeared to be comparable to the speed of exact calculation, and the estimates produced were fairly accurate. The method therefore seems promising for situations in which exact calculation of the estimates is impossible. An example that comes to mind is CML estimation of the item parameters in models for continuous responses (Müller, 1987; Verhelst, 1995).

As regards polytomous items, in Chapter 1 the item information function for trinary PCM items was investigated, that is, for PCM items with maximum score M = 2. The condition for this function to be either bi- or unimodal was established, and the location and value of the maxima were derived. The item information function depends on the item parameters $\alpha, \delta_1$, and $\delta_2$, and it is unimodal if $\alpha(\delta_2 \Leftrightarrow \delta_1) \leq 4 \ln 2$. A practical conclusion in this chapter was that in item banking, the construction of more items with M = 1 and M = 2 might be preferable to the construction of less items with larger M.

In Chapter 2 the relation between polytomously scored item responses and Guttman dependence was thoroughly investigated. One of the conclusions was that a variable satisfying the GRM can never be distributed as the total score on a set of independent binary 1-PL items. Consequently, it is also impossible for the response curves of a GRM item to coincide with those of a PCM item satisfying

Huynh's (1994) condition.

In Chapter 3 it was demonstrated that the GRM and the PCM are mathematically unsuited for use with sequential scoring. An interesting topic for further research would now be the investigation of the applicability of the PCM, the GR-M, and the SM, to parallel scoring. Since it has been shown that, of the PCM, the GRM, and the SM, only the GRM is suitable for graded scoring (Jansen and Roskam, 1986), and that only the SM is suitable for sequential scoring (Chapter 3), it would be interesting to find out whether perhaps only the PCM would be useful for parallel scoring. Although it was demonstrated in Chapter 2 that the GRM cannot model parallel processes in which 1-PL curves are assumed for the binary subtasks, still there is the possibility that these models can describe parallel processes with other ICCs. It remains to be shown that the GRM and the SM are unsuited for parallel processes, for any functional form of the ICCs.

Chapter 4, finally, investigated the possibility of distinguishing between responses generated under the PCM, the GRM, and the SM. It appeared that these models could be well distinguished, both in a forced-choice experiment and in a decision procedure.

Throughout this dissertation, the importance of the scoring rule has been stressed on several occasions. This topic will now be pursued in some more detail. First, a distinction is made between rating scales and educational tests. Many item response models require as input numbers of at least an ordinal nature. With rating scales, the responses themselves are of an ordinal nature. With educational test items however the response in general does not consist of a number. Here the response can be, for example, a short essay, or a mathematical derivation, or an enumeration of certain facts, or a written answer to a question. The process of transforming these responses into numbers is known as scoring. Three different types of scoring were distinguished: graded scoring, sequential scoring and parallel scoring. It was argued in Chapter 3 that the very same response on the very same test item may be scored in several ways, resulting in possibly different scores. An example was given of an item that could be subjected both to parallel and to sequential scoring; another example is an essay question, which can be subjected either to graded scoring (one overall judgement given by an expert in the field), or to parallel scoring (one credit given for each feature in a list of features which the essay does exhibit).

Because of this ambiguity in the relation between response and item score, modeling item responses is in fact a two-stage process. First the response has to be scored. Then the score has to be modeled. Ideally, the scoring rule will follow the solution strategy, and the model will follow the scoring rule. If the scoring rule is modified, the model has to change as well. This connection between scoring rule and model has also been noted by Samejima (1997), who offers a detailed example of the possible steps that can be taken in producing a mathematical proof for $a^2 = b^2 + c^2 \Leftrightarrow 2ab\cos A$. But, as she remarks, ".. any grading system is arbitrary. If our experimental setting is improved and allows observation of the examinee's performance in more finely graded steps, then $\mathrm{m}_i$ will be larger". The symbol $\mathrm{m}_i$ denotes the maximum score. When the same response is graded in two different ways, one of these resulting in, say, a dichotomous score and the other in a polytomous score, two different models will be needed to describe these scores. Hence this is another example of the relation between scoring rule and model.

If one of the three above-mentioned scoring rules is applied, and if one has to choose between the PCM, the GRM, and the SM, this task seems to be an easy one. Indeed, the logistic GRM is eminently suited for graded scoring, and the logistic SM for sequential scoring. With these models applied to these scoring rules, the interpretation of the parameters is straightforward. The PCM, however, needs some further consideration.

The PCM allows $\delta_k$ to be larger that $\delta_{k+1}$, but no satisfactory substantive interpretation has been found for this phenomenon. Furthermore, the PCM is unsuited for sequential scoring, as was demonstrated in Chapter 3; and it is also unsuited for graded scoring, including rating scales (Jansen and Roskam, 1986). This holds for the rating scale version of the PCM as well. But with the PCM as a model for parallel scoring, the interpretation of the parameters remains problematic. First, it is once more repeated that the score on a PCM item is distributed as the total score on a set of independent binary 1-PL items only under a condition on the item parameters. This condition takes the form of a critical distance that has to be exceeded. This condition was also derived by Andrich (1985), who presented a table of lower bounds for $\lambda$ (the half-distance between consecutive parameters), being half the size of Huynh's critical distances. The condition entails a constraint on the PCM family: many of its members do not satisfy it. Those members of the PCM that do not satisfy the condition, are unsuited for parallel scoring. But, second, even for those members that do satisfy the condition, the parameters have no obvious relation to the process that is being modeled: the $\delta$'s in Equation 1.1 in Chapter 1 do not reflect the difficulties of independent binary subtasks.

In order for the PCM parameters to reflect the difficulty of the independent subtasks, a reparameterization must be carried out. This reparameterization has been proposed by Verhelst and Verstralen (1991), and is only defined for those members of the PCM that satisfy the above-mentioned condition. Let $\boldsymbol{\delta} = (\delta_1 \ldots \delta_M)$ represent the usual parameterization, and let $\boldsymbol{\eta} = (\eta_1 \ldots \eta_M)$ be the parameters in the reparameterized model. Only one item is considered, so there is no need for an item subscript $j$. The score probabilities in the reparameterized model are given by $\Pr(X = k; \theta, \boldsymbol{\eta}) \propto \exp(k\theta)\gamma_k(\boldsymbol{\eta})$, where $\gamma_k(\boldsymbol{\eta})$ are the elementary symmetric functions of the vector $\boldsymbol{\eta}$, which were defined in Equation 5.5. The subset of the family of PCM models for which the reparameterization is defined, could be called a constrained partial credit model (CPCM). The parameters $\boldsymbol{\eta}$ in the CPCM are the difficulty parameters of a set of independent binary 1-PL subtasks.

The reparameterized CPCM is completely symmetric in its parameters. As an example, for a polytomous item with $M = 2$, the parameter vector $\boldsymbol{\eta} = (.5, .8)$ will yield exactly the same response curves as the parameter vector $(.8, .5)$. Using this model, it is therefore impossible to decide which parameter value corresponds to which subtask. However, if a parallel scoring rule is used, the subtask scores must be known. And if the subtask scores are known, application of the binary Rasch model to the subtask scores would do exactly the same job as the CPCM. In fact, it would even do the job better, because with the RM it is possible to connect (estimated) parameter values back to subtasks.

The above is not meant to completely discourage the use of the PCM. The PCM can, for example, be profitably applied in the detection of violations of the local independence assumption in a set of binary items. The critical distance has be be satisfied if this model is to describe the total score on a set of independent binary 1-PL items. Andrich (1985) argues that, if the PCM is applied to a set of binary items, smaller estimated values of $\delta_2 \Leftrightarrow \delta_1$ may be indicative of local dependence. Wilson (1988) used this idea to investigate whether the assumption of local independence was tenable within several substantively related subtests of a larger test. Furthermore, the PCM has several very desirable properties. It can be derived from fundamental measurement theoretic requirements (Fischer, 1995). And, being an exponential family model, it is mathematically tractable. However, the fact remains that there is no guarantee that exponential family models actually describe an empirical process and have interpretable parameters. On the other hand, models that may be more realistic can perhaps not be adequately estimated, as is exemplified by the discussion concerning the estimation of the parameters in the 3-PL model (Thissen and Wainer, 1982). The problems with the PCM therefore are a striking illustration of the dilemma between requiring a model to be mathematically tractable, and at the same time realistic.

# Samenvatting (Summary in Dutch)

In dit proefschrift wordt ingegaan op enkele modellen uit de item respons theorie. Item respons modellen worden gebruikt om het verband te beschrijven tussen enerzijds iemands vaardigheid op een bepaald terrein, en anderzijds de responsen van deze persoon op een aantal items die op dat terrein betrekking hebben. De term vaardigheid kan zeer breed worden opgevat: bij vaardigheid kan worden gedacht aan een intellectuele of cognitieve vaardigheid zoals algemene ontwikkeling, topografische kennis of luistervaardigheid in een vreemde taal, maar ook aan psychologische constructen als bijvoorbeeld dominantie, depressiviteit of zelfvertrouwen. Items zijn dan examenopgaven of de vragen in een psychologische test. Responsen zijn de antwoorden op de items.

Al de bovengenoemde 'vaardigheden' hebben gemeen dat zij latent zijn, dat wil zeggen zij zijn niet direct observeerbaar. Een kenmerk als lengte is niet latent want van twee personen is direct duidelijk wie langer is. Wie daarentegen beter is in hoofdrekenen, is niet direct duidelijk. Om verschillen tussen personen op het latente kenmerk 'vaardigheid in hoofdrekenen' zichtbaar te maken, is het eerst nodig deze personen een aantal hoofdrekensommen te laten maken. Om verschillen in dominantie zichtbaar te maken is het nodig om ofwel deze personen enige tijd te observeren in hun contacten met anderen (de responsen zijn dan scores in observatiecategorieën), ofwel hen een psychologische test over dominantie te laten invullen (en dan te hopen dat de antwoorden waarheidsgetrouw zijn).

Omdat de beschouwde vaardigheden latent zijn, wordt item respons theorie ook wel latente trek theorie genoemd (in het Engels: latent trait theory). De latente vaardigheid wordt vaak aangeduid met het symbool $\theta$. De antwoorden (responsen) op testvragen (items) zijn wel observeerbaar. In een item respons model worden de geobserveerde responsen in verband gebracht met de ongeobserveerde latente trek. De moeilijkheid van de items dient ook in het model verdisconteerd te worden. Over het verband tussen itemmoeilijkheid, vaardigheid en geobserveerde responsen kunnen verschillende aannames gemaakt worden. Wanneer een gegeven antwoord slechts goed of fout kan zijn, zal men bijvoorbeeld aannemen dat personen met een heel lage vaardigheid maar een kleine kans hebben om de vraag goed te beantwoorden, en personen met een hoge vaardigheid een grote kans. Daar tussenin, neemt men aan, zal de kans op een goed antwoord geleidelijk toenemen met de vaardigheid. Het verband tussen vaardigheid en de kans op het geven van het goede antwoord beschrijft men dan met behulp van een wiskundige formule.

Behalve modellen voor goed/fout responsen (dit worden ook wel dichotome responsen genoemd) zijn er ook modellen voor polytoom gescoorde responsen. Een polytoom gescoorde respons neemt een van de waarden $0, 1, \ldots, M$ aan. Als $M$ bijvoorbeeld gelijk is aan 3, dan betekent een score 0 dat het gegeven antwoord helemaal fout was, een score 3 dat het antwoord helemaal goed was, en de scores 1 en 2 duiden erop dat het antwoord gedeeltelijk juist was. Een ander onderscheid tussen modellen is of zij rekening houden met de mogelijkheid van gokken, zoals bij multiple choice opgaven.

Als het model geformuleerd is, dan dienen de onbekende grootheden in het model te worden geschat. Deze onbekende grootheden worden parameters genoemd. In item respons modellen zijn er in het algemeen twee soorten parameters: itemparameters (bijvoorbeeld de moeilijkheid van de vragen) en persoonsparameters (de vaardigheid op één of meerdere dimensies). Deze parameters zijn onbekend en zij dienen uit de geobserveerde responsen te worden geschat. Voor dit probleem zijn verschillende oplossingen bedacht. De meeste oplossingen komen erop neer dat men probeert eerst de itemparameters te schatten, en daarna, in een tweede stap, de persoonsparameters. Bij het schatten van de persoonsparameters beschouwt men dan de geschatte itemparameters als bekend. Deze tweetrapsprocedure introduceert echter extra onnauwkeurigheid in de vaardigheidsschattingen. In sommige gevallen is directe schatting van de vaardigheid $\theta$ mogelijk met behulp van zogenaamde conditionele maximum likelihood (CML) schatters. Voor grote aantallen personen is CML schatting van $\theta$ echter nogal tijdrovend. In hoofdstuk 5 van het proefschrift wordt daarom een methode onderzocht om de exacte CML schatters voor $\theta$ te vervangen door een benadering. De onderzochte werkwijze is voorgesteld door Geyer and Thompson (1992); het is een Markov chain Monte Carlo methode. Deze methode wordt in dit proefschrift toegepast op het conditionele Raschmodel voor dichotome data (Rasch, 1960). Het blijkt dat de methode redelijk nauwkeurig is, maar evenals de exacte berekeningswijze veel rekentijd kost.

De hoofdstukken 1 tot en met 4 van het proefschrift gaan niet over het schatten van $\theta$ maar over de vraag welk model en welke items in een bepaalde situatie het beste kunnen worden gebruikt. In hoofdstuk 1 wordt de iteminformatiefunctie, $I(\theta)$, onderzocht voor trinaire items in het partial credit model (Masters, 1982). Het partial credit model (PCM) is een model voor polytoom gescoorde items; trinaire items hebben als mogelijke score $0, 1$ of 2. Voor elk afzonderlijk item in een toets kan uit het aangenomen model een informatiefunctie worden afgeleid. Deze functie geeft, voor elke waarde van $\theta$, de hoeveelheid 'informatie' die een antwoord op dit item levert over de waarde van $\theta$. Als iemand drie items krijgt voorgelegd die alle drie veel te moeilijk zijn, dan zijn hoogstwaarschijnlijk alle drie de gegeven antwoorden fout en weet de onderzoeker nog steeds niet of de onderzochte persoon net onder het niveau van de items zit, of ver daaronder. Het is beter om items aan te bieden die enerzijds niet al te moeilijk en anderzijds ook weer niet veel te makkelijk zijn. Dan is er een gerede kans dat sommige items goed zullen worden beantwoord en andere fout, zodat de foutenmarge bij het schatten van $\theta$ binnen de perken blijft. De eigenschappen van de iteminformatiefunctie voor dichotome items, dat wil zeggen voor items met mogelijke scores 0 en 1, zijn goed bekend: het is eenvoudig af te leiden waar het maximum van deze functie ligt en voor welk vaardigheidsniveau het item derhalve geschikt is. De eigenschappen van informatiefuncties voor polytoom gescoorde items zijn minder goed bekend. In hoofdstuk 1 wordt aangetoond dat de iteminformatiefunctie voor trinaire items onder het PCM één- of tweetoppig is. De voorwaarde waaronder de functie eentoppig is wordt afgeleid. Bovendien wordt de locatie en waarde van het maximum bepaald. Het maximum voor eentoppige functies blijkt hoger dan dat voor tweetoppige functies. Dit resultaat kan worden gebruikt bij zogenaamde computergestuurde adaptieve testafname: hier wordt de test per computer afgenomen, na elk antwoord wordt een voorlopige vaardigheidsschatting gemaakt, en het volgende item dat (door de computer) wordt aangeboden is het item met de hoogste informatiewaarde in het gebied van de voorlopige schatting voor $\theta$. Kennis van de waarde en locatie van dat maximum kan de keuze van een geschikt item vergemakkelijken.

De hoofdstukken 2 tot en met 4 gaan over relaties tussen drie families van item respons modellen voor polytome responsen. De beschouwde modellen zijn steeds het PCM, het graded respons model (GRM; Samejima, 1969) en het sequentiële model (SM; Tutz, 1990). Deze modellen worden wiskundig gedefinieerd in de betreffende hoofdstukken. Zij verschillen in de wiskundige functie die ze aannemen voor de kans op een itemscore. De verschillen tussen de drie modellen zijn echter vaak klein, dat wil zeggen, het is vaak zo dat, gegeven de vaardigheid $\theta$, de kansverdelingen voor de score op een item elkaar niet zo veel ontlopen. Dan dient zich de vraag aan of de drie modellen niet inwisselbaar zijn, of dat misschien met één model kan worden volstaan. Op deze vraag hebben de hoofdstukken 2 tot en met 4 betrekking.

In de hoofdstukken 2 en 3 staat het begrip Guttmanschaal centraal. Een Guttmanschaal is een verzameling dichotome items met de volgende eigenschap: als de items zijn gesorteerd op volgorde van opklimmende moeilijkheid, dan heeft iemand die item $k$ goed beantwoordt, ook alle voorgaande items goed. De antwoorden op de items in een Guttmanschaal zijn dus afhankelijk: als iemand 4 items uit een Guttmanschaal goed heeft beantwoord, dan zijn dat de 4 gemakkelijkste items. Het is duidelijk dat een Guttmanschaal in veel gevallen geen realistisch beeld van de werkelijkheid geeft. Niettemin kan worden aangetoond (zie hoofdstuk 2) dat *elk* item respons model voor een polytoom gescoord item kan worden geschreven in de wiskundige vorm van de somscore op een Guttmanschaal. Dit resultaat wordt vervolgens gebruikt om aan te tonen dat de score op een 'graded respons item' nooit kan worden geschreven als de somscore op een verzameling onafhankelijke dichotome Rasch items (Rasch, 1960).

Het eerste van de twee bovengenoemde resultaten betekent niet dat alle item respons modellen nu zonder meer geschikt zijn om Guttman schalen te beschrijven. Dit wordt duidelijk in hoofdstuk 3, waarin onderzocht wordt of het PCM, GRM en SM geschikt zijn voor gebruik bij sequentieel gescoorde responsen. Bij sequentieel scoren wordt steeds één onderdeel van het gegeven antwoord bekeken. Is dit onderdeel van het antwoord onjuist dan wordt er geen punt voor toegekend, en bovendien stopt dan ook het scoringsproces. Is het onderdeel juist beantwoord, dan wordt een punt toegekend en wordt het volgende onderdeel bekeken. Een voorbeeld: om de vraag $[(7 \ast 8) \Leftrightarrow 4] \ast 3$ goed te beantwoorden moet eerst $7 \ast 8 = 56$ worden berekend, daarna $56 \Leftrightarrow 4 = 52$, en tenslotte $52 \ast 3 = 156$. Stel nu dat iemand opschrijft $7 \ast 8 = 72$, en vervolgens $72 \Leftrightarrow 4 = 68$, en $68 \ast 3 = 204$. Bij sequentieel scoren worden dan 0 punten toegekend. Het zou ook mogelijk zijn om 2 punten toe te kennen voor de antwoorden die goed zijn, gegeven de eerste fout, maar dat is geen sequentieel scoren. In dat geval spreekt men van parallel scoren.

Aangezien bij sequentieel scoren het scoringsproces wordt beëindigd na de eerste fout, is er een grote overeenkomst tussen een sequentieel gescoord polytoom item en de scores op een Guttmanschaal: in beide gevallen zijn bij een score $k$ de eerste $k$ items of onderdelen juist beantwoord. Het SM is speciaal geconstrueerd voor sequentieel scoren. In hoofdstuk 3 wordt aangetoond dat men bij het toepassen van het PCM of het GRM op sequentieel gescoorde responsen een specificatiefout maakt, d.w.z. deze modellen bezitten niet alle eigenschappen die nodig zijn om sequentieel gescoorde responsen adequaat te beschrijven. Het praktische belang hiervan is dat nu bekend is dat voor sequentiële responsen, van deze drie modellen alleen het SM geschikt is. Jansen and Roskam (1986) hebben reeds aangetoond dat voor ratingschalen van deze drie modellen alleen het GRM geschikt is. Wanneer nu nog wordt aangetoond dat voor, bijvoorbeeld, parallel gescoorde responsen alleen het PCM geschikt is, dan hebben alle drie de modellen theoretisch bestaansrecht.

In hoofdstuk 4 wordt voor responsen die met behulp van een computer kunstmatig zijn gegenereerd onder een van de drie bovengenoemde families van modellen (PCM, GRM en SM), onderzocht of het mogelijk is het model te herkennen volgens welk zij zijn gegenereerd. Voor dit onderzoek wordt een methode uit de signaaldetectietheorie gebruikt. Bij deze methode worden steeds twee 'signalen' (responsen) gegenereerd die afkomstig zijn uit verschillende modellen. In de helft van de gevallen behoren de twee modellen in de simulatiestudie tot dezelfde familie, in de andere gevallen niet. Voor de manier waarop de onderzochte modellen zijn geselecteerd wordt verwezen naar hoofdstuk 4. Bij eerdere toepassingen van deze methode (Maydeu-Olivares et al., 1994; Van Engelenburg, 1997) bleek het moeilijk de twee gegenereerde responspatronen correct te classificeren. In deze eerdere onderzoeken werd steeds geprobeerd één responsvector te classificeren, dat wil zeggen de gesimuleerde responsen van één persoon op een hele test. In het huidige onderzoek worden niet de responsen van één persoon maar de gesimuleerde responsen van $n$ personen geclassificeerd. Het blijkt dan dat twee modellen afkomstig uit eenzelfde familie moeilijker van elkaar te onderscheiden zijn dan twee modellen afkomstig uit verschillende families. Dit leidt tot de conclusie dat de drie onderzochte families van modellen behalve theoretisch, ook praktisch gezien bestaansrecht hebben.

In de epiloog tenslotte wordt het belang van de scoringsregel voor de modelkeuze aangestipt. Voorts wordt daar geconstateerd dat er voorlopig een spanning blijft bestaan tussen het verlangen naar een model dat enerzijds de werkelijkheid goed beschrijft, en anderzijds wiskundig eenvoudig te hanteren is.

# Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F. (Eds.), *2nd International symposium on information theory*, pages 267–281. Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on automatic control*, *AC-19*, 716–723.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimates. *Journal of the Royal Statistical Society, series B*, *32*, 283–301.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, series B*, *34*, 42–54.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81.

Andersen, E. B. and Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, *42*, 357–374.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Andrich, D. (1982). An extension of the Rasch Model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105–113.

Andrich, D. (1985). artikel met die dependence afstanden. *Psychometrika*, *999*, 111–111.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19*, 101–119.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347–365.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord., F. M. and Novick, M. R. (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443–459.

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Chang, H. H. and Mazzeo, J. (1994). The unique correspondence of the item response function and the item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391–405.

Cliff, N. (1983). Evaluating Guttman scales: Some old and new thoughts. In Wainer, H. and Messick, S. (Eds.), *Principals of modern psychological measurement. A Festschrift for Frederic M. Lord*, pages 283–301. Hillsdale, NJ: Lawrence Erlbaum Associates.

De Vries, H. H. (1988). Het partial credit model en het sequentiele Rasch model met stohastisch design. Technical report, Universiteit van Amsterdam, Amsterdam, the Netherlands. (The partial credit model and the sequential Rach model with stochastic design).

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* SIAM Monograph No.38. Philadelphia: Society for industrial and applied mathematics.

Ellis, J. L. and van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika*, *58*, 417–429.

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

Fischer, G. H. (1995). The derivation of polytomous Rasch models. In Fischer, G. H. and Molenaar, I. W. (Eds.), *Rasch models: foundations, recent developments, and applications*, chapter 16, pages 293–305. New York: Springer.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, *7*, 149–153.

Ford, L. R. J. (1957). Solution of a ranking problem from binary comparisons. *American Mathematical Monthly*, *64*, 241–252.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, *85*, 1–11.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo Maximum Likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, series B*, *54*, 657–699.

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Gustafsson, J. E. (1979). *PML: a computer program for conditional estimation and testing in the Rasch model for dichotomous items.* University of Gotheburg, Institute of Education. Report Nr. 85.

Guttman, L. (1950). The basis for scalogram analysis. In Stouffer, S., Guttman, L., Suchman, E., P.F.Lazarsfeld, S.A.Star, and J.A.Clausen (Eds.), *Measurement and Prediction. Studies in social psychology in World War II, Vol. IV*, pages 60–90. Princeton: Princeton University Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains, and their applications. *Biometrika*, *57*, 97–109.

Haughton, D. (1996). A review of some aspects of information criteria for model selection. *Kwantitatieve Methoden*, *17*(52), 53–68.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., and Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*, 679–694.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., and Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331–348.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika*, *59*, 111–119.

Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. *Psychometrika*, *61*, 31–39.

Jackson, D. N. (1984). *Personality Research Form.* Port Huron, MI: Research Psychologists Press. .

Jansen, P. G. W. and Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika*, *51*, 69–91.

Kendall, M. and Stuart, A. (1979). *The advanced theory of statistics.* New York: Macmillan. Vol.2, 4th ed.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* New York: Wiley.

Levine, M. V., Drasgow, F., Williams, B., McCusker, C., and Thomasson, G. L. (1992). Measuring tyhe difference between two models. *Applied Psychological Measurement*, *16*, 261–278.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, No. 140.* Partly also in: G. F. Summers (ed). Attitude measurement, Chicago: Rand McNally, 1970.

Liou, M. (1994). More on the computation of higher-order derivatives of the elementary symmetric functions in the Rasch model. *Applied Psychological Measurement, 18*, 53–62.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NY: Lawrence Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*, 233–245.

Martin-Löf, P. (1973). Statistika modeller. Anteckningar från seminarier läsaret 1969-1970 utarbetade av Rolf Sundberg (Statistical Models. Notes from seminars 1969-1970 by Rolf Sundberg). Institutet för försäkringsmatematik och matematisk statistik vid Stockholms universitet.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Maydeu-Olivares, A., Drasgow, F., and Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18*, 245–256.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, series B, 51*, 127–138.

Mellenbergh, G. (1995). Conceptual Notes on Models for Discrete Polytomous Item Responses. *Applied Psychological Measurement, 19*, 91–100.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087–1092.

Mislevy, R. J., Johnson, E. G., and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154.

Mokken, R. J. (1970). *A Theory and Procedure of Scale Analysis. With applications in political research.* The Hague: Mouton.

Mokken, R. J. and Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417–430.

Molenaar, I. W. (1983). Item Steps. Heymans Bulletin HB-83-630-EX, University of Groningen, Vakgroep Statistiek en Meettheorie FSW, Grote Kruisstraat 2/1, Groningen, The Netherlands.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*, 165–181.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E. (1993). Information Functions of the Generalized Partial Credit Model. *Applied Psychological Measurement*, *17*, 351–363.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical recipes in C. The art of scientific computing* (Second ed.). Cambridge: Cambridge University Press.

Proth, J. M. and Hillion, H. P. (1990). *Mathematical Tools in Production Management.* New York: Plenum Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Reprinted in 1980 by The University of Chicago Press).

Rigdon, S. E. and Tsutakawa, R. K. (1986). Estimation for the Rasch model when both ability and difficulty parameters are random. *Journal of Educational Statistics*, *12*, 76–86.

Ripley, B. (1987). *Stochastic Simulation.* New York: Wiley.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*, 349–359.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, *No. 17.*

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph Supplement, No. 17.*

Samejima, F. (1997). Graded response model. In Van der Linden, W. J. and Hambleton, R. K. (Eds.), *Handbook of modern item response theory*, pages 85–100. New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*, 461–464.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Sijtsma, K. and Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79–105.

Thissen, D. (1988). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory.* Scientific Software Inc, Mooresville, USA. (Version 5.1)[Computer program].

Thissen, D. and Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*, 397–412.

Thissen, D. M. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55.

Tutz, G. (1997). Sequential models for ordered responses. In Van der Linden, W. and Hambleton, R. (Eds.), *Handbook of Modern Item Response Theory*, pages 139–152. New York: Springer.

Van der Ark, A. (1994). Aanpassing van de PRF-E. Item-bias detectie volgens de schenderstheorie bij een meervoudige persoonlijkheidsvragenlijst. Master's thesis, University of Amsterdam, Vakgroep Psychologische Methoden. (In Dutch).

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item respone theory.* PhD thesis, University of Amsterdam.

Verhelst, N. (1995). A Rasch model for continuous responses. Arnhem, The Netherlands: National Institute for Educational Measurement (Cito).

Verhelst, N. and Verstralen, H. H. F. M. (1991). The partial credit model with non-sequential solution strategies. Measurement and Research Department Reports 91-5, CITO, Arnhem, The Netherlands.

Verhelst, N. D., Glas, C. A. W., and de Vries, H. H. (1997). A steps model to analyze partial credit. In Van der Linden, W. J. and Hambleton, R. K. (Eds.), *Handbook of Modern Item Response Theory*, pages 123–138. New York: Springer.

Verhelst, N. D., Glas, C. A. W., and van der Sluis, A. (1984). Estimation problems in the Rasch model: the basic symmetric functions. *Computational Statistics Quarterly*, *1*, 245–262.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Wichmann, B. A. and Hill, I. D. (1982). An efficient and portable pseudo-random number generator. Algorithm AS 183. *Applied Statistics. Journal of the Royal Statistical Society, series C*, *31*, 188–190.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, *12*, 353–364.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Scientific Software Inc, Chicago, USA. [Manual for computer program].