COMPARISON OF EXAMINATION GRADES USING ITEM RESPONSE THEORY: A CASE STUDY

Oksana B. Korobko

Promotiecommissie:

Promotores:
Prof. dr. C.A.W. Glas

Assistent Promotor:
Dr. ir. B.P. Veldkamp

Referent:
Dr. J.W.Luyten

Overige leden:
Prof. dr. H. Kelderman
Prof. dr. W.J. van der Linden
Prof. dr. C.W.A.M. Aarts
Prof. dr. K. Sijtsma

Comparison of Examination Grades
using Item Response Theory:
a Case Study
O.B. Korobko
Ph.D. thesis
University of Twente
The Netherlands
7 September 2007

ISBN: 978-90-365-2527-5

COMPARISON OF EXAMINATION GRADES USING ITEM RESPONSE THEORY: A CASE STUDY

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus
prof. dr. W.H.M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 7 september 2007 om 16.45 uur

door

Oksana Borisovna Korobko
geboren op 1 Februari 1973
te Kherson, Oekraïne

Dit proefschrift is goedgekeurd door promottoren:

Promotor: prof. dr. C.A.W. Glas
Ass. Promotor: dr. ir. B.P. Veldkamp

To Emily

Acknowledgment

The presentation of this thesis is an indicator that an important milestone in my scientific journey has been reached. Now, I have a great opportunity to look back on my achievement at the University of Twente, and to express my gratitude to all those people who guided, supported and stayed beside me during my PhD study. First of all, I would like to thank Prof. Dr. Cees A.W. Glas. It has been a privilege to be under the guidance of such knowledgeable, inspiring, and patient supervisor. His motivation, enthusiasm and real help always inspired me to reach more in science. I appreciate Bernard P.Veldkamp for his guidance, help and corrections during my last year of my Ph.D. study. Also, I would like to thank Roel J. Bosker who gave me the opportunity to start my Ph.D. study, and supervised me on my first year.

My first few years as a Ph.D. student were completed at the O&M department and I am very grateful to all people with whom I worked together, especially: to my roommates Melanie Ehren and Karin Falkenburg; to the secretaries of O&M department (at that time) Lisenka van het Reve and Carola Groeneweg for their always help; to Hans Luyten, who helped me with the data collection; and also to Marinka Kuijpers, Ralf Maslowski, Maria Hendriks, Lyset Rekers-Mombarg, Bob Witziers, Rien Steen, Adrie Visscher, Elvira Annevelink, Birgit Schyns, and Gerdy ten Bruggencate for making my stay in the department so pleasant. I have very much appreciated the opportunity to complete my Ph.D study in OMD department for my last year. I would like to thank all my colleagues for their support and friendly environment in the department, especially: Jonald Pimentel and his family for friendship and support through all my years as a Ph.D student; to Anna Dagohoy, Leonardo Sotaridona as my fellow Ph.D. students during my first years; to Anke Weekers, Naveed Khalid, Caio Azevedo, Rinke Klein Entink, Hanneke Geerlings, Caroline Timmers en Iris Egberink for help, outings and conversations we had. I also

thank secretaries of OMD department Birgit and Lorette for their help.

I very appreciate my all friends: Oksana and Roman Stepanyan for their friendship, always help, mutual assistance and understanding; Andre Zvelindovsky for all the best what he has done for me; Irina Shostak for friendship and unselfish help before and during my Ph.D. study; Mikhael and Marina Scherb for their humor and friendship; Sveta and Rob Van Dijk for their pleasant company and friendship; Oksana Ribak my best friend from my Master student's time till now, in spite of distance. I am very thankful to my parents who provided me the opportunity to get the university degree at very difficult time for my country. I would like to thank my sister Victoria and her family for support and understanding. Also, I am very grateful to my husband's all big family for their hospitality during our visits and friendship. Finally, I would like to dedicate this dissertation to my lovely daughter Emily and my husband Sasha. Their love and support always cheer me up and motivate me to finish this work. I am very grateful for your understanding and patience.

Oksana B. Korobko
Enschede, September 2007

Contents

List of Tables

1

Introduction

Psychometrics is the theory of educational and psychological measurement. It concerns the measurement of knowledge, abilities, attitudes, and personality traits. Psychometric measurement is primarily concerned with the study of differences between individuals and between groups of individuals and has been used in psychology, health and educational research.

In educational science methods for the comparison of student achievement, school effectiveness and school differences can be based on school grades. However, the fact that there is many variation among subjects, courses, teachers, instructors and grading standards makes comparison of student's achievement difficult. In this thesis, we choose the Grade Point Average (GPA) on final examinations in the Netherlands as an example. Students can choose different subjects for their final examination, so they have different examination packages. Therefore, GPAs need a standardization that accounts for the difficulty of subjects and the proficiency of students. Using this data set as a guiding example, the problem is studied from a variety of perspectives.

There are many methods for standardization of GPA. They can be roughly divided into two groups: observed score methods (Kelly, 1976; Elliot & Strenta, 1988; Caulkins, Larkey & Wei, 1996; Smits, Mellenbergh & Vorst, 2002) and IRT-based methods (Young, 1990, 1991; Johnson, 1997, 2003). This research mostly will be focussed on IRT-based methods as they are more recent. IRT methods separate the influence of the difficulty level of the examination subjects and the proficiency level of the students via the introduction of item difficulty parameters and latent proficiency parameters. First, it will be assumed that the grades on all subjects can be

explained using a unidimensional representation of proficiency of students. Usually IRT models apply to discrete data (Rasch, 1960; Samejima, 1969; Bock, 1972; Lord, 1980; Masters, 1982). However, in some situations responses to the items may be continuous. For example, in this study the original data are continuous examination grades from 0 till 10 with two decimal places. IRT models for continuous responses are outlined by such authors as Mellenbergh (1994), Moustaki (1996) and Skrondal and Rabe-Hesketh (2004). The results obtained using unidimensional IRT models for both continuous and discrete data will be compared with a well-established observed score standardization method proposed by Kelly (1976).

In many situations, it may be plausible that there is more than one proficiency factor underlying the grades. For instance, there might be a specific proficiency factor for the science subjects and another one for language subjects. Therefore, it will be investigated whether the introduction of an IRT model with Q proficiency dimensions results in a better model for the grades. The IRT model is equivalent to a factor analysis model. The correlation between the proficiency factors represent the extent to which the proficiency dimensions are dependent. A high positive value for the factor loading means that the q-th dimension is important for the subject, a value close to zero means that the dimension does not play an important role. First a simple structure of factor loadings will be introduced where each examination is loading on one dimension only. The unidimensional subscales were searched for with the program OPLM (Verhelst, Glas & Verstralen, 1995). The pattern of loadings is both used for a categorical and a continuous interpretation of the data. Next, it will be investigated whether some subjects may be loading on more than one dimension. This more complicated factor structure will be investigated first for discrete data in combination with marginal maximum likelihood (MML) estimation.

Up till this point, the interaction between the choice of an examination subject and the proficiency parameters has not been taken into account. Implicitly, this means that it is assumed that the missing data process can be ignored. That is, it is assumed that the missing values (the grades on the examinations subjects that were not taken) are missing at random and the parameters of the distribution of the observed data and the distribution of the missing data indicators are distinct (Rubin, 1976). Free choice of examination subjects may however lead to a stochastic design that might violate the assumption of ignorability. If ignorability does not hold, the inferences made using an IRT model ignoring the missing data process can be severely biased (Bradlow & Thomas, 1998; Holman & Glas, 2005). Several authors have shown that selection bias can be removed when the distribution of missing

data indicator is modeled concurrently with the observed data using an IRT model (Moustaki & O'Muircheartaigh, 2000; Moustaki & Knott, 2000; Holman & Glas, 2005). Therefore, the multidimensional IRT model was enhanced with a so-called selection model for the missing data indicators.

As it was mentioned above, most IRT models pertain to discrete data. A unidimensional IRT model for continuous item responses (Mellenbergh, 1994) has been taken as a basis for developing an MML estimation and testing procedure for a multidimensional IRT model for continuous data. The Lagrange Multiplier (LM) test by Aitchison and Silvey (1958) is applied to evaluate the underlying assumptions of subpopulation invariance, the form of the item response function, local stochastic independence and the factor structure of the model. As an example of the proposed methods an analysis of one of the biggest packages of total data set is presented. Further, a number of simulation studies were carried out to assess the Type I error rate and the power of the proposed LM tests.

The thus far outlined studies have been done in the framework of marginal maximum likelihood (MML). As an alternative, a Bayesian framework is considered. A comprehensive estimation method using a Markov chain Monte Carlo (MCMC) computational method is developed that can simultaneously estimate the parameters for models for discrete and continuous responses for a broad class of models. The method combines approaches by Shi and Lee (1998), Béguin and Glas (2001) and Fox and Glas (2001,2002,2003). An analysis of the scaling of students' scores on a number of examination subjects is presented as an example of the proposed method. The data set used for this research contains grades of the students which are nested in different schools. One of the research questions addressed was how much of the variance in the students' proficiency is attributable to the schools. Therefore, the MCMC analysis of the IRT models was done with a two-level model for the proficiency parameters. That is, the overall covariance matrix was partitioned into a within schools covariance matrix and a between schools covariance matrix. The intra class correlation coefficients, which are the proportion of between school variance to the total variance, give the information about the proportion of variance attributable to the schools (see, for instance Bryk and Raudenbush, 1992). Another research question investigated concerned the proportion of variance attributable to gender. A second analysis was carried out with gender as a predictor for each of the four proficiency dimensions.

1.1. Overview of the Thesis

The chapters in this thesis are self-contained, hence they can be read separately. Therefore, some overlap could not be avoided and the notations, the symbols and the indices may slightly vary across chapters.

In Chapter 2, three methods for obtaining estimates of adjusted GPAs are discussed: a method proposed by Kelley (1976), an IRT model with a unidimensional representation of proficiency, and a multidimensional IRT model with a simple structure multidimensional representation of proficiency. For all three methods, the grades are either interpreted as continuous or categorical. The performance of the methods is investigated using data from the Central Examinations in Secondary Education in the Netherlands. Though the multidimensional IRT model fit the data significantly better than the other models, all three methods produced very similar results. The impact of the schools on the outcome data is small.

Chapter 3 presents three methods for the estimation of proficiency measures that are comparable over students and subjects based on IRT: a method based on a model with a unidimensional representation of proficiency, a method based on a model with a multidimensional representation of proficiency and a method based on a multidimensional representation of proficiency where the stochastic nature of the choice of examination subjects is explicitly modelled by a selection model. The results of the comparison using the data from the Central Examinations in Secondary Education show that the unidimensional item response model produces unrealistic results, which do not appear when using the two multidimensional IRT models. Further, it is shown that both multidimensional models produce acceptable model fit. However, the model that explicitly takes the choice process into account produces the best model fit.

Chapter 4 presents MML estimation and testing procedures for IRT models for continuous data. The model assumptions evaluated are subpopulation invariance (the violation is often labeled differential item functioning), the form of the item response function, local stochastic independence and the factor structure of the model. An analysis pertaining to scaling the students' grades is given as an example of the methods proposed. A number of simulation studies is presented that assess the Type I error rate and the power of the proposed tests.

In Chapter 5 a comprehensive Bayesian estimation method using a Markov chain Monte Carlo (MCMC) computational method was developed that can be used to simultaneously estimate the parameters for models for discrete and continuous re-

sponses. To illustrate the estimation procedure, estimates of both a model without and with a selection model are presented. Finally, it will be shown how the proportion of variance in the grades explained by the students' schools and the effect of covariates (in this case Gender) can be estimated.

Finally, a summary of the main results is given and some suggestion for further research are made.

2

Comparing School Performance using Adjusted GPA Techniques

ABSTRACT: Methods are presented for comparing school performance using the grades obtained on final central examinations where students choose different subjects. It must be expected that the comparison between the grades is complicated by the interaction between the students pattern and level of proficiency on one hand, and the choice of examination subjects on the other hand. Three methods for obtaining estimates of school performance adjusting for this interaction are discussed: a method proposed by Kelley (1976), an item response model (IRT) with a unidimensional representation of proficiency, and multidimensional IRT model with simple structure multidimensional representation of proficiency. For all three methods, the grades are either interpreted as continuous or categorical. The performance of the methods is investigated using data from the Central Examinations in Secondary Education in the Netherlands. Though the multidimensional IRT model fit the data significantly better than the other models, all three methods produced very similar results. The impact of the schools on the outcome data is insignificant, but for discrete data and multidimensional models differences between schools almost vanished.

This chapter has been submitted for publication as: O.B. Korobko, B.P. Veldkamp, and C.A.W. Glas, Comparing school performance using the adjusted GPA techniques

2.1. Introduction

School effectiveness research and the trend towards public reporting of school final grades have given rise to a need for value added measures of school performance, in which the average student achievement of schools is corrected for differences between the students at school entry (Fitz-Gibbon, 1994; Willms, 1992). Differences between average grades obtained in the final examination play a role to assess the achievement of each school. The analysis of school performance are usually done in the framework of multilevel modelling techniques (c.f. Goldstein, 1995; Snijders & Bosker, 1999). The grade point average (GPA) on examinations is often entered as a variable in these models. However, if the students have different examination packages, GPAs are probably not comparable. The main problem with using GPAs as proxies for educational achievement is the incorrect assumption that all course grades mean essentially the same thing. However, there is always substantial variation among topics, courses, teachers, instructors and grading standards. A related problem is that students generally choose subjects that fit their proficiency level. One of the problems addressed here is whether the fact that students generally choose the examination subjects in which they feel competent distorts the comparison of average examination results between schools and whether GPAs need a standardization over subjects that accounts for the confounding of the difficulty of subjects and the proficiency of students.

Methods for standardization of GPAs can be roughly divided into two classes: observed score methods (Kelly, 1976; Elliot & Strenta, 1987, 1988; Caulkins, Larkey & Wei, 1996; Smits, Mellenbergh & Vorst, 2002) and IRT-based methods (Young, 1990, 1991; Johnson, 1997, 2003). Kelly (1976) proposes an heuristic method to re-scale the grades in such a way that the GPAs of the subjects are the same in a situation where all students take all examinations and all examinations have the same difficulty. The method by Smits, Mellenbergh and Vorst (2002) does not re-scale the observed responses but imputes unobserved grades accounting for the difficulty of the examination topics and the overall proficiency level of the students. Smits, Mellenbergh and Vorst, (2002) compared seven different missing grade imputation methods. The simple GPA-adjustment techniques produced unrealistic results for imputed grades, since imputed values for some subjects were higher than the observed values. More complicated imputation techniques, like Multiple Imputation (MI) produced more realistic results. Also Schafer & Olsen (1998) pointed out that simple mean substitution can seriously dampen relationships among variables.

IRT-based methods (Young, 1990, 1991; Johnson, 1997, 2003) separate the influence of the difficulty level of the examination topics and the proficiency level of the students via the introduction of difficulty parameters and latent proficiency parameters. This may have two drawbacks. First, the used IRT models pertain to discrete observations while the grades may be better represented as continuous responses. And second, proficiency may not be unidimensional at all. Therefore, the present article investigates the impact of using IRT models with a multidimensional representation of proficiency, and the impact of using discrete or continuous representation of grades.

This article is organized as follows. After this section, an example of an observed score method, the method proposed by Kelly (1976) and IRT-based methods are presented. The methods will be compared using data from the Central Examinations in Secondary Education in the Netherlands, which were collected by Dutch Inspection of Education. The methods will be used for a comparison schools. Finally, the last section gives a discussion and some conclusions.

2.2. Design and Methods

Data are used from 6,142 approximately 17-year old students in pre-university schools in the Netherlands, the only curriculum track (of the four available) that prepares students for direct entry into a university. The data were collected by the Inspectorate of Education. The students sit examinations in 6 or 7 subjects to be chosen from a total of 16. These external examinations are based on standardized achievement tests, and for this study only the results from the first session are used (unsatisfactory marks might be “repaired” in a re-session).

Our analysis relate to a subset of the pre-university students that took their final examination in the school year 1994/1995. The original data set comprised 16,118 students. Students that did not take an examination in both Dutch and English were excluded from the analysis. Furthermore, students taking an examination in one of the “unusual” subjects (see Table 2.1) were excluded as well. However, most students were excluded to restrict the analysis to 60 fairly common combinations of examination subjects out of a potential 8,000. The students that had chosen one of the 25 most common combinations were included, but none of the 25 most common combinations included the subjects Classical Greek or Fine Art and only one combination included Latin. Extra students were added in order to make sure that the data set contained sufficient information on these three subjects as well.

Table 2.1: Usual and unusual subjects; percentage of students taking an examination

Usual subjects		Unusual subjects	
Subjects	Percentage	Subjects	Percentage
Dutch language	99.9	Frisian language	0.0
Latin	14.6	Russian	0.0
Classical Greek	6.2	Spanish	0.2
French	37.6	Handicrafts	1.9
German	45.4	Music	1.6
English	99.1	Philosophy	0.7
History	49.5	Social studies	2.3
Geography	33.9		
Applied Math	63.0		
Advanced Math	44.7		
Physics	46.7		
Chemistry	38.2		
Biology	37.0		
General Economy	58.7		
Business Economy	36.0		
Arts	7.8		

These were the students with the 10 most common combinations of Latin with other subjects (except for the one already included), the students with the 13 most common combinations of Greek with other subjects (one of these also included Latin) and the 12 most common combinations of Fine Art with other subjects.

Given the subjects chosen, we can distinguish three groups of students:

1. The linguistically oriented students (20%). These students take examinations in French and German languages and not more than one of the subjects Applied Mathematics, Advanced Mathematics, Physics and Chemistry.

2. The science oriented students (33%). These students take examinations in 2at least three of the subjects Applied Mathematics, Advanced Mathematics, Physics and Chemistry and no examinations in French or German languages.

3. Other students (47%).

One might view the problem of comparing the difficulty of examination as a test

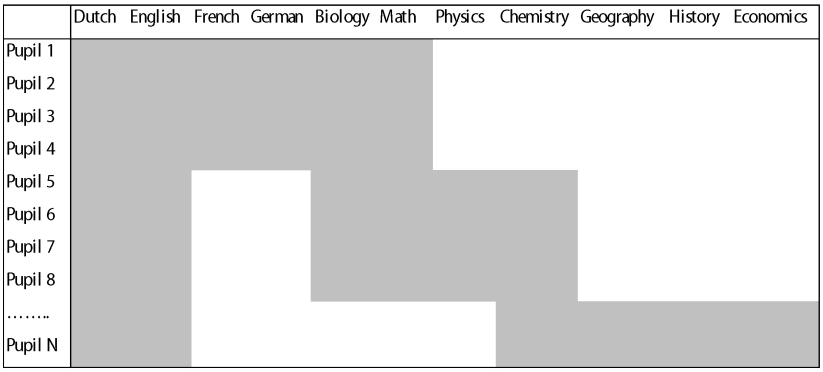

Figure 2.1: Design of the study

equating problem, with an incomplete design with 60 tests, in which each subject is an item. The "anchor items" in this study are the subjects Dutch Language and English language, that are taken by all students. The design is graphically depicted in Figure 2.1. We restrict ourselves in this example to 3 very simple combinations of 6 out 11 subjects.

2.2.1. Methods Based on Item Response Theory

An IRT Model for Categorical Data

The original examinations grades are categorized into four categories labelled $j = 0, ..., m_i$, where $m_i = 3$. The original grades ranged from 1 ("poor") to 10 ("excellent"), but for the purpose of this study they were re-scaled to a four point scale, where the points are 0 (original grade 0 to 5.4, which is unsatisfactory), 1 (original grade 5.5 to 6.4, which is just satisfactory), 2 (original grade 6.5 to 7.4, which is good), and 3 (original grade 7.5 to 10, which is very good).

The data will be analyzed using the generalized partial credit model (Muraki, 1992). For the unidimensional case, it is assumed that the probability that the grade of student n ($n = 1, ..., N$) on examination subject i ($i = 1, ..., K$), denoted by X_{ni}, is

in category j is given by

$$\Pr(X_{ni} = j | d_{ni} = 1) = \frac{\exp\left(j\alpha_i\theta_n - \sum_{h=1}^{j}\beta_{ih}\right)}{1 + \sum_{h=1}^{m}\exp\left(h\alpha_i\theta_n - \sum_{p=1}^{h}\beta_{ip}\right)}, \tag{2.1}$$

where θ_n is the unidimensional proficiency parameter that represents the overall proficiency. So it is assumed here that one unidimensional proficiency parameter θ can explain all examination grades. The parameters β_{ij} $(j = 1, ..., m_i)$ model the difficulty of examination subject i, and the parameter α_i defines the extent to which the probability is related to the proficiency θ. Following Bock and Zimowski (1997), it will be assumed that distinct groups of students have distinct normal distributions of their proficiency parameters θ. In the present case, it is assumed that every group of students taking a specific examination package have a normal proficiency distribution with a specific mean. The variance is the same for all groups. The parameters are estimated using maximum marginal likelihood (see, Bock & Aitkin, 1981).

An IRT Model for Continuous Data

The examination grades originally range from 0 till 10 with two decimal places. They can be analyzed with IRT models for continuous responses as outlined by such authors as Mellenbergh (1994), Moustaki (1996) and Skrondal and Rabe-Hesketh (2004). These models are equivalent to a unidimensional factor model. Consider a two-dimensional data matrix X with entries x_{ni}, for $n = 1, ..., N$, and $i = 1, ..., K$. The matrix contains the responses of students to items. It is assumed that the response of the student n on the item i is normally distributed, that is

$$P(x_{ni} \mid \theta_n, \alpha_i, \beta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{(x_{ni} - \tau_{ni})^2}{2\sigma_i^2}\right). \tag{2.2}$$

The expectation of the item response is a linear function of the explanatory variables,

$$\tau_{ni} = \alpha_i\theta_n - \beta_i \tag{2.3}$$

where α_i is a factor loadings and β_i is a location parameter. We assume that the density of person parameter θ_n is a normal distribution with the expectation μ_θ and the variance σ_θ. Further, we assume that the variance $\sigma_i^2 = 1$, for all i. That is, we assume that all the observed responses have the same scale. The parameters can, for instance, be estimated using maximum marginal likelihood estimation as implemented in the M-plus program (Muthén & Muthén, 2003).

Multidimensional IRT Models for Categorical and Continuous Data

In the previous models it was assumed that the probability of the grades of student n on examination subject i is by (2.1) and (2.2) for categorical responses and continuous responses, respectively. However, there may be more than one factor underlying the examination grades. For instance, there might be a special proficiency factor for the science proficiency and another one for the language proficiency. Of course, it must be expected that these factors correlate positively, and probably quite high. If Q proficiency dimensions are needed to model the grades, the proficiency of student n can no longer be represented by a unidimensional scalar θ_n, but must be represented by a vector of proficiency $\left(\theta_{n1}, ..., \theta_{nq}, ..., \theta_{nQ}\right)$. The probability of a grade in category j is now given by

$$\Pr\left(X_{ni} = j\right) = \frac{\exp\left(j\left(\sum_{q=1}^{Q} \alpha_{iq}\theta_{nq}\right) - \sum_{h=1}^{j} \beta_{ih}\right)}{1 + \sum_{h=1}^{m} \exp\left(h\left(\sum_{q=1}^{Q} \alpha_{iq}\theta_{nq}\right) - \sum_{p=1}^{h} \beta_{ip}\right)}. \tag{2.4}$$

For continuous responses, the expectation of the item response is given by

$$\tau_{ni} = \sum_{q=1}^{Q} \alpha_{iq}\theta_{nq} - \beta_i = \alpha_i'\theta_n - \beta_i,$$

where α_i is a vector, which are usually called factor loadings and β_i is a location parameter. Both for the categorical and continuous model, we assume that the density of θ_n is described by a Q-variate normal distribution with a covariance matrix Σ_θ. The correlation between the proficiency dimensions that are parameters of this multivariate normal distribution represent the extent to which the dimensions are dependent. In addition, it will be assumed that the proficiency parameters of groups of students taking a specific package of examination subjects have specific means. So it will be assumed that the mean of these distributions depends on the package and that the covariance matrix of the proficiency parameters is common over groups.

Takane and de Leeuw (1987) show that the model for categorical data is equivalent with a full-information factor analysis model. Therefore, the parameters $\alpha_{i1}, ...,$ α_{iQ} are often called factor-loadings, and the proficiency parameters $\theta_{n1}, ..., \theta_{nq}, ..., \theta_{nQ}$ can be viewed as factor scores. Note that the factor loadings are specific for an examination subject and they model the relation between the probability of obtaining a grade and the level on the Q proficiency dimensions. A high positive value of α_{iq}

means that the q-th dimension is important for the subject, a value close to zero means that the dimension does not play an important role. Finally, the relation between the Q proficiency parameters is modelled by assuming that the proficiency parameters $\theta_1, ..., \theta_q, ..., \theta_Q$ are independent between persons and, for every person drawn from a Q-variate normal distribution with a mean μ and a covariance matrix Σ. To identify the model, it is will be assumed that the mean of the proficiency parameters $\theta_1, ..., \theta_q, ..., \theta_Q$ of the first package is equal to zero. For further identification restrictions refer to Béguin and Glas (2001). In the present application a simple structure of factor loadings was used, that is it: each item is loading on one dimension only.

2.2.2. Kelly's Method

The IRT-based methods will be compared with a method proposed by Kelly (1976). This method gives us the standardization of the subject grades such that the difficulty of subjects and the strictness of possible raters is corrected for. "Standardization is used to approximate a student's grade in a subject to that which would be obtained in the ideal situation when all students took all subjects, and all subjects were marked by the same examiners" (Kelly, 1976). The method is conditional on the students' total grades $\overline{x}_n = \sum_i d_{ni} x_{ni}$. That is, these grades are considered an estimate of overall proficiency and are not affected by the standardization. The students, subject grades x_{ni} are standardized to grades x^*_{ni} in such a way that the mean difficulties of the subjects become the same. So the method boils down to weighting the subjects in such a way that their difficulties are the same, without altering the total grade distribution.

Two algorithms are available to achieve this. Kelly (1976) proposed an iterative method. In each iteration, a consensus standard is established for each subject by equating the mean grade in that subject with the mean of the mean grades the same students obtained in all other subjects. Define

$$y_{ni} = \left[\sum_{j=1, j\neq i}^{K} d_{nj} x_{nj} \right] / \left[\sum_{j=1, j\neq i}^{K} d_{nj} \right], \tag{2.5}$$

So y_{ni} is the mean of the grades of individual n in the subjects endorsed, excluding subject i. The correction for subject i is defined as

$$\overline{\delta}_i = \overline{y}_i - \overline{x}_i$$

where $\overline{x}_i$ is the mean of the grades in each subject, that is,

$$\overline{x}_i = \left[\sum_{n=1}^{N} d_{ni} x_{ni}\right] / \left[\sum_{n=1}^{N} d_{ni}\right],$$

and $\overline{y}_i$ is the mean of the grades y_{ni}, that is,

$$\overline{y}_i = \left[\sum_{n=1}^{N} d_{ni} y_{ni}\right] / \left[\sum_{n=1}^{N} d_{ni}\right].$$

Then the students' subject grades are adjusted to obtain grades

$$x^*_{ni} = x_{ni} - \delta_i.$$

The process is re-iterated with these adjusted grades as input and the iterations are repeated until convergence. Note that the method re-weights the mean grades for each subject until they are the same, but for each student the mean grade remains the same. Therefore, the adjustments δ_i can be seen as the difficulties of the subjects. So, the correction indicates how difficult this subject is in relation to the other subjects. A positive correction indicates difficult subjects and negative correction indicates easy subjects.

Lawley (see, Kelly, 1976) has shown that this iterative procedure is equivalent to a set of linear equations that can be solved analytically. Both methods were used in the present article and the results were equivalent.

Kelly's method received criticism from Newton (1997), who argues that the method cannot be used to obtain the between-subjects comparisons. If the sample of students was divided into identifiable subgroups, such as male and female candidates, we would obtain different corrections for different subgroups. If these differences were statistically significant,this would invalidate the method, because grading does not take into account gender. According to Newton (1997) "these techniques would only be in the running as indices of between-subject comparability if our public examinations measured a different kind of quality to that which they currently assess" and further "The Subject-Pair Analysis (SPA) does not assume that factors such as motivation and teaching standards are comparable between subjects". The students demonstrate different level of achievement in different subjects, and Kelly's method can provide false conclusions concerning grading standards. Newton also criticizes the term "general academic ability" as used by Kelly. This problem of multidimensionality is easily solved within the framework of IRT.

2.2.3. Methods for Comparison the Schools

A basic measure for degree of dependency in clustered data (in our case the students nested in different schools) is the intraclass correlation coefficient. It gives the proportion of the variance in the students' grades attributable to the schools. The intraclass correlation coefficient (ICC) is defined as

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}, \tag{2.6}$$

where σ^2 stands for the within-schools variance and τ^2 stands for the between-schools variance (see, for instance, Snijders and Bosker, 1999). The sum $\tau^2 + \sigma^2$ is the total variance.

For each school the average examination grade per subject (averaging over students) is estimated using available observed grades and, if these are not available, the imputed expected grades based on the unidimensional and multidimensional IRT models for continuous and categorical data. For unobserved grades (subjects not endorsed by students) the grades were computed by first computing the posterior expectation and variance under the model. Because these expectations are in fact estimates, the uncertainty of these estimates must be taken into account. This was done by the method of plausible value imputation (see Mislevy, Beaton, Kaplan & Sheehan, 1992): for every unobserved subject of every student one value was drawn from its posterior distribution and the variance components and intraclass correlations were computed.

2.3. Results

2.3.1. Kelly's Method and Unidimensional IRT Model for Categorical and Continuous Data

The results of applying Kelly's method and unidimensional IRT models are given in Table 2.2 for categorical and Table 2.3 for continuous data, respectively. The first column in these tables presents the examination subjects. The second column presents the mean of the observed grades. The third column presents the correction δ_i for each of the subjects as obtained by Kelly's method. This correction can be interpreted as the difficulty of each subject. Most difficult subjects, such as Advanced Mathematics, Applied Mathematics and Physics, obtain positive corrections, and less

difficult subjects, such as Latin, Arts and French, obtain negative corrections. The correction for subjects like Dutch, English, Geography and Business Economics is near zero, so these subjects have a difficulty near the overall mean. In these tables, these corrections are given in decreasing order. The fourth column presents the corrected mean, which we obtained by applying Kelly's method. The next column shows the expected average examination grades, given the data, and computed using IRT models under the assumption that all students take all examinations. That is, if a student did not take a subject, an expected grade was imputed that was computed on the basic of the estimated proficiency of the student and the "item-parameters" of the subject. For unidimensional IRT models, both for continuous and categorical responses the mean of the expected grades are not much different from the observed grades. The last column presents the mean item parameter $\overline{\beta}$ obtained by IRT model. This mean can be seen as the overall location of the examination on the latent scale. The rank order of the item parameters are between brackets. The correction obtained by Kelly's method and the mean parameter $\overline{\beta}$ obtained by a unidimensional IRT model can be interpreted as the difficulty of the examination subject. The correlation between the correction obtained by Kelly's method and mean IRT parameters is very high, for categorical data correlation is 0.96, and for continuous data correlation is 0.88. It is interesting that Chemistry and Biology are in the top 5 of the most difficult subjects for continuous data, but for categorical data the correction for the mean of grades for these subjects are negative for Kelly's method.

Both Kelly's method and IRT models are based on models assuming an unidimensional proficiency structure. In the first method, the difficulty of the subjects is represented by the adjustment δ_i needed to scale the difficulty of the subjects, in the second method by expected grades computed under the assumption that all students took all subjects. In Tables 2.2 and 2.3 it can be seen that the rank orders of the corrections δ_i (the third column) and the item parameters under IRT Models are very similar. Further, it can be seen that the most difficult subjects as Advanced Math and the least difficult subject is Latin.

Several methods are available to obtain overall proficiency grades for students. Four methods were compared: EAP estimates of the ability parameters (denoted by $\hat{\theta}$), plausible values drawn from the posterior distribution (denoted by $\tilde{\theta}$), and expected GPAs evaluated using either $\hat{\theta}$ or $\tilde{\theta}$, denoted by GPA($\hat{\theta}$) and GPA($\tilde{\theta}$), respectively.

Table 2.4 shows the correlations between the methods. Correlations between observed (raw) GPA, expected GPA and proficiency estimates for continuous obser-

Table 2.2: Correction and Corrected means obtained by Kelly Method and Estimation Grades under 1-dimensional IRT Model (categorical data)

Subjects	Mean	Correction	Corrected Mean	Expected Grades	$\overline{\beta}$
Advanced Math	1.37	0.31	1.68	1.20	0.39(1)
Applied Math	1.16	0.22	1.38	1.23	0.28(3)
Physics	1.50	0.16	1.66	1.32	0.38(2)
General Economy	1.27	0.12	1.39	1.33	0.26(4)
Dutch	1.38	0.08	1.46	1.38	0.21(5)
English	1.50	-0.04	1.46	1.50	0.01(8)
Geography	1.31	-0.04	1.27	1.44	0.11(6)
Business Economy	1.41	-0.05	1.36	1.48	0.05(7)
Chemistry	1.76	-0.11	1.65	1.56	-0.10(9)
Biology	1.76	-0.11	1.65	1.62	-0.27(12)
German	1.51	-0.14	1.37	1.60	-0.18(10)
History	1.59	-0.22	1.38	1.66	-0.23(11)
Classical Greek	2.18	-0.22	1.98	1.86	-0.42(15)
French	1.64	-0.25	1.39	1.71	-0.29(13)
Arts	1.60	-0.36	1.24	1.67	-0.29(14)
Latin	2.48	-0.67	1.81	2.29	-1.05(16)

vations are given in the first part of this table. The correlation between Raw GPA and GPA($\hat{\theta}$) and GPA($\tilde{\theta}$) is very high, 0.98 and 0.97 respectively. The correlation between the estimates of proficiency $\hat{\theta}$ and the plausible values $\tilde{\theta}$ is 0.92. Overall the difference between the various estimation methods is quite high.

The second part of the table presents the analogous correlations under a discrete model. Correlation between Raw GPA and GPA($\hat{\theta}$) is very high, 0.99. Overall, the pattern is similar to the pattern for the continuous case: the correlations are quite high.

The bottom part of the Table 2.4 represents the correlation matrix between continuous and discrete raw GPA, expected GPA's and estimated proficiencies. Also here, the correlations are high and in most cases are more than 0.90.

Table 2.3: Correction and Corrected means obtained by Kelly Method and Estimation Grades under 1-dimensional IRT Model (continuous data)

Subjects	Mean	Correction	Corrected Mean	Expected Grades	$-\beta$
Advanced Math	6.32	0.53	6.85	6.16	6.01(1)
Physics	6.46	0.48	6.94	6.31	6.16(3)
Chemistry	6.77	0.19	6.96	6.61	6.47(9)
Biology	6.71	0.13	6.84	6.61	6.55(10)
General Economy	6.14	0.13	6.27	6.16	6.21(4)
Applied Math	6.02	0.07	6.10	6.04	6.12(2)
Dutch	6.30	0.06	6.35	6.30	6.30(5)
Business Economy	6.31	0.00	6.31	6.35	6.39(7)
English	6.42	-0.09	6.33	6.42	6.42(8)
Geography	6.24	-0.19	6.06	6.33	6.39(6)
German	6.48	-0.23	6.25	6.53	6.59(11)
History	6.55	-0.33	6.22	6.59	6.65(13)
Classical Greek	7.27	-0.33	6.93	6.95	6.97(15)
French	6.66	-0.43	6.24	6.72	6.77(14)
Arts	6.54	-0.46	6.08	6.62	6.63(12)
Latin	7.73	-0.88	6.85	7.54	7.53(16)

2.3.2. Comparison of the Results for Categorical and Continuous Multidimensional IRT Models

A multidimensional IRT model for discrete responses was fitted with a method by Béguin and Glas (2001). The method identifies the dimensions by fitting unidimensional IRT models by discarding items, or, in the present case, examination subjects. These examination subjects are entered as unique indicators of a dimension in the multidimensional IRT model, that is, these examination subjects load on one dimension only. The unidimensional subscales were searched for with the program OPLM (Verhelst, Glas & Verstralen, 1995). The R_{1c} statistic (Glas, 1988) was used as a criterion for model fit.

Using this partitioning of examinations into subscales, the parameters of the multidimensional model for discrete data were estimated using maximum marginal likelihood by a dedicated program, and the parameters of the multidimensional model for continuous data were estimated using maximum marginal likelihood estimation

Table 2.4: Correlations between raw GPA, expected GPA and proficiency estimated using unidimensional models

Continuous Observations					
	Raw GPA	GPA($\widehat{\theta}$)	GPA($\widetilde{\theta}$)	$\widehat{\theta}$	$\widetilde{\theta}$
Raw GPA	1.00				
GPA($\widehat{\theta}$)	0.98	1.00			
GPA($\widetilde{\theta}$)	0.97	0.98	1.00		
$\widehat{\theta}$	0.98	0.95	0.94	1.00	
$\widetilde{\theta}$	0.89	0.87	0.78	0.92	1.00
Discrete observations					
	Raw GPA	GPA($\widehat{\theta}$)	GPA($\widetilde{\theta}$)	$\widehat{\theta}$	$\widetilde{\theta}$
Raw GPA	1.00				
GPA($\widehat{\theta}$)	0.99	1.00			
GPA($\widetilde{\theta}$)	0.95	0.96	1.00		
$\widehat{\theta}$	0.95	0.96	0.92	1.00	
$\widetilde{\theta}$	0.83	0.85	0.93	0.87	1.00
Discrete by Continuous Observations					
	Continuous				
Discrete	Raw GPA	GPA($\widehat{\theta}$)	GPA($\widetilde{\theta}$)	$\widehat{\theta}$	$\widetilde{\theta}$
Raw GPA	0.96	0.95	0.93	0.94	0.86
GPA($\widehat{\theta}$)	0.96	0.93	0.92	0.96	0.88
GPA($\widetilde{\theta}$)	0.92	0.90	0.89	0.92	0.85
$\widehat{\theta}$	0.93	0.91	0.89	0.94	0.86
$\widetilde{\theta}$	0.81	0.80	0.78	0.83	0.76

by M-plus program (Muthén & Muthén, 2003). Table 2.5 gives us results of Multidimensional IRT models for continuous and categorical data. The table shows the extent to which the subjects depend on the proficiency level of three dimensions: Language, Science and Economy. The first column presents the subjects, the next three columns present the factor loadings α_{iq} for three dimensions Language, Science and Economy for categorical data and last three columns present the factor loading α_{iq} for three dimensions Language, Science and Economy for continuous data. The stars indicate fixed factor loadings. The categorical and continuous data have the same simple structure: each item is loading on one factor only. Highest loadings on

the Language dimension are obtained for German, French and English, and lowest loading on this dimension is for Dutch Language. This is probably due to the fact that Dutch is the mother tongue for the students and so that the specific linguistic component of this subject may be small. For the Science dimension, Physics, Chemistry and Advanced Mathematics have the highest loadings. For the Economy dimension, General Economy and Business Economy have the highest loadings. Arts loaded low on any dimension and was assigned to the third dimension. The results are analogous for categorical and continuous data.

Table 2.5: Factor Loading per Subjects for the 3-Factor Solution IRT (simple structure) and Correlation Matrices

	Categorical data			Continuous data		
Subjects	Language	Science	Economy	Language	Science	Economy
Dutch	0.49	0.00*	0.00*	0.22	0.00*	0.00*
Latin	0.82	0.00*	0.00*	0.39	0.00*	0.00*
Classical Greek	0.75	0.00*	0.00*	0.41	0.00*	0.00*
French	1.33	0.00*	0.00*	0.62	0.00*	0.00*
German	1.64	0.00*	0.00*	0.60	0.00*	0.00*
English	1.21	0.00*	0.00*	0.62	0.00*	0.00*
History	0.00*	0.00*	0.87	0.00*	0.00*	0.43
Geography	0.00*	0.85	0.00*	0.00*	0.36	0.00*
Applied Math	0.00*	0.74	0.00*	0.00*	0.56	0.00*
Advanced Math	0.00*	0.96	0.00*	0.00*	0.62	0.00*
Physics	0.00*	1.41	0.00*	0.00*	0.65	0.00*
Chemistry	0.00*	1.45	0.00*	0.00*	0.68	0.00*
Biology	0.00*	0.98	0.00*	0.00*	0.38	0.00*
General Economy	0.00*	0.00*	1.10	0.00*	0.00*	0.55
Business Economy	0.00*	0.00*	1.17	0.00*	0.00*	0.55
Arts	0.00*	0.00*	0.36	0.00*	0.00*	0.19
			Correlation matrix			
Language	1.00			1.00		
Science	0.52	1.00		0.50	1.00	
Economy	0.57	0.97	1.00	0.54	0.95	1.00

The correlation matrices between the dimensions are given at the bottom of the table. Note that the correlation between the Science dimension and the Economy dimension is very high: 0.97 for categorical data, and 0.95 for continuous data. Correlations between the other dimensions are much lower.

Table 2.6 presents estimated average grades for examination subjects under mul-

tidimensional IRT models for categorical and continuous data. The third and the fifth columns present mean item parameters $\overline{\beta}$ for categorical grades and β for continuous grades, respectively. Note that Latin was the least difficult subject, while Advanced Math and Physics are the most difficult subjects. The correlation between continuous and categorical expected grades was 0.96 and between continuous and categorical item parameters β is 0.98. That means that these two different IRT methods produced very similar results. Further, the rank orders of the subjects are very similar to the rank orders in Table 2.2 and Table 2.3.

Table 2.6: Examination grades and item parameters estimated under 3-factor IRT model

	Categorical data		Continuous data	
Subjects	Estimated Grade	$\overline{\beta}$	Estimated Grade	β
Dutch	1.38	0.22(5)	6.30	6.30(5)
Latin	2.32	-1.09(16)	7.49	7.43(16)
Classical Greek	1.91	-0.47(15)	6.94	6.91(15)
French	1.61	-0.20(11)	6.64	6.62(12)
German	1.49	-0.06(9)	6.47	6.45(10)
English	1.50	0.00(8)	6.42	6.42(9)
History	1.67	-0.24(12)	6.58	6.69(14)
Geography	1.42	0.12(7)	6.32	6.42(8)
Applied Math	1.20	0.30(4)	6.01	6.13(3)
Advanced Math	1.29	0.41(1)	6.19	5.95(1)
Physics	1.41	0.39(2)	6.36	6.11(2)
Chemistry	1.61	-0.11(10)	6.65	6.41(7)
Biology	1.64	-0.28(13)	6.63	6.52(11)
General Economy	1.29	0.31(3)	6.14	6.22(4)
Business Economy	1.42	0.13(6)	6.33	6.35(6)
Arts	1.67	-0.30(14)	6.63	6.64(13)

The fit of the IRT models (unidimensional and multidimensional) for continuous and categorical data was evaluated using likelihood ratio tests. A test of unidimensional IRT against multidimensional IRT for continuous data yielded a chi-square value of 1790.8 with 3 degrees of freedom. A test of unidimensional IRT against multidimensional IRT for categorical data yielded a chi-square value of 867.6, also with 3 degrees of freedom. So the multidimensional IRT models fitted better than the unidimensional IRT models. However, the impact of this better model fit as displayed in Table 2.6 was quite small.

Table 2.7 presents the correlations between observed (raw) GPA and expected

GPA estimated using multidimensional IRT models for continuous and categorical data. In this table Raw GPA pertains to observed grades only, GPA($\hat{\theta}$) pertains to expected GPA obtained using estimated proficiencies of students, and GPA($\tilde{\theta}$) pertains to expected GPA obtained using plausible values drawn from the posterior distributions of the parameters. The first three GPAs relate to the continuous IRT model, and the last three GPA's to the discrete IRT model. The correlation between Raw GPA and GPA ($\hat{\theta}$) for the continuous case is very high. For the discrete case the correlation between Raw GPA and GPA ($\hat{\theta}$) is lower, but still as high as high 0.95.

Table 2.7: Correlations between raw GPA and expected GPA estimated using multidimensional models

	Continuous			Discrete		
	Raw GPA	GPA($\widehat{\theta}$)	GPA($\widetilde{\theta}$)	Raw GPA	GPA($\widehat{\theta}$)	GPA($\widetilde{\theta}$)
Raw GPA	1.00					
GPA($\widehat{\theta}$)	0.99	1.00				
GPA($\widetilde{\theta}$)	0.95	0.96	1.00			
Raw GPA	0.96	0.94	0.90	1.00		
GPA($\widehat{\theta}$)	0.93	0.92	0.88	0.95	1.00	
GPA($\widetilde{\theta}$)	0.92	0.89	0.86	0.88	0.89	1.00

2.3.3. Estimation of Variance Attributable to Schools via Imputation

Finally, various estimates of the variance attributable to the schools were estimated using ICCs as defined by (2.6). The ICCs are shown in Table 2.8. Note that the ICC for continuous observed grades is highest: 0.080. All ICCs for the continuous grades estimated using unidimensional and multidimensional methods are lower. This suggests that the choice pattern of the examination topics is related to the school attended since the schools explain more observed variance than adjusted variance. Note that if we correct for the unreliability of the estimates by using plausible values, the ICCs decrease even further.

The impact of the school on the outcomes for the discrete grades is systematically lower than for the continuous grades. This means that categorization seems to attenuate the differences between schools. The overall conclusion is that the impact of the schools on the outcomes is not very large.

Table 2.8: Intra-class correlations estimated using different methods

	Continuous	Discrete
Raw GPA	0.0800	0.0740
	Unidimensional	
	Continuous	Discrete
GPA($\widehat{\theta}$)	0.0729	0.0662
GPA($\widetilde{\theta}$)	0.0704	0.0623
$\widehat{\theta}$	0.0712	0.0661
$\widetilde{\theta}$	0.0604	0.0526
	Multidimensional	
	Continuous	Discrete
GPA($\widehat{\theta}$)	0.0722	0.0584
GPA($\widetilde{\theta}$)	0.0675	0.0592
$\widehat{\theta}_1$	0.0719	0.0538
$\widetilde{\theta}_1$	0.0425	0.0534
$\widehat{\theta}_2$	0.0566	0.0413
$\widetilde{\theta}_2$	0.0392	0.0404
$\widehat{\theta}_3$	0.0684	0.0434
$\widetilde{\theta}_3$	0.0476	0.0431

2.4. Discussion and Conclusion

The problem addressed here concerned comparison of students and schools based on average examination grades. The complicating factor is that students only sit examinations in subjects they have chosen themselves. As a consequence more proficient students may choose examinations in subjects that are more difficult. Kelly's method and unidimensional IRT methods show very similar results, both for continuous and discrete grades. The rank order of the estimates of the difficulty of the examination subjects is very high. The most difficult subjects according these methods are Advanced Mathematics, Applied Mathematics and Physics, least difficult subjects are French, Arts and Latin.

However, it is not a-priori plausible that the proficiency structure assessed by the examinations is unidimensional. Three dimensional IRT models with a simple structure where each subject loads on one dimension only were considered. The results of the three factor models for categorical and continuous grades are very similar. Highest loadings on the Language dimension are attained by the examina-

tions in German, French and English Language. For the Science dimension, Physics, Chemistry and Advanced Math have the highest loadings. A third dimension had highest loadings for General Economy and Business Economy, and was therefore labeled as an Economy dimension. However, the correlation between the Science dimension and the Economy dimension is very high.

Overall, the multidimensional IRT model fitted the data significantly better than unidimensional IRT model, despite the fact that the obtained expected grades for multidimensional and unidimensional IRT models are very close.

A drawback of the methods discussed here is that every subject should load on one dimension only. Latin had a low loading on the Language dimension but could probably load on other dimensions also. So the next step is developing multidimensional IRT models that can support a more complicated factor structure.

3

Modelling the Choice of Examination Subjects

ABSTRACT: Methods are presented for comparing grades obtained in a situation where students can choose between different subjects. It must be expected that the comparison between the grades is complicated by the interaction between the students' pattern and level of proficiency on one hand, and the choice of the subjects on the other hand. Three methods for the estimation of proficiency measures that are comparable over students and subjects based on item response theory are discussed: a method based on a model with a unidimensional representation of proficiency, a method based on a model with a multidimensional representation of proficiency and a method based on a multidimensional representation of proficiency where the stochastic nature of the choice of examination subjects is explicitly modeled. The methods are compared using the data from the Central Examinations in Secondary Education in the Netherlands. The results show that the unidimensional item response model produces unrealistic results, which do not appear when using the two multidimensional item response models. Further, it is shown that both multidimensional models produce acceptable model fit. However, the model that explicitly takes the choice process into account produces

This chapter has been submitted for publication as: O.B. Korobko, C.A.W. Glas, R.J. Bosker, and H. Luyten, Comparing the difficulty of examination subjects with Item Response Theory

the best model fit.

3.1. Introduction

The problem of grade adjustment for the comparison of students and schools has a long history (see, for instance, Linn, 1966). Johnson (1997, 2003) notes that combining student grades through simple averaging schemes to obtain grade point averages (GPAs) results in systematic bias against students enrolled in more rigorous curricula. The practice has important consequences for the course selection by the students, and it may be one of the major causes of grade inflation. Caulkins, Larkey and Wei (1996) note that the use of GPA is based on the incorrect assumption that all course grades mean essentially the same thing. There is, however, substantial variation among majors, courses, and instructors in the rigor with which grades are assigned. A lower GPA may not necessarily mean that the student performs less well than students who have higher GPAs; the student may simply be taking courses and studying in fields with more stringent grading standards.

The appropriateness of GPAs is also a point of debate in school effectiveness research and in the trend towards public reporting of school results. School results are generally corrected for differences between the students at school entry (Fitz-Gibbon, 1994; Willms, 1992), but the comparability of the actual outcome measures, such as examination results, has received less attention, with the exception of Kelly (1976), Newton (1997), and Smits, Mellenbergh and Vorst (2002). In many countries (such as the Netherlands, where the data used here emanate) a student's examination result has a direct consequence for the admittance to university. Therefore, students generally choose the examination subjects in which they feel competent. The focal problem addressed by Kelly (1976), Newton (1997), and Smits, Mellenbergh and Vorst (2002) is whether the fact that students generally choose subjects that fit their proficiency distorts the comparison of average examination results between schools. Parents, local authorities and politicians, however, may interpret these differences in GPAs as absolute objectivity, ignoring the influence of the differences in the difficulty of the subjects and the students' choice behavior.

Most more recent methods for adjusting GPA are based on item response theory (IRT). The objective of these methods is to account for the relative difficulty of the courses or examinations and the differences in the proficiency levels of the students (Young, 1990, 1991; Johnson, 1997, 2003). In the present article, this approach is expanded in two directions. First, it is assumed that the courses or examinations

load on more than one dimension. (In the sequel, we will use the term examinations as a generic name that also includes assessments of courses and the like). Using a real-data example it is shown that a multidimensional representation of proficiency leads to more plausible results and better model fit. Second, it is argued that the free choice of examinations may lead to a violation of the ignorability principle (Rubin, 1976) and, as a consequence, to biased estimates of the difficulties of the examination subjects. It is shown that this bias can, to a certain extent, be accounted for by introducing a stochastic model for the choice variables.

This article is organized as follows. First, three IRT models will be described: a unidimensional and a multidimensional model for the grades only, and a multidimensional model pertaining to the grades and the choice variables simultaneously. As an example, an analysis of data collected by Dutch Inspectorate of Education will be presented. Then, a method for the evaluation of model fit will be described and the fit of the three models will be compared. Finally, the last section presents a discussion and some conclusions.

3.2. Methods

3.2.1. Grade Point Average Adjustment

One might view the problem of comparing the difficulty of examinations as an item scaling problem with incomplete data, that is, as a test equating problem (see, for instance, Kolen and Brennan, 1995), where an item score is the (discrete, polytomous) score on an examination subject. We define a choice variable as

$$d_{ni} = \begin{cases} 1 & \text{if student } n \text{ did chose examination subject } i \\ 0 & \text{if student } n \text{ did not chose examination subject } i \text{ ,} \end{cases} \tag{3.1}$$

for students $n = 1, ..., N$ and examination subjects $i = 1, ..., K$. An important aspect of the problem discussed in this article is that the design (that is, the values of the choice variables d_{ni}) is not fixed in advance, but it is student driven and therefore stochastic. The consequences of the stochastic nature of the design will be returned to below.

The objective is to compute adjusted GPAs in such a way that they are comparable. This is done by estimating the GPA for a situation where all students take all examinations. Since they do not actually take all examinations, we impute expected

grades for the missing observations, that is

$$GPA = \frac{1}{K}\sum_{i=1}^{K}(d_{ni}X_{ni} + (1 - d_{ni})E(X_{ni})), \tag{3.2}$$

where X_{ni} is the observed grade if $d_{ni} = 1$ and an arbitrary value if $d_{ni} = 0$, and $E(X_{ni})$ is the expectation under a model used to describe the students' proficiency.

3.2.2. Item Response Theory

The expectations $E(X_{ni})$ in (3.2) will be computed using IRT models for the proficiency of the students and the difficulty of the examination subjects. Three models will be discussed. In the first model, it will be assumed that the grades on all subjects have a unidimensional representation of proficiency. In the second model, this assumption is broadened to the assumption that the subjects relate to more than one proficiency dimension. The third model is motivated by the expectation that there is an interaction between the students' pattern and level of proficiency on one hand, and the choice of examination subjects on the other hand. Therefore, the third model has a multidimensional representation of proficiency where the choice-variables are explicitly modelled.

Model 1

Model 1 is the unidimensional version of the generalized partial credit model (Muraki, 1997). The probability that the grade X_{ni} is in category j ($j = 0, ..., m$) is given by

$$p\,(X_{ni} = j|d_{ni} = 1; \theta_n) = \frac{\exp\left(j\alpha_i\theta_n - \sum_{h=1}^{j}\beta_{ih}\right)}{1 + \sum_{h=1}^{m}\exp\left(h\alpha_i\theta_n - \sum_{k=1}^{h}\beta_{ik}\right)}, \tag{3.3}$$

where θ_n is the unidimensional proficiency parameter that represents the overall proficiency of student n. So it is assumed here that all examination grades relate to one unidimensional proficiency parameter θ_n. The parameters β_{ij} ($j = 1, ..., m_i$) are the locations on the latent scale where the probabilities of scoring in category $j-1$ and j are equal. These parameters model the difficulty of examination subject i. ($\beta_{i0} = 0$ to identify the model). Parameter α_i defines the extent to which the response is related to the proficiency θ_n.

The parameters of the model can be estimated using maximum marginal likelihood (MML, see Bock & Aitkin, 1981). In MML, the model is enhanced with the

assumption that the proficiency parameters are drawn from one normal distribution or from more than one normal distribution (the latter is known as multiple-group IRT, see Bock and Zimowski, 1997). In the example presented below, it cannot be a priori assumed that the average level of proficiency is independent of the chosen examination package. Therefore, it will be assumed that students choosing the same examination package (that is, students with the same pattern on the choice variables $d_{n1}, ..., d_{ni}, ..., d_{nK}$) are drawn from a normal distribution with a mean μ_p (where p is the index of the package) and a variance σ^2.

In MML, a likelihood function is maximized where the students' proficiency parameters are integrated out of the likelihood. The marginal log-likelihood for Model 1 is given by

$$L_1 = \sum_p \sum_{n|p} \log \int \prod_i p(x_{ni}|d_{ni};\theta) g(\theta;\mu_p,\sigma^2) d\theta, \tag{3.4}$$

where x_{ni} is the observed grade, $p(x_{ni}|d_{ni};\theta)$ is equal to (3.3) evaluated at x_{ni} if $d_{ni} = 1$, and $p(x_{ni}|d_{ni};\theta) = 1$ if $d_{ni} = 0$. Further, $g(\theta;\mu_p,\sigma^2)$ is the normal density with parameters μ_p and σ^2. The model can be identified by choosing $\mu_1 = 0$ and $\sigma^2 = 1$.

The estimates can be computed using the software packages Multilog (Thissen, Chen & Bock, 2002) or Parscale (Muraki & Bock, 2002). These packages compute concurrent MML estimates of all the structural parameters in the model (the β-parameters and the means μ_p), and this is the approach that is also pursued in the present article.

After the parameters of the examinations are estimated by MML, the missing examination scores can be estimated by their posterior expectations, that is, by

$$E\,(X_{ni} \mid \mathbf{x}_n) = \sum_{j=1}^{m} j \int p(X_{ni} = j|d_{ni} = 1;\theta)\; p(\theta|\,\mathbf{x}_n) d\theta, \tag{3.5}$$

where $p(\theta|\mathbf{x}_n)$ is the distribution of θ given the observations $\mathbf{x}_n$, and $p(X_{ni} = j|d_{ni} = 1;\theta)$ is defined by (3.3). These expected scores are then imputed in (3.2).

Model 2

In the previous model it was assumed that the grade of student n depended on a unidimensional proficiency parameter θ_n. However, there may be more than one proficiency factor underlying the grades. For instance, there might be a specific proficiency factor for the science subjects and another one for language subjects. If

Q proficiency dimensions are needed to model the grades, the proficiency can be represented by a vector of proficiency parameters $\left(\theta_{n1}, ..., \theta_{nq}, ..., \theta_{nQ}\right)$. The probability of a grade in category j is now given by

$$p\left(X_{ni} = j | d_{ni} = 1; \theta_n\right) = \frac{\exp\left(j\left(\sum_{q=1}^{Q} \alpha_{iq}\theta_{nq}\right) - \sum_{h=1}^{j} \beta_{ih}\right)}{1 + \sum_{h=1}^{m} \exp\left(h\left(\sum_{q=1}^{Q} \alpha_{iq}\theta_{nq}\right) - \sum_{k=1}^{h} \beta_{ik}\right)}. \tag{3.6}$$

In addition, it will be assumed that the proficiency parameters θ_n, $\theta_n = (\theta_{n1}, ..., \theta_{nq}, ..., \theta_{nQ})$ of groups of students taking a specific package of examination subjects have a Q- variate normal distribution with a mean μ_p and a covariance matrix $\mathbf{\Sigma}$. So it is assumed that the mean depends on the examination package, and that the covariance matrix of the proficiency parameters is common for all students. Takane and de Leeuw (1987) show that the model is equivalent with a full-information factor analysis model. Therefore, the parameters $\alpha_{i1}, ..., \alpha_{iQ}$ are often called factor-loadings, and the proficiency parameters $\theta_{n1}, ..., \theta_{nq}, ..., \theta_{nQ}$ can be viewed as factor scores. Note that the factor loadings are specific for an examination subject and that they model the relation between the probability of obtaining a grade and the level on the Q proficiency dimensions. A high positive value of α_{iq} means that the q-th dimension is important for the subject, a value close to zero means that the dimension does not play an important role.

The model is identified by setting $\mu_1 = 0$ and setting the diagonal of $\mathbf{\Sigma}$ equal to one. For a discussion of these and alternative identification restrictions refer to Béguin and Glas (2001). The marginal log-likelihood of the model becomes

$$L_2 = \sum_{p} \sum_{n|p} \log \int \prod_{i} p(x_{ni}|d_{ni}; \theta) g(\theta; \mu_p, \Sigma) d\theta, \tag{3.7}$$

where x_{ni} is the observed grade, $p(x_{ni}|d_{ni}; \theta)$ is equal to (3.6) evaluated at x_{ni} if $d_{ni} = 1$, and $p(x_{ni}|d_{ni}; \theta) = 1$ if $d_{ni} = 0$, and $g(\theta; \mu_p, \Sigma)$ is the Q-variate normal density. The parameters of the multidimensional model can be estimated using MML (see Bock, Gibbons & Muraki, 1988) and the computer packages TESTFACT (Wood et al., 2002), ConQuest (Wu, Adams & Wilson, 1997) or Mplus (Muthén & Muthén, 2003) can be used to compute the estimates.

Using the MML estimates of the parameters of the examinations, the missing examination scores can be estimated by their posterior expectations analogously to

(3.5), but the expectations are now with respect to a Q-variate posterior distribution $p(\theta|\mathbf{x}_n)$, that is, by

$$E\,(X_{ni} \mid \mathbf{x}_n) = \sum_{j=1}^{m} j \int .. \int p(X_{ni} = j|d_{ni} = 1;\theta)\; p(\theta|\,\mathbf{x}_n)d\theta. \tag{3.8}$$

Model 3

In Model 2 there is no interaction between the choice of an examination subject and the proficiency parameters. That is, it is assumed that the process causing the missing data does not need to be considered in the estimation process. Rubin (1976) identified two conditions under which the missing data process can be ignored. A missing data mechanism is ignorable if the missing values are missing at random (MAR) and if the parameters of the distribution of the observed data (say λ) and the distribution of the missing data (say φ) are distinct. MAR holds if the probability of the missing data pattern $p(d\,|x_{mis}, x_{obs}, \varphi)$ does not depend on missing data, that is, if $p(d\,|x_{mis}, x_{obs}, \varphi) = p(d\,|x_{obs}, \varphi)$. Distinctness entails that there are no functional dependencies between φ and the parameters of interest λ, or that φ and λ have independent priors. If ignorability does not hold, the inferences made using an IRT model ignoring the missing data process can be severely biased (Bradlow & Thomas, 1998; Holman & Glas, 2005).

A general method to deal with non-ignorable missing data proposed by Heckman (1979) is the introduction of a selection model for the observations. Several authors have applied this approach in the framework of IRT (Moustaki & O'Muircheartaigh, 2000; Moustaki & Knott, 2000; Holman & Glas, 2005) and have shown that selection bias can be removed when the distribution of d_{ni} is modelled concurrently with the observed data using an IRT model. Their approach is adapted to the present problem as follows. As in Model 2, the scores on the examination subjects are modelled by (3.7). Further, it is assumed there exists a latent variable θ_{Q+1}, that governs the choice of the examination subjects, that is, the realizations of the choice variable defined by (3.1). If the students' proficiency level is highly correlated with the choice of examination subjects, then θ_{Q+1} will be highly correlated with $\theta_1, ..., \theta_Q$ also. The dependence between the latent variables is modelled by assuming that $\theta_1, ..., \theta_{Q+1}$ have a multivariate normal distribution, again with a specific mean for every group of students and a common covariance matrix. The marginal likelihood of the model is

$$L_3 = \sum_{p} \sum_{n|p} \log \int \prod_{i} \left[p(x_{ni}|d_{ni};\theta^{(Q)})p(d_{ni};\theta_{Q+1})\right] g(\theta;\mu_p,\Sigma)d\theta, \tag{3.9}$$

where $\theta^{(Q)} = (\theta_1, ..., \theta_Q)$, and $\theta = (\theta_1, ..., \theta_{Q+1})$.

The correlation between θ_{Q+1} and the proficiency dimensions $\theta_1, ..., \theta_Q$ describe the extent to which the choice of an examination subject depends on the proficiency level. So if, for example, the correlation between θ_1 and θ_{Q+1} is positive, a high level on proficiency dimension θ_1 is positively related with endorsing subjects that load high on dimension θ_1. Further, the magnitude of the correlations between $\theta_1, ..., \theta_Q$ and θ_{Q+1} give an indication of the extent to which ignorability is violated. If these correlations are close to zero, the choice behavior is not related to proficiency, and the missing data are ignorable. If, on the other hand, these correlations are substantial the choice variable is highly related to the proficiency for the students. Holman and Glas (2005) show that the bias in the parameter estimates is positively related to the correlation between the latent proficiency and the parameters of the IRT model for the missing data indicator d_{ni} and that this bias vanishes when the observations and the realizations of the missing data indicators are concurrently modelled by the multidimensional IRT model described here.

The final consideration is about the model for the choice variables d_{ni}. Since the students can only chose a limited number of subjects, it is reasonable to assume that the probability of choosing a subject as a function of the proficiency dimension θ_{Q+1} is single peaked: Students will probably chose subjects within a certain region of the proficiency dimension θ_{Q+1} and avoid subjects that are too difficult or too easy. An IRT choice model that may reflect this, is given by

$$p\,(d_{ni} = 1) = \pi_{i1}(\theta_{(Q+1)n}) - \pi_{i2}(\theta_{(Q+1)n}) \tag{3.10}$$

where

$$\pi_{ij}(\theta_{(Q+1)n}) = \frac{\exp\left(\theta_{(Q+1)n} - \gamma_{ij}\right)}{1 + \exp\left(\theta_{(Q+1)n} - \gamma_{ij}\right)}, \tag{3.11}$$

and $\gamma_{i1} < \gamma_{i2}$ to guarantee that $\Pr\,(d_{ni} = 1)$ is positive. The model considered here is closely related to models by Verhelst and Verstralen (1993) and Andrich and Luo (1993, also see Andrich, 1997). These two models share with the model given by (3.10) the property of a single-peaked response probability, only the functional form of the probability is chosen differently. The model by Verhelst and Verstralen (1993) is derived from the partial credit model; the model by Andrich and Luo (1993) has a hyperbolic cosine function probability function. The motivation for the present model is its simple functional form.

The model given by (3.10) is also related to a special case of the graded response model by Samejima (1969, 1993). The graded response model pertains to a polyto-

mously scored response variable, for instance, a response variable y_{ni} that assumes the values 0, 1 or 2. The response probabilities are given by

$$\begin{aligned} \Pr(y_{ni} = 0) &= 1 - \pi_{i1}(\theta), \\ \Pr(y_{ni} = 1) &= \pi_{i1}(\theta) - \pi_{i2}(\theta), \\ \Pr(y_{ni} = 2) &= \pi_{i2}(\theta), \end{aligned}$$

with $\pi_{ij}(\theta)$ as defined by (3.11). So model (3.11) can be derived from the graded response model by noting that the responses $y_{ni} = 0$ and $y_{ni} = 2$ are extreme cases and collapsed to $d_{ni} = 0$.

Estimation procedures for a model that is a mixture of the logistic IRT model defined by (3.6), the collapsed graded response model defined by (3.10), and a $(Q+1)$-variate normal model for the proficiency parameters are not readily available. In Appendix A, the marginal maximum likelihood (MML) procedure used to calculate the estimates reported below is outlined. Estimation of the missing examination scores is analogously to their estimation in Model 1 and Model 2, except that the expectations are now with respect to a $Q + 1$-variate posterior distribution $p(\theta|\mathbf{x}_n)$.

3.2.3. Model Fit

Likelihood ratio testing is the standard methodology for model comparison, and this methodology will also be applied below. However, these tests are rather global and give, for instance, no information with respect to the fit of specific examination subjects. In principle, IRT models can be evaluated by Pearson-type statistics, that is, statistics based on the difference between observations and their expectations under the null-model. Such statistics are available for unidimensional models for dichotomous observations (Orlando & Thissen, 2000; Glas & Suarez-Falcon, 2003), for unidimensional models for polytomous observations (Glas, 1998, 1999), and for multidimensional models for such observations (te Marvelde, Glas, Van Landeghem, & Van Damme, 2006). In the present article, a comparable fit statistic will be presented that is targeted at the special application considered here.

Most item fit statistics are based on splitting up the sample of respondents into subgroups with different proficiency distributions and evaluating whether the item response frequencies in these subgroups differ from their expected values. Orlando and Thissen (2000) point out that the splitting criteria should be directly observable

(for instance, number correct scores) rather than estimated (for instance, estimated proficiency). Following this suggestion, we split up the sample of students using a splitter examination labelled s. Two subgroups are formed, one subgroup of students that did choose subject s (so $d_{ns} = 1$) and subgroup of students that did not choose subject s (so $d_{ns} = 0$). The test is based on the assumption that this criterion splits the sample up in two subgroups with different proficiency distributions. We compute the average grade on number of students with a grade j on examination i in both subgroups as

$$S_{i0} = \left[\sum_n (1 - d_{ns}) d_{ni} x_{ni}\right] / \left[\sum_n (1 - d_{ns}) d_{ni} m_i\right]$$

and

$$S_{i1} = \left[\sum_n d_{ns} d_{ni} x_{ni}\right] / \left[\sum_n d_{ns} d_{ni} m_i\right].$$

where m_i is the maximum grade on examination i, so $m_i = 3$ for all i. These average grades can be compared to their expected values given by

$$E_{i0} = \left[\sum_n (1 - d_{ns}) d_{ni} E\left(X_{ni} \mid \mathbf{x}_n\right)\right] / \left[\sum_n (1 - d_{ns}) d_{ni} m_i\right]$$

and

$$E_{i1} = \left[\sum_n d_{ns} d_{ni} E\left(X_{ni} \mid \mathbf{x}_n\right)\right] / \left[\sum_n d_{ns} d_{ni} m_i\right],$$

where $E\left(X_{ni} \mid \mathbf{x}_n\right)$ is given by (3.8).

In Appendix it is shown that a Pearson-type fit-statistic bases on the squared differences between observed and expected values can be used to evaluate whether the observed and expected response frequencies are acceptably close given (the observed value on the choice variable of the splitter examination). In Appendix it is also outlined that the statistic has an asymptotic χ^2 distribution with one degree of freedom. However, the application presented below has a very large sample size and the power of the test becomes very large. Therefore, the test will be used to compare the relative model fit of nested models. More specifically, the test will be used to evaluate whether Model 3 fits the data systematically better than Model 2.

3.3. An Example

3.3.1. The Data

The data used to illustrate the advantage of IRT in the present setting are from approximately 18-year old students of pre-university schools in the Netherlands. This is the only curriculum track (of the four available) that gives students the opportunity for direct entry into a university. The external examinations are standardized nationwide achievement tests. The students take examinations in 7 or 8 subjects chosen from the list of subjects displayed in Table 3.1. The data used in this study were collected by the Dutch Inspection of Education. For this study only the results from the first session of the examinations were used (unsatisfactory marks might be "repaired" in a re-session). The data are a subset of the data of pre-university students that took their final examination in the school year 1994/1995. The original data set comprised 16,118 students. To keep the presentation of the results relatively simple, the analysis was restricted to 60 fairly common combinations of examination subjects. The resulting data set consisted of the examination results of 6142 students. The distribution of the students over examination subjects in the original data and the selected data are shown in Table 3.1.

Below, the appropriateness of the methods for computing adjusted GPAs will be assessed by evaluating the consequences of the method in subgroups. From the combinations of different subjects chosen by the students, we distinguish three main groups:

1. The linguistically-oriented students (20%). These students definitely take examinations in French and German language, and not more than one of the subjects like Applied Mathematics, Advanced Mathematics, Physics and Chemistry.

2. The science-oriented students (33%). These students definitely take examinations in at least three of the subjects like Applied Mathematics, Advanced Mathematics, Physics and Chemistry and no examinations in French or German languages.

3. All other students (47%).

The original grades ranged from 1 ("poor") to 10 ("excellent"), but for the purpose of our study these were re-scaled to a four point scale, where the points are 0 (original grade 0 to 5.4, which is unsatisfactory), 1 (original grade 5.5 to 6.4, which is just satisfactory), 2 (original grade 6.5 to 7.4, which is good), and 3 (original grade 7.5 to 10, which is very good). The overall observed mean examination scores and the mean examination scores observed in the three subgroups are displayed in Table 3.2.

The following observations are of interest. Note that students with a science-

Table 3.1: Distribution of students over examination subjects in original data set ($N = 16,118$) and analysis data set ($N = 6,142$)

Subjects Selected	Percentage Original Data	Percentage Selected Data	Subjects Not Selected	Percentage Original Data
Dutch language	99.9	100.0	Frisian language	0.0
Latin	14.6	10.3	Russian	0.0
Classical Greek	6.2	4.1	Spanish	0.2
French	37.6	36.6	Handicrafts	1.9
German	45.4	44.5	Music	1.6
English	99.1	100.0	Philosophy	0.7
History	49.5	48.8	Social studies	2.3
Geography	33.9	31.3		
Applied Math	63.0	65.1		
Advanced Math	44.7	40.2		
Physics	46.7	42.5		
Chemistry	38.2	39.9		
Biology	37.0	33.3		
General Economy	58.7	66.6		
Business Economy	36.0	37.9		
Arts	7.8	5.5		

oriented package score lower on Dutch and English language than the students with a language-oriented package. On the other hand, students with a science-oriented package score substantially higher on Applied Mathematics than students with a language-oriented package. This is a first indication that the proficiency dimension might not be unidimensional. Further, the score of French and German may be boosted relative to the score on Dutch and English language by the absence of the students with a science oriented package, who seem to have a lower language proficiency than the other students. The IRT analyses presented below will clarify these observations.

3.3.2. Results

Model 1

Model 1 was estimated by MML, that is, by maximizing (3.4). The estimates of the parameters α_i and β_{ij} ($j = 1, ..., m_i$) are given in Table 3.3. The last column of the table gives the average of the estimates of the parameters β_{ij} ($j = 1, ..., m_i$), denoted by $\overline{\beta}$. This average is an estimate of the global position of subject j on the latent scale,

Table 3.2: Observed examination scores per subject and per package

Subjects	Overall	Science	Language	Mixed
Dutch Language	1.38	1.29	1.53	1.37
Latin	2.47	2.36	2.44	2.70
Classical Greek	2.18	2.07	2.18	2.32
French	1.63	—	1.68	1.57
German	1.50	—	1.51	1.50
English	1.50	1.38	1.64	1.51
History	1.58	1.87	1.52	1.55
Geography	1.31	1.88	1.11	1.34
Applied Math	1.15	2.28	0.71	0.91
Advanced Math	1.37	1.37	—	1.36
Physics	1.50	1.47	—	1.59
Chemistry	1.76	1.76	—	1.75
Biology	1.75	1.71	—	1.91
General Economy	1.27	1.73	0.92	1.25
Business Economy	1.41	1.84	1.39	1.30
Arts	1.60	1.75	1.56	1.57

and serves as an indication of the average difficulty of the subject.

Note that Dutch language and Art are the least discriminating with respect to the overall proficiency and Physics and Chemistry have the highest discrimination. Inspection of the values of $\overline{\beta}$ in the last column shows that Advanced Mathematics is now slightly more difficult than Applied Mathematics. This result is contrary to the result in Table 3.2, where the overall average of Advanced Mathematics is higher than the overall average of Applied Mathematics (1.37 versus 1.15). This phenomenon is of course explained by the fact that the students with a language-oriented package do not take Advanced Mathematics.

Using the MML estimates, posterior expectations as defined by (3.5) were imputed for the missing examination scores. The results are given in Table 3.4. The average scores in the table can be interpreted as the average scores obtained if all students endorsed all subjects. The most dramatic effect is the decrease of the average scores for the classical languages Latin and Greek. The explanation may be that the small percentage of the students that actually choose these subjects (10.3% and 4.1%) are highly proficient. Adding imputed values for the other students (of lower proficiency) can only lower this average. This explanation is in line with the experience in Dutch education.

Table 3.3: Parameter estimates for Model 1

	N	α	β_1	β_2	β_3	$\bar{\beta}$
Dutch language	6142	0.38	-0.97	0.10	1.38	0.17
Latin	637	0.65	-1.95	-1.18	-0.17	-1.10
Classical Greek	256	0.68	-1.36	-0.65	0.57	-0.48
French	2250	0.91	-1.24	-0.28	0.71	-0.27
German	2739	0.99	-1.22	-0.05	0.86	-0.14
English	6142	0.63	-0.79	-0.03	0.79	-0.01
History	2997	0.83	-1.31	-0.24	0.92	-0.21
Geography	1928	0.71	-1.22	0.14	1.41	0.11
Applied Math	4002	0.59	-0.42	0.33	1.01	0.31
Advanced Math	2471	0.91	-0.51	0.44	1.23	0.39
Physics	2614	1.38	-0.90	0.29	1.51	0.30
Chemistry	2452	1.37	-1.14	-0.15	0.97	-0.11
Biology	2048	1.03	-1.77	-0.25	1.35	-0.22
General Economy	4092	0.87	-0.82	0.23	1.30	0.24
Business Economy	2330	0.90	-1.10	-0.01	1.21	0.03
Arts	338	0.37	-1.60	-0.22	1.06	-0.26

An unexpected result were the imputed means for French and German language for the students with a science-oriented package. In Table 3.3 it can be verified that these students did not choose these two languages in their examination package. In Table 3.4 it can be verified that these students score relatively low on Dutch and English language, (1.29 and 1.38, respectively) yet the imputed means on French and German language are quite close to the mean scores for the students with a language-oriented package. The opposite phenomenon occurred with the imputed values for Advanced Mathematics, Physics, Chemistry, and Biology for the students with a language-oriented package. The imputed means were all close to the means for the other students, yet their (generally observed) score on Applied Mathematics was as low as 0.87. This is highly unexpected. In the sequel, it will become clear that this phenomenon is attributable to the multidimensionality of the proficiency variables.

Model 2

A three-dimensional version of Model 2 was fitted with a method developed by Béguin and Glas (2001). The method identifies the dimensions by fitting unidimensional IRT models by discarding items, or, in the present case, examination subjects. These examination subjects are entered as unique indicators of a dimension

Table 3.4: Examination scores per subject and per package estimated under Model 1

	Science	Language	Mixed
Dutch Language	1.29	1.53	1.37
Latin	1.90	1.86	1.83
Classical Greek	1.61	1.58	1.57
French	1.58	1.68	1.56
German	1.55	1.51	1.51
English	1.38	1.64	1.51
History	1.63	1.53	1.55
Geography	1.53	1.31	1.42
Applied Math	1.76	0.87	1.01
Advanced Math	1.37	1.37	1.38
Physics	1.47	1.41	1.46
Chemistry	1.74	1.46	1.54
Biology	1.67	1.48	1.56
General Economy	1.57	1.04	1.28
Business Economy	1.58	1.44	1.38
Arts	1.69	1.61	1.63

in the multidimensional IRT model, that is, these examination subjects load on one dimension only. Examination subjects that do not fit one dimension uniquely are allowed to load on all dimensions. In the present application, the unidimensional subscales were searched for with the program OPLM (Verhelst, Glas & Verstralen, 1995). The R_{1c} statistic (Glas, 1988) was used as a criterion for model fit. Then, given the factor loadings fixed to zero in the previous stage, an MML estimate was made of the subject parameters and the correlation matrix.

The results are shown in Table 3.5 under the heading “Factor Solution Model 2”. The factor loadings fixed to zero are marked by an asterisk. The three dimensions that appeared can be interpreted as “Language”, “Science”, and “Economy”. Arguments for this interpretation are the fact that “Dutch” loads high on the first, and very low on the second dimension, while “Advanced Mathematics” loads mildly negative on the first, and high on the second dimension. Examination subjects that do not load according to expectation are Latin and Classical Greek (that both load high on the Language and Economy dimension). Note that History loads on all three dimensions, Arts loads low on the Language dimension, and Geography has a high loading on the Science dimension. The correlation matrix of the three latent dimensions is shown at

the bottom of the table. Note that the correlation between the Science and Economy dimension is substantially higher than the other two correlations.

Next, using the MML estimates of the parameters of Model 2, the missing scores could be estimated by their posterior expected values. The results are given in Table 3.6 under the heading Model 2. Again, the average scores in the table can be interpreted as the average scores obtained if all students endorsed all subjects. An important implausible finding using Model 1 was that the expected grades for the language-oriented group on Advanced Mathematics, Science, Chemistry and Biology were higher than the grades of the science-oriented group. Inspection of the analogous estimated averages computed using Model 2 displayed in Table 3.6 6 show that these estimates do not suffer from this phenomenon. Also the estimates for French and German language for the science oriented group are now lower than the analogous estimates in Table 3.4.

Model 3

Above, it was argued that the free choice of examination subjects lead to a stochastic design that might violate the assumption of ignorability. Therefore, Model 3 was derived from Model 2 by adding a special dimension to model the missing data indicators d_{ni} as defined in (3.1). The MML estimates of the parameters of Model 3 were obtained by maximization of (3.9); the results are shown in Table 3.5 under the heading "Factor Solution Model 3". Note that the patterns of the factor loadings and the correlation matrices for the first three dimensions for Model 2 and Model 3 are analogous. Since the fourth dimension, that is, the latent dimension describing the choice process is modelled by the model given by Formula (3.11), displaying the factor loadings is little informative, since they are all equal to one. Therefore, the average of the two subject parameters, that is, $\overline{\gamma}_i = (\gamma_{i1} + \gamma_{i2})/2$ are displayed for all subjects in the last column labelled "Choice". The parameters $\overline{\gamma}_i$ can be seen as an estimate of the location of the subject on this fourth proficiency dimension. Note that the parameters for Dutch and English cannot be estimated, because these two examination subjects are obligatory and so all the choice variables d_{ni} for these examination subjects are structurally equal to one and the parameters γ_{ij} related to these subjects cannot be estimated.

The interpretation of the mean parameters $\overline{\gamma}_i$ is as follows. The fourth dimension correlates positively with the three proficiency dimensions, and highest with the Science dimension. This dimension can be viewed as an overall proficiency dimension, and the choice of subjects is assumed governed by proficiency. Since the

Table 3.5: Factor Loading per Subject for the Three- and Four-Factor Solution and Correlation Matrices

	Factor Solution Model 2			Factor Solution Model 3			
	L*	S	E	L	S	E	Choice
Dutch Language	2.22	-0.05	0.45	1.91	-0.09	0.44	—
English	6.97	0.00*	0.00*	5.46	0.00*	0.00*	—
Latin	3.14	-0.22	1.89	2.88	-0.32	1.67	-0.76
Classical Greek	2.31	0.03	2.75	2.32	-0.20	1.46	-1.12
French	7.26	0.00*	0.00*	6.11	0.00*	0.00*	-0.89
German	9.27	0.00*	0.00*	7.79	0.00*	0.00*	-0.62
History	2.04	1.31	2.06	1.86	-0.23	2.18	-0.19
Geography	0.00*	6.12	0.00*	0.00*	3.23	0.00*	0.26
Applied Math	0.00*	3.63	0.00*	0.00*	4.69	0.00*	0.01
Advanced Math	-0.76	5.84	0.09	-0.64	4.25	0.12	0.43
Physics	0.00*	9.03	0.00*	0.00*	6.01	0.00*	0.76
Chemistry	0.00*	8.86	0.00*	0.00*	6.57	0.00*	0.89
Biology	0.00*	6.62	0.00*	0.00*	5.09	0.00*	1.24
General Economy	0.00*	0.00*	7.78	0.00*	0.00*	3.42	-0.31
Business Economy	0.00*	0.00*	6.99	0.00*	0.00*	4.26	-0.13
Arts	1.23	0.03	1.06	1.15	0.08	0.49	0.56
Correlation matrix							
Language	1.00			1.00			
Science	0.51	1.00		0.43	1.00		
Economy	0.45	0.81	1.00	0.48	0.84	1.00	
Choice dimension				0.12	0.74	0.56	1.00
Fixed factor loadings							
* L, S, and E denote Language, Science, and Economy respectively.							

'difficulty parameters' $\overline{\gamma}_i$ are an estimate of the location of the subjects on the fourth proficiency dimension, they represent the ordering of the examination subjects on this dimension. That is, "difficult subjects" as Biology ($\overline{\gamma}_i = 1.24$), Chemistry ($\overline{\gamma}_i = 0.89$) and Advanced Mathematics ($\overline{\gamma}_i = 0.43$) are endorsed by the more proficient students. Note that also Arts ($\overline{\gamma}_i = 0.56$) scores high on this dimension.

As for the other two models, also for Model 3 the missing scores were estimated by their posterior expectations. The averages computed assuming all students endorsed all subjects are given in Table 6 under the heading Model 3. Also the estimates under Model 3 do not show the implausible results obtained under Model 1. In Table 3.6, it can also be verified that the estimates for Model 2 and Model 3 did

Table 3.6: Examination scores per subject and per package estimated under Model 2 and Model 3

	Model 2			Model 3		
	Science	Language	Mixed	Science	Language	Mixed
Dutch Language	1.29	1.53	1.37	1.29	1.53	1.37
Latin	1.80	1.78	1.78	1.85	1.79	1.83
Classical Greek	1.54	1.48	1.56	1.49	1.62	1.64
French	1.45	1.68	1.54	1.44	1.69	1.58
German	1.41	1.51	1.49	1.35	1.50	1.51
English	1.38	1.64	1.51	1.38	1.64	1.51
History	1.66	1.52	1.59	1.72	1.62	1.69
Geography	1.88	1.26	1.50	1.85	1.11	1.41
Applied Math	1.97	0.88	1.09	1.86	0.82	1.03
Advanced Math	1.36	0.81	1.03	1.36	0.81	1.17
Physics	1.47	0.92	1.15	1.46	0.75	1.04
Chemistry	1.75	1.00	1.25	1.75	0.85	1.17
Biology	1.69	1.06	1.31	1.72	0.97	1.28
General Economy	1.50	1.03	1.31	1.45	1.02	1.24
Business Economy	1.49	1.29	1.37	1.43	1.22	1.30
Arts	1.64	1.59	1.65	1.57	1.87	1.85

not substantially differ.

3.3.3. Model Fit

First, the fit of the three models was compared using likelihood ratio tests. A test of Model 1 against Model 2 yielded an chi-square value of 2070.1 with 135 degrees of freedom. So Model 1 had to be rejected. To facilitate the test of Model 3 against Model 2, both models have to refer to the same data. For Model 3, these data comprise of the subject scores and the choice variables. Therefore, Model 2 was enhanced with an independent the choice model by the from of (3.11) for the variables d_{ni}. Then the likelihood was computed as the product of the likelihood under Model 2 multiplied by the likelihood of the choice model for the variables d_{ni}. The test of this enhanced model against Model 3 is equivalent with testing whether the covariances between the latent variables associated with the observations and the latent variables associated with d_{ni} are zero. The test statistic has a value of 312.2, with 3 degrees of freedom. The conclusion is that Model 3 fitted significantly better than Model 2. However, as noted above, the impact of this better model fit was quite small.

Table 3.7: Model fit evaluated using T_i-statistic

Splitter	Advanced Math		History		Business Economy	
	Model 2	Model 3	Model 2	Model 3	Model 2	Model 3
Dutch Language	84.0	31.0*	58.9	9.3*	106.8	38.3*
Latin	9.0	1.3*	12.0	6.4*	13.3	4.4*
Classical Greek	0.7	1.2	1.0	0.2*	1.0	1.1
French	2.1	0.0*	25.2	15.4*	41.3	34.2*
German	0.1	1.5	52.9	47.7*	63.0	50.2*
English	48.9	34.9*	69.4	48.8*	125.1	98.8*
History	1.5	1.4*			79.8	56.4*
Geography	4.4	2.5*	176.5	113.0*	89.4	33.0*
Applied Math	25.7	70.6	17.8	20.2	90.0	35.2*
Advanced Math			22.7	6.9*	17.4	21.1
Physics	19.7	4.2*	10.5	2.4*	23.7	12.2*
Chemistry	67.3	22.8*	22.6	6.1*	0.9	0.4*
Biology	7.4	5.3*	125.3	117.8*	7.3	4.9*
General Economy	9.4	4.5*	91.4	49.0*		
Business Economy	1.8	1.2*	4.0	2.6	2.6	5.5
Arts	11.5	9.4*	13.4	13.6	15.0	9.0*

* indicates better fit for Model 3

Likelihood ratio tests are global tests that give an impression of overall model fit. They do no provide information on the fit of the individual examination subjects. Therefore, the T_i-statistic as defined in (3.15) was computed for all examinations $i = 1, ..., 16$, both under Model 2 and Model 3. Three splitter-examinations were used: Advanced Mathematics, History and Business Economy. The results are displayed in Table 3.7. Above it was argued that the absolute values of the test statistics were less interesting due to the large sample sizes. The statistics have an asymptotic χ^2-distribution with one degree of freedom, and in Table 3.7 it can be seen that most are significant (the 5% critical value is 3.84). More informative for the comparison of the two models is their difference in model fit. In Table 3.7, all instances where Model 3 fitted better than Model 2 are marked with an asterisk. In 36 of the 45 cases, Model 3 fitted better than Model 2. So also here the overall conclusion is that Model 3 showed the best fit.

3.4. Discussion and Conclusion

The problem addressed concerns the comparison of examination grades in case students have chosen different subjects. The complicating factor is that students only sit examinations in those subjects they have chosen themselves. As a consequence more proficient students may choose examinations in subjects that are more or the less bright students may choose less difficult subjects. However, it is not a-priori plausible that the proficiency structure assessed is unidimensional. This was corroborated by the implausible result that the language oriented students had better expected grades in Advanced Mathematics, Physics and Chemistry than the science oriented students when a unidimensional model was used to compute overall scores. Therefore, a multidimensional IRT model for polytomous items, the generalized partial credit model, was fitted to the data. The three-dimensional model had a substantially better fit than the unidimensional model. Furthermore, the implausible result of the high expected grades in Mathematics and Science for the language oriented students now vanished.

Another problem addressed related to the fact that it is not a-priori plausible that the missing data (the grades for the examination subjects that were not chosen) are missing at random. In other words, it was expected that the missingness indicators correlated with the proficiency level in such a way that this might bias the estimates of the difficulty of the examination subjects. It was attempted to remove this bias by using a four-dimensional IRT model, where the first three dimensions are related to the observed grades, while the fourth dimension is related to the observed choice of students. Though this model fitted the data significantly better than the three-dimensional model, the expected grades computed using the two models were very close.

Appendix

3.A. MML Estimates for the Choice Model and an LM Test for Model Fit

In this appendix it will be shown how the choice model defined by (3.10) and (3.11) can be incorporated in the existing MML estimation procedures for multidimensional IRT models as developed by Bock, Gibbons and Muraki (1988). The purpose is to concurrently compute estimates of the item parameters and the mean and covariance matrix of the distribution of the person parameters θ, by maximizing a likelihood function that is marginalized with respect to these person parameters θ. So let ξ be the vector of the estimands, that is $\xi = (\alpha, \beta, \gamma, \mu_\theta, \mathbf{vec}(\Sigma_\theta))$, where α and β are vectors of the item parameters of the response model given by (3.6), γ is a vector of the item parameters of the choice model given by (3.10) and (3.11), and μ_θ and $\mathbf{vec}(\Sigma_\theta)$ are a vector of the mean and a vector of the variances and covariances of θ. Usually, the model is identified by setting μ_θ equal to zero, and fixing a number of elements in α (for details refer to Holman and Glas, 2005), so these parameters are not estimated but fixed.

The marginal log-likelihood function is given by

$$\log L_{\mathbf{Y}}(\xi) = \sum_n \log p(\mathbf{y}_n; \xi)$$

where $\mathbf{y}_n$ is a vector with elements y_{ni} and y_{ni} can either be the grade x_{ni} or the choice d_{ni} of the student n. The probability of $\mathbf{y}_n$ is given by

$$p(\mathbf{y}_n; \xi) = \int \cdots \int \prod_k p(y_{nk}|\theta_n)\, g(\theta_n|\Sigma_\theta)\, d\theta,$$

where $g(\theta_n|\Sigma_\theta)$ is the density of θ_n which assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance Σ_θ. The maximum of the log-likelihood function can be found by solving $\partial \log L(\xi, \mathbf{Y}) / \partial \xi = \mathbf{0}$.

The first order derivatives of the log-likelihood function for MML estimation can be found using Fisher's identity (Louis, 1982). In the IRT framework, Fisher's identity is given by

$$\mathbf{h}(\xi) = \frac{\partial}{\partial \xi} \log L_{\mathbf{Y}}(\xi) = \sum_n E(b_n(\xi)|\mathbf{y}_n, \xi), \tag{3.12}$$

where the expectation is with respect to the posterior expectation $p(\theta_n|\mathbf{y}_n, \xi)$, and

$$b_n(\xi) = \frac{\partial}{\partial \xi} \log p(\mathbf{y}_n, \theta_n; \xi) = \frac{\partial}{\partial \xi} [\log p(\mathbf{y}_n|\theta_n, \alpha, \beta, \gamma) + \log g(\theta_n; \mu_\theta, \Sigma_\theta)]$$

(see, for instance, Glas, 1992). Notice that the derivative is a sum of the logarithm of the probability the response pattern and the logarithm of the density of the student proficiency parameter. The power of Fisher's identity is that the derivatives are simply to derive, while the derivation of $\mathbf{h}(\xi)$ is a cumbersome enterprize. For instance, it is well know that the maximum likelihood estimate of a covariance matrix is obtained as

$$\Sigma_\theta = \frac{1}{N} \sum_n \theta_n \theta_n^t.$$

Inserting this into (3.12) results in

$$\Sigma_\theta = \frac{1}{N} \sum_n E\left(\theta_n \theta_n^t | \mathbf{y}_n, \xi\right) \tag{3.13}$$

where

$$E\left(\theta_n \theta_n^t | \mathbf{y}_n, \xi\right) = \int \ldots \int \theta_n \theta_n^t f[\theta_n | \mathbf{y}_n, \Sigma_\theta] \, d\theta_n$$

and the posterior density has a form

$$f[\theta_n | \mathbf{y}_n, \Sigma_\theta] = \frac{\prod_k p(y_{nk}|\theta_n) \, g(\theta_n | \Sigma_\theta).}{\int \cdots \int \prod_k p(y_{nk}|\theta_n) \, g(\theta_n | \Sigma_\theta) \, d\theta_n}.$$

The likelihood equations for the item parameters of the choice model, γ_{i1} and γ_{i2}, are also easily found by using Fisher's identity. The probability of the choice pattern $\mathbf{d}_n$ of student n given θ and the item parameters γ is

$$P(\mathbf{d}_n|\theta, \gamma) = \prod_{i=1}^{k} P_i(\theta)^{d_{ni}} (1 - P_i(\theta))^{1-d_{ni}}$$

where $P_i(\theta) = \pi_{i1}(\theta) - \pi_{i2}(\theta)$ where π_{i1} and π_{i2} are given by Formula (3.11), and the logarithm of this function is

$$\log P(\mathbf{d}_n|\theta_n, \gamma) = \sum_{i=1}^{k} [d_{ni} \log(P_i(\theta_n)) + (1 - d_{ni}) \log(1 - P_i(\theta_n))].$$

Using the short-hand notation $P_i = P_i(\theta_n)$,for $j = 1$ and $j = 2$ we have

$$\frac{\partial \log P_i}{\partial \gamma_{ij}} = \left[d_{ni} \frac{1}{P_i} \frac{\partial P_i}{\partial \gamma_{ij}} - \frac{(1 - d_{ni})}{(1 - P_i)} \frac{\partial P_i}{\partial \gamma_{ij}} \right]$$

with

$$\frac{\partial Pi}{\partial \gamma_{ij}} = \frac{\partial \pi_{i1}}{\partial \gamma_{ij}} - \frac{\partial \pi_{i2}}{\partial \gamma_{ij}}$$

Let δ_{kj} denote the Kronecker symbol, that is equal to one if $k = j$ and equal to zero if $k \neq j$. Then

$$\frac{\partial \pi_{ik}}{\partial \gamma_{ij}} = -\delta_{kj}(\pi_{ik}(1 - \pi_{ik}))$$

and

$$\frac{\partial \pi_{i1}}{\partial \gamma_{ij}} - \frac{\partial \pi_{i2}}{\partial \gamma_{ij}} = -\delta_{1j}(\pi_{i1}(1 - \pi_{i1})) + \delta_{2j}(\pi_{i2}(1 - \pi_{i2}).$$

Substitution of this expression into (3.12) and dropping the short-hand notation we obtain the equations

$$\mathbf{h}(\gamma_{i1}) = \frac{\partial}{\partial \gamma_{i1}} \log L_{\mathbf{Y}}(\xi) = -\sum_n E\left((d_{ni} - P_i(\theta_n)) \frac{\pi_{i1}(\theta_n)(1 - \pi_{i1}(\theta_n))}{P_i(\theta_n)(1 - P_i(\theta_n))} \middle| \mathbf{y}_n, \xi \right) = 0,$$

and

$$\mathbf{h}(\gamma_{i2}) = \frac{\partial}{\partial \gamma_{i2}} \log L_{\mathbf{Y}}(\xi) = \sum_n E\left((d_{ni} - P_i(\theta_n)) \frac{\pi_{i2}(\theta_n)(1 - \pi_{i2}(\theta_n))}{P_i(\theta_n)(1 - P_i(\theta_n))} \middle| \mathbf{y}_n, \xi \right) = 0,$$

for $i = 1, ..., k$. In the analysis presented in this article, these equations were solved simultaneously with the well-known MML equations for the item parameters of the multidimensional version of the GPCM and the MML equation for the covariance matrix given by (3.13).

Fit of IRT models can be evaluated using the Lagrange Multiplier (LM) test (Glas, 1998, 1999; Glas & Suarez-Falcon, 2003; te Marvelde, Glas, Van Landeghem, & Van Damme, 2006). The LM test can be used to test an IRT model against a more general alternative, which is an IRT model with additional parameters. The LM statistic is evaluated using the MML-estimates of the special IRT model only. The statistic is asymptotically chi-square distributed with degrees of freedom equal to the number of fixed parameters. With a proper choice of alternative model, the LM test becomes

a test based on residuals, that is, a test based on differences between observed and expected frequencies.

As the alternative model we choose

$$p\left(X_{ni}=j|d_{ni}=1;\theta_n\right)=\frac{\exp\left(j\left(\sum_{q=1}^{Q}\alpha_{iq}\theta_{nq}\right)-\sum_{h=1}^{j}\beta_{ih}+d_{ns}j\delta\right)}{1+\sum_{h=1}^{m}\exp\left(h\left(\sum_{q=1}^{Q}\alpha_{iq}\theta_{nq}\right)-\sum_{k=1}^{h}\beta_{ik}+d_{ns}h\delta\right)}, \tag{3.14}$$

for $j=1,\ldots,m_i$. In the model under the null hypothesis, the additional parameters δ_j are equal to zero and the model is equivalent with the multidimensional GPCM given by (3.6).

Following Glas (1999) it can be inferred that the first order derivative of the likelihood with respect to δ_j is given by

$$v_i=\sum_n d_{ns}d_{ni}x_{ni}-\sum_n d_{ns}d_{ni}E\left(X_{ni}|\theta)\ \mathbf{x}_n\right),$$

A test for the null-hypothesis $\delta=0$ can be based on the fit-statistic

$$T_i=v_i^2/w_i \tag{3.15}$$

where w_i is the variance matrix of v_i, which is the opposite of the second order derivatives of the likelihood function with respect to the δ-parameter. If the statistic T_i is evaluated using $\delta=0$, that is, using the MML-estimates of the null-model, T_i has an asymptotic χ^2 distribution with one degree of freedom (see Glas, 1999; te Marvelde, Glas, Van Landeghem, & Van Damme, 2006).

4

Test Statistics for Models for Continuous Item Responses

ABSTRACT: The theory of estimating and testing item response models for continuous responses is developed in a marginal maximum likelihood framework. It is shown that the fit to the model can be evaluated using Lagrange multiplier tests. The tests focus on the assumed form of the response functions, differential item functioning, local stochastic independence and the factor structure underlying the responses. The tests are illustrated with an example of the analysis of data from central examinations in secondary education in the Netherlands. Using simulation studies, it is shown that the tests have good properties in terms of control of Type I error rate and power.

4.1. Introduction

Item response theory (IRT) models are stochastic models for two-way data; say the responses of students to items. An essential feature of these models is parameter separation, that is, the influences of the items and students on the responses are modeled by distinct sets of parameters. IRT provides the theoretical underpinning for computer adaptive testing, the use of incomplete assessment designs, equating and linking of assessments, evaluation of differences between groups and differential item functioning. Most applications of IRT models pertain to categorical data (Rasch, 1960; Samejima, 1969; Bock, 1972; Lord, 1980; Masters, 1982). However, also situations may arise where the responses to the items are continuous. An example is the so-called analogous-scale item format where a respondent marks the position on a line to express his or here opinion about some topic.

IRT models for continuous responses are outlined by such authors as Mellenbergh (1994), Moustaki (1996) and Skrondal and Rabe-Hesketh (2004). The present article focuses on testing the models for continuous responses. A method for testing model fit will be proposed in the framework of multidimensional IRT models for continuous responses and marginal maximum likelihood (MML) estimation. The model assumptions evaluated are subpopulation invariance (the violation is often labeled differential item functioning), the form of the item response function, local stochastic independence and the factor structure of the model. An analysis pertaining to scaling students' scores on a number of examination topics will be given as an example of the methods proposed. Finally, a number of simulation studies will be presented that assess the Type I error rate and the power of the proposed tests.

4.2. The Model

Consider a two-dimensional data matrix X with entries x_{nk}, for $n = 1, ..., N$, and $k = 1, ..., K$. The matrix contains the responses of students to items. It is assumed that the response of the student n on the item k is normally distributed, that is

$$P(x_{nk} \mid \theta_n, \alpha_k, \beta_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left(-\frac{(x_{nk} - \tau_{nk})^2}{2\sigma_k^2}\right). \tag{4.1}$$

The expectation of the item response is a linear function of the explanatory variables,

$$\begin{aligned} \tau_{nk} &= \sum_{h=1}^{H} \alpha_{kh}\theta_{nh} - \beta_k \\ &= \alpha_k'\theta_n - \beta_k, \end{aligned} \tag{4.2}$$

where α_k is a vector of the parameters $\alpha_{k1}, ..., \alpha_{kh}, ..., \alpha_{kH}$ which are usually called factor loadings and β_k is a location parameter. Further, $\theta_n = (\theta_{n1}, ..., \theta_{nh}, ..., \theta_{nH})$ is the H-dimensional proficiency parameter of student n. We assume that the density of θ_n is described by the normal distribution with the expectation μ_θ and the covariance matrix Σ_θ. The distribution will be denoted by $g(\theta_n; \mu_\theta, \Sigma_\theta)$. The model is in part identified by the restriction $\mu_\theta = 0$. Additional restrictions must be imposed to completely identify the model. The restrictions will be returned to below. Further, we assume that the variance $\sigma_k^2 = 1$, for all k. That is, we assume that all the observed responses have the same scale.

In the case of discrete responses, the data are the response patterns of the students, and these counts are seldom, if ever, transformed. In the case of continuous responses, transformations can be applied to the responses. For instance, if the model given by (4.1) is used to analyze response times, the observations x_{nk} should be the logarithms of the response times.

4.3. Estimation

To introduce the test statistics and to derive their asymptotic distributions, first some theory on MML estimation for IRT models (see Bock & Aitkin, 1982) must be outlined.

Preliminaries

Let η be a vector of model parameters, that is, η consists of the vectors $\alpha, \beta, \mu_\theta$, and $vec(\Sigma_\theta)$, where $vec(\Sigma_\theta)$ is a vector containing the diagonal and lower-diagonal elements of Σ_θ. The marginal log-likelihood function can then be written as

$$\log L(\eta, \mathbf{X}) = \sum_n \log \Pr(\mathbf{x}_n; \eta) \tag{4.3}$$

where $\mathbf{x}_n$ is the response pattern of the student n. The MML estimation equations are derived upon equating the vector of derivatives of the log-likelihood function to zero.

The first order derivatives can be derived using Fisher's identity (Louis, 1982). In the framework of IRT, Fisher's identity is given by

$$\mathbf{h}(\eta) = \frac{\partial}{\partial \eta} \log L(\eta, \mathbf{X}) = \sum_n E\left(b_n(\eta) | \mathbf{x}_n, \eta\right), \tag{4.4}$$

where the expectation is with respect to the posterior expectation $p(\theta_n | \mathbf{x}_n, \eta)$. Further

$$b_n(\eta) = \frac{\partial}{\partial \eta} \log p(\mathbf{x}_n, \theta_n; \eta) = \frac{\partial}{\partial \eta} \left[\log p(\mathbf{x}_n | \theta_n, \alpha, \beta) + \log g(\theta_n; \mu_\theta, \Sigma_\theta)\right]. \tag{4.5}$$

Notice that the derivative is a sum of the logarithm of the probability of the response pattern and the logarithm of the density of the student ability parameter. The power of Fisher's identity is that the derivatives are very easy to derive, while the derivation of $\mathbf{h}(\eta)$ is a cumbersome enterprise (Glas, 1999; te Marvelde, et al., 2006). Moreover, direct derivation of the matrix of second order derivatives needed for the computation of the standard errors of the estimates is even more demanding. However, using Fisher's identity repeatedly, Louis (1982) shows that the Fisher information matrix

$$H(\eta, \eta) = -\frac{\partial^2 L(\eta, \mathbf{X})}{\partial \eta \partial \eta'}, \tag{4.6}$$

is the sum over students n of terms

$$-E\left(B_n(\eta, \eta) | \mathbf{x}_n, \eta\right) - E\left(b_n(\eta) b_n(\eta)' | \mathbf{x}_n, \eta\right) + E\left(b_n(\eta) | \mathbf{x}_n, \eta\right) E\left(b_n(\eta) | \mathbf{x}_n, \eta\right)', \tag{4.7}$$

where

$$B_n(\eta, \eta) = \frac{\partial^2 \log Pr(\mathbf{x}_n, \theta_n; \eta)}{\partial \eta \partial \eta'}.$$

Glas (1998, 1999) and Glas and Suarez-Falcon (2003) show that in the case of the two- and three-parameter logistic model and the nominal response model, the second derivatives can be approximated by

$$H(\eta, \eta) \approx \sum_n E\left(b_n(\eta) | \mathbf{x}_n, \eta\right) E\left(b_n(\eta) | \mathbf{x}_n, \eta\right)'. \tag{4.8}$$

Below, the precision of this approximation will be evaluated empirically.

The exact expressions for the information matrix derived using (4.7) and (4.8) are given in Appendix.

4.3.1. Application to the IRT model for Continuous Responses

The logarithm of the marginal likelihood function for responses following the model given by (4.1) is

$$\log L(\eta|x) = \sum_n \log \int \cdots \int \prod_k p(x_{nk}|\theta_n, \alpha_k, \beta_k)\, g(\theta_n; \Sigma_\theta)\, d\theta_n. \tag{4.9}$$

where η is the ensemble of item and population parameters, $\eta = (\alpha', \beta', vec(\Sigma_\theta))'$, and $vec(\Sigma_\theta)$ is defined as a vector of the diagonal and lower-diagonal elements of Σ_θ. Further, $g(\theta_n; \Sigma_\theta)$ is the density of θ_n which assumed to be a multivariate normal distribution with mean vector $\mu_\theta = \mathbf{0}$ and variance-covariance Σ_θ. The maximum of (4.9) as a function of η is found as the solution of the equations $\partial \log L(\eta|x)/\partial\eta = 0$. These equations are easily found using Fisher's identity, which is given by (4.4). For instance, since θ_n has a multivariate normal distribution, equating the first order derivatives of $g(\theta_n; \Sigma_\theta)$ with respect to $vec(\Sigma_\theta)$ to zero gives

$$\Sigma_\theta = \frac{1}{N} \sum_n \theta_n \theta_n'.$$

Application of (4.4) and (4.5) gives the estimation equation

$$\Sigma_\theta = \frac{1}{N} \sum_n E\left(\theta_n \theta_n' | x_n, \eta\right) \tag{4.10}$$

where

$$E\left(\theta_n \theta_n' | x_n, \eta\right) = \int \ldots \int \theta_n \theta_n' f[\theta_n | \mathbf{x}_n, \Sigma_\theta]\, d\theta_n,$$

where the posterior density has a form

$$f[\theta_n | \mathbf{x}_n, \Sigma_\theta] = \frac{\prod_k p(x_{nk}|\theta_n, \alpha_k, \beta_k)\, g(\theta_n; \Sigma_\theta)}{\int \cdots \int \prod_k p(x_{nk}|\theta_n, \alpha_k, \beta_k)\, g(\theta_n; \Sigma_\theta)\, d\theta_n}. \tag{4.11}$$

The likelihood equations for β_k ($k = 1, \ldots, K$) as found in an analogous way. The observations x_{nk} have a normal distribution with expectation τ_{nk} and this expectation is linear in β_k. Equating the first order derivatives of $p(x_{nk}|\theta_n, \alpha_k, \beta_k)$ to zero gives

$$\sum_n x_{nk} = \sum_n \tau_{nk},$$

and application of (4.4) and (4.5) gives the estimation equation

$$\sum_n x_{nk} = \sum_n E\left(\tau_{nk} | \mathbf{x}_n, \eta\right), \tag{4.12}$$

for $k = 1, ..., K$. Similarly, the likelihood equations for α_{kh} ($k = 1, ..., K$, $h = 1, ..., H$) are obtained as

$$\sum_n x_{nk} E\left(\theta_{nh} | \mathbf{x}_n, \eta\right) = \sum_n E\left(\tau_{nk}\theta_{nh} | \mathbf{x}_n, \eta\right). \tag{4.13}$$

All these expressions can be solved simultaneously. In practice this is done by the Newton-Raphson algorithm, the EM (expectation-maximization) algorithm (Dempster, Laird & Rubin, 1977), or a combination of the two, where the EM algorithm is used as a first approximation and the Newton-Raphson algorithm is used when the estimates are sufficiently close to the desired maximum (Bock and Aitkin, 1981).

4.3.2. Identification of the Model

To identify the model the restriction $\mu_\theta = 0$ was imposed. Analogous to multidimensional IRT models for discrete responses, the model can be identified further in two ways (see, for instance, Béguin & Glas, 2001). The first approach requires setting the covariance matrix to the identity matrix and introducing the constrains $\alpha_{jq} = 0$ $j = 1...q-1$ and $q = j+1...Q$. The latent ability dimensions are independent of each other. The first item loads on the first dimension only. The second item loads on the first two dimensions only, and so on until item Q loads on the first $Q-1$ dimensions. All other items load on all dimensions.

The second approach to identify the model is setting the mean equal to the zero and considering the covariance matrix as a parameter of proficiency distribution that must be estimated. Further, the model is identified by imposing the restrictions, $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, ..., Q$ and $q = 1, ..., Q$. So, here the first item defined the first dimension, the second item defines the second dimension, and the third item defines the third dimension. The covariance matrix $\mathbf{\Sigma}_\theta$ describes the relation between the defined latent dimensions.

The transformation between the two parameterizations can be done as follows. Let $\mathbf{A}^o$ and $\mathbf{A}$ be the matrices of discrimination parameters for the first and the second approaches, respectively. According to Béguin and Glas (2001), θ can be transformed to θ^o by $\theta^o = \mathbf{L}^{-1}\theta$, where $\mathbf{L}$ is the Cholesky decomposition of $\mathbf{\Sigma}_\theta$. Since $\mathbf{L}$ is lower triangular and $\mathbf{A}\theta = \mathbf{A}\mathbf{L}\theta^o = \mathbf{A}^o\theta^o$, the restrictions $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, ..., Q$ and $q = 1, ..., Q$, are transformed into $\alpha^o_{jq} = 1$ for $j = 1, ..., Q-1$

and $q = j + 1, ..., Q$. Let us defining the lower triangular matrix $\mathbf{F}$ as the first Q rows of $\mathbf{A}^o$ and using $\theta = \mathbf{F}\theta^o$, we obtained $\Sigma_\theta = \mathbf{F}\mathbf{F}'$ and $\mathbf{A} = \mathbf{A}^o\mathbf{F}^{-1}$, which in turn produces restrictions $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, ..., Q$ and $q = 1, ..., Q$.

4.3.3. Computation

For solving the estimation equations the EM can be used. This general iterative algorithm for ML estimation in incomplete data problems handles missing data, first, by replacing missing values by a distribution of missing values, second, by estimating new parameters given this distribution, and, third, by re-estimating the distribution of the missing values assuming the new parameter estimates are correct. This process is iterated until convergence is achieved. The multiple integrals that appear above can be evaluated using adaptive Gauss-Hermite quadrature (Schilling & Bock, 2005). A critical point related to using Gauss-Hermite quadrature is the dimensionality of the latent space, that is, the number of latent variables that can be analyzed simultaneously. Wood et al. (2002) indicates that the maximum number of factors is 10 with adaptive quadrature, 5 with non-adaptive quadrature and 15 with Monte Carlo integration.

4.4. Testing the Model

4.4.1. Preliminaries

The Lagrange Multiplier (LM) test by Aitchison and Silvey (1958) is grounded on the following rationale. Consider some general parameterized model, and a special case of the general model, the so-called restricted model. The restricted model is derived from the general model by imposing constraints on the parameter space. In many instances, this is accomplished by fixing one or more parameters of the general model to constants. The LM test is based on the evaluation of the first-order partial derivatives of the log-likelihood function of the general model, evaluated using the maximum likelihood estimates of the restricted model. The unrestricted elements of the vector of first-order derivatives are equal to zero, because their values originate from solving the likelihood equations. The magnitudes of the elements of the vector of first-order partial derivatives corresponding to restricted parameters determine the value of the statistic: the closer they are to zero, the better the model fits.

More formally, let us consider a null-hypothesis about a model with parameters η_0. This model is derived from the general model with parameters η by fixing one or more parameters to known constants. We can make a partition of η_0 as $\eta_0 = (\eta_{01}', \eta_{02}')'$, and postulate constants described by vector η_{02} that is, $\eta_{02} = \mathbf{c}$. In the applications presented below, the restricted model is the IRT model, so $\eta_{01} = (\alpha', \beta', vec(\Sigma_\theta))'$, and the constants will be zero, that is, $\mathbf{c} = \mathbf{0}$. The partial derivatives of the log-likelihood function of first and second order are $\mathbf{h}(\eta) = \partial \log L(\eta)/\partial\eta$ and $\mathbf{H}(\eta, \eta) = -\partial^2 \log L(\eta)/\partial\eta\partial\eta'$ accordingly. Then, the LM statistic is given by

$$LM = \mathbf{h}(\eta_0)'\mathbf{H}(\eta_0, \eta_0)^{-1}\mathbf{h}(\eta_0). \tag{4.14}$$

For the case of a partitioned η, at the point of the LM estimates η_{01}, the free parameters have partial derivatives equal to zero, $\mathbf{h}(\eta_{01}) = \mathbf{0}$. This simplifies (4.14) to

$$LM(\mathbf{c}) = \mathbf{h}(\mathbf{c})' \, \mathbf{W}^{-1}\mathbf{h}(\mathbf{c}), \tag{4.15}$$

where

$$\mathbf{W} = \mathbf{H}_{22}(\mathbf{c}, \mathbf{c}) - \mathbf{H}_{21}(\mathbf{c}, \eta_{01})\,\mathbf{H}_{11}(\eta_{01}, \eta_{01})^{-1}\,\mathbf{H}_{12}(\eta_{01}\mathbf{c}), \tag{4.16}$$

and the partitioning of $\mathbf{W}$ is according to the partition $\eta_0 = (\eta_{01}', \mathbf{c}')'$. The LM statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in η_2 (Aitchison & Silvey, 1958). The LM test is equivalent with the efficient score test (Rao, 1947) and the modification index that is commonly used in structural equation modelling (Sörbom, 1989). Sörbom (1989) shows that the value of the LM statistic is proportional to the expected increase of the conditional likelihood should the additional parameters be estimated. In the next section, we will introduce four LM statistics targeted at differential item functioning, the shape of the item response curve, local independence and the factor structure.

4.4.2. Differential Item Functioning

Differential item functioning (DIF) is a difference in item response behavior between equally proficient members of two or more groups. As an example, consider the difference in response behavior between boys and girls. It could be that performance of boys on science and mathematical items is better than performance for girls. On the other hand, the performance of girls on language items could be better than the performance of boys. By itself, however, this does not indicate differential item functioning. Differential item functioning arises when, for a certain item, the level

of performance of equally proficient boys and girls is different, probably because the item refers to irrelevant knowledge that is more ubiquitous in one population than in the other.

There are several techniques for detection of DIF and most of them are based on the evaluation of differences in response probabilities between groups conditional on a measure of proficiency. In the framework of the LM test, this is accomplished as follows. First we define a background variable to distinguish between the groups. Only two groups are considered here, say the reference and the focal group. The generalization to more groups is straight forward. Define Y_n as

$$Y_n = \begin{cases} 1 & \text{if } n \text{ belongs to the focal group} \\ 0 & \text{if } n \text{ belongs to the reference group.} \end{cases} \tag{4.17}$$

Let η_{01} be a vector of the item parameters α**,** β and the parameters of the population distribution of the abilities of the students $vec(\Sigma_\theta)$. So $\eta_{01} = (\alpha,\beta,vec(\Sigma_\theta))$. In the alternative model, the expectation of the item response, τ_{nk}, is a linear function of item parameters as in (4.2) and an additional parameter δ_k, that is,

$$\tau_{nk} = \alpha_k'\theta_n - \beta_k + \delta_k Y_n. \tag{4.18}$$

If $\delta = 0$, the null model holds. In the alternative model, δ_k is a free parameter. Note that δ_k can be interpreted as a shift in the item parameter β_k in the focal group. To test whether this parameter significantly differs from zero, the LM statistic defined by (4.14) and (4.15) can be used with $\eta_{02} = \mathbf{c} = \delta_k$.

The expression for $\mathbf{h}(\tau_{nk})$ can be found in the same way as (4.12) was derived. It easily follows that the first derivatives of the log-likelihood function, $\mathbf{h}(\delta_k) = \partial \log L(\eta)/\partial\delta_k$, are given by

$$\mathbf{h}(\delta_k) = \sum_n x_{nh} Y_n - \sum_n Y_n E\left(\tau_{nk}\,|x_n\right). \tag{4.19}$$

Substitution of this expression into (4.16) provides the expression for the LM statistic

$$LM = \frac{\left(\sum\limits_n x_{nk} Y_n - \sum\limits_n Y_n E\left(\tau_{nk}\,|x_n\right)\right)^2}{\mathbf{W}}, \tag{4.20}$$

where $\mathbf{W}$ is defined by (4.16). Note that $\mathbf{W}$ is now is a scalar. $\mathbf{W}$ can be interpreted as the variance of $\mathbf{h}(\eta_{02})$ give the parameter estimates. Expressions for $\mathbf{W}$ can be derived using either (4.7) or (4.8). The LM statistic has an asymptotic χ^2-distribution with one degree of freedom.

4.4.3. Shape of the Item Response Function

The shape of the response function, that is, the appropriateness of $p(x_{nk}|\theta_n, \alpha_k, \beta_k)$ in describing the response probabilities, could, in principle, be evaluated by partitioning the space of θ into a number of subsets and comparing the observed and expected responses averaged over the respondents in the subsets. However, in the framework of IRT models for discrete responses, Orlando and Thissen (2000) remarked that the grouping of respondents based on an estimate of θ rather than on some directly observable statistic violates the assumption of the traditional χ^2-goodness-of-fit-test, and, as a result, the distribution of such statistics remains unclear. As an alternative, they proposed statistics where the grouping of respondents is based on directly observable number-correct scores rather than on estimates of θ. An analogous approach will also be pursued here.

For a test targeted at item k, we partition the sample of respondents using a number of boundaries for the total score obtained on all the other items. So let the item of interest be labelled k and the other items are labelled $j = 1, 2, ..., k-1, k+1, ..., K$. Let $\mathbf{x}^{(k)}$ be the response pattern without item k, and let $r\left(\mathbf{x}^{(k)}\right)$ be the number-correct score on this partial response pattern,

$$r\left(\mathbf{x}^{(k)}\right) = \sum_{j \neq k} x_j. \tag{4.21}$$

$r\left(\mathbf{x}^{(k)}\right)$ is often called a rest-score. The range of possible scores $r\left(\mathbf{x}^{(k)}\right)$ is partitioned into S_k intervals. Furthermore, define

$$\mathbf{w}\left(s, \mathbf{x}^{(k)}\right) = \begin{cases} 1 & \text{if } r_{s-1} \leq r\left(\mathbf{x}_n^{(k)}\right) < r_s, \\ 0 & \text{otherwise,} \end{cases} \tag{4.22}$$

for $s = 1, ..., S_k$ with $r_0 = -\infty$ and $r_{S_k} = \infty$. So $w\left(s, \mathbf{x}^{(k)}\right)$ is an indicator function assuming a value equal to 1 if the number correct score of response pattern $\mathbf{x}^{(k)}$ is in score range s. The expectation of the item response under the alternative model has the form

$$\tau_{nk} = \alpha_k' \theta_n - \beta_k + \sum_{s=1}^{S-1} \mathbf{w}\left(s, \mathbf{x}^{(k)}\right) \delta_s. \tag{4.23}$$

Note that $\mathbf{w}\left(s, \mathbf{x}^{(k)}\right)$ is equal to one for only one of the S score segments, so the summation defined in (4.23) only selects one of the parameters δ_s. The parameter δ_s gauges the shift in item parameter β_k for score group s. Finally, note that there is no

parameter δ_S; that is, the highest score level is used as a base line. If δ_S would also be present, the model defined by (4.23) would no longer be identified.

The application of the LM statistic to test the model is analogous to the application to DIF. If $\delta = (\delta_1, ..., \delta_{S-1}) = \mathbf{0}$, the null model holds. In the alternative model, δ is a free parameter that can be interpreted as a shift in the item parameter β_k. To test whether this parameter significantly differs from zero, the LM statistic defined by (4.14) and (4.15) can be used with $\eta_{02} = \mathbf{c} = \delta$. The statistic has an asymptotic χ^2-distribution with $S - 1$ degrees of freedom.

The expression for $\mathbf{h}(\delta)$ can be found in the same way as (4.12) was derived. The expression for first derivative with respect to δ_s is

$$\mathbf{h}(\delta_s) = \sum_n \mathbf{w}\left(s, \mathbf{x}^{(k)}\right) x_{nk} - \sum_n \mathbf{w}\left(s, \mathbf{x}^{(k)}\right) E\left(\tau_{nk} \,|x\right). \tag{4.24}$$

Note that the first order derivative is the difference between the observed scores and expected scores of persons in subgroup s. The simplest form of the test emerges if only two score levels are considered, that is, if $S_k = 2$. In that case, one could set the cut-off score r_1 somewhere in the middle of the score range, say, $r_1 = 0$, and test whether students with a high rest-score $r\left(\mathbf{x}^{(k)}\right)$ perform better or worse as expected on the target item k. The distribution of this version of the test statistic has one degree of freedom.

4.4.4. Local Independence

The assumption of local stochastic independence requires the association between the items to vanish given the parameters. If, for instance, we want to test whether an item response depends on the previous item, we define the indicator function

$$\mathbf{w}\left(s, x_{(k-1)}\right) = \begin{cases} 1 & \text{if } r_{s-1} \le x_{(k-1)} < r_s, \\ 0 & \text{otherwise,} \end{cases} \tag{4.25}$$

for $s = 1, ..., S_k$ with $r_0 = -\infty$ and $r_{S_k} = \infty$. As before, the simplest form of the test emerges if only two score levels are considered, and test whether students with a high score on the previous item perform better or worse than expected on the target item.

The expression for expectation of the item response, τ_{nk}, has a form

$$\tau_{nk} = \alpha_k' \theta_n - \beta_k + \sum_{s=1}^{S-1} \mathbf{w}\left(s, x_{(k-1)}\right) \delta_s. \tag{4.26}$$

and last contribution describes the effect of item $k-1$ on item k.

Parameter δ_k reflects the alternative model and we can get the expression for first derivative

$$\mathbf{h}(\eta_{02}) = \sum_n x_{nk}\mathbf{w}\left(s, x_{(k-1)}\right) - \sum_n \mathbf{w}\left(s, x_{(k-1)}\right) E\left(\tau_{nk}\,|x_n\right). \tag{4.27}$$

Note that, analogous to the test for the shape of the response functions, in this case the first order derivative is equal to the difference between the observed scores and expected scores of persons in subgroup s again. Also in this case, the simplest form of the test emerges if only two score levels are considered, that is, if $S_k = 2$. In that case, one could set the cut-off score r_1 somewhere in the middle of the score range of item $k-1$ and test whether students with a high score on item $k-1$ perform better or worse as expected on the target-item k.

4.4.5. Tests for the Factor Structure

Above it was argued that the model can be identified by setting a mean and a covariance matrix equal to zero and the identity matrix, respectively, and introducing the constraints $\alpha_{jq} = 0$, for $j = 1, ..., Q-1$ and $q = j+1, ..., Q$. In this approach, the model is identified by assuming that the responses on the first item are uniquely determined by the first ability dimension, the responses on the second items are uniquely determined by a mixture of the first and the second ability dimension, and so forth. In general, these identification restrictions will not support a very reliable interpretation of the ability dimensions. Therefore, in an exploratory factor analysis, the factor solution is usually visually or analytically rotated. Often a rotation scheme is devised to approximate Thurstone's simple-structure criterion (Thurstone, 1947), where the factor loadings are split into two groups, the elements of the one tending to zero and the elements of the other tending toward unity. In the framework of multidimensional IRT models, Béguin and Glas (2001) suggest an approach that has much in common with Thurstone's approach. The idea is to identify the dimensions with subscales of items loading on one dimension only, either by identifying these $S \leqslant Q$ subscales a priori, or by identifying them using an iterative search based on fitting S unidimensional IRT models. The procedure can be characterized as a top-down procedure, that starts with the set of all items and discards the non-fitting items using fit-statistics for unidimensional IRT models. The test statistics discussed above can be used for the evaluation of item fit. After identification of S sets of scaled items, there will usually be a set of remaining items that load on all ability dimensions. Next,

restrictions will be imposed on the matrix of discrimination parameters **A** to reflect the structure of the subscales found, that is, if item j belongs to subscale q, $\alpha_{jq} = 1$, and $\alpha_{jq'} = 0$, for $q' = 1, ..., Q,\ q' \neq q$. Finally, if more dimensions than subscales are specified, the remaining dimensions $q = S+1, , ..., Q$ are identified using $\alpha_{j'q} = 1$,and $\alpha_{j'q'} = 0$ $(q' = 1, ..., Q,\ q' \neq q)$, for some item j' not belonging to a subscale.

To test this factor structure, we adopt the null-hypothesis H_0: $\alpha_{jl} = 0$. The model is estimated under the null-hypothesis, that is, the likelihood equation (4.13) is not solved for the parameter α_{jl}, but this parameter is restricted to zero. Given the MML estimates of the free parameters, analogous to (4.13) the first order derivatives with respect to α_{jl} are given by

$$h\left(\alpha_{jl}\right) = \sum_{n} E\left(\left(x_{nj} - \tau_{nj}\right)\theta_{nl}\,|\mathbf{x}_n\right),$$

and the Lagrange Multiplier statistic is computed as

$$LM\left(\alpha_{jl}\right) = \frac{h\left(\alpha_{jl}\right)^2}{H_{22}\left(\alpha_{jl},\alpha_{jl}\right) - H_{21}\left(\alpha_{jl},\eta\right)H\left(\eta,\eta\right)^{-1}H_{12}\left(\alpha_{jl},\eta\right)},$$

where h and H are the partial derivatives of the log-likelihood function of the first and second order. The statistic has an asymptotic χ^2-distribution with one degree of freedom

4.5. An Empirical Example

The methods presented above are illustrated with an analysis of a data set of Dutch Central examinations. The data are a subset of the data of pre-university students that took their final examination in the school year 1994/1995. The students choose an examination package that consisted of different examination topics. For our illustration we choose one of the packages, which consisted of seven topics: Dutch, English and German Language, History, Mathematics, General Economy and Business Economy. The sample consisted of 445 students. The examination scores were on a scale of 0 to 10, with two significant digits after the decimal point.

The objective was to fit a model with a low dimensionality and a simple factor structure. First a unidimensional model was fitted and for every topic the test statistic for the shape of the response function was computed using two subgroups. This lead

Table 4.1: Parameter estimates for examination topics (Starred entries are fixed)

Topic	α_{k1}	α_{k2}	α_{k3}	β_k
Dutch	0.17	-0.04	0.37	-6.21
German	1.00*	0.00*	0.00*	-6.41
English	0.95	0.00*	0.00*	-6.44
History	0.38	0.10	0.52	-6.42
Mathematics	0.00*	1.00*	0.00*	-5.82
Gen. Econ.	0.00*	0.00*	1.00*	-6.16
Bus. Econ.	0.00*	0.28	0.57	-6.19
Covariance Matrix				
	0.786			
	0.325	0.571		
	0.430	0.558	0.605	
Correlation Matrix				
	1.000			
	0.485	1.000		
	0.623	0.949	1.000	

to the conclusion that the unidimensional model did not fit. Then a two-dimensional model was hypothesized, where one dimension should represent a language ability and the other a mathematics ability. However, for the outcome of the model tests showed that also this model was not tenable. The estimates of the final three-dimensional model are shown in Table 4.1.

The first two dimensions are identified by fixing the rows of the topics German and Mathematics. In Table 4.1, the fixed elements are marked with an asterisk. Note that German only loads on the first dimension, while Mathematics only loads on the second dimension. Therefore, these dimensions can be labeled as language ability and mathematics ability. The third dimension was identified as a dimension related to Economy, but also Dutch Language and History had significant loadings on this dimension. The estimated β-parameters are shown in the last column. Since the mean of the ability distribution is scaled to zero, the opposite of the β -parameters reflect the average scores on the topics. Note that the average score on Mathematics was the lowest and the average score on English Language was the highest. The estimated covariance matrix and the associated correlation matrix are given at the bottom of the table. Note that the Economics dimension correlated highly with the Mathematics dimension. The other two correlations are only moderate.

Next, model fit was evaluated using the four tests outlined above. Differential item functioning (or rather, differential test functioning, in the present case) was evaluated for the background variable Gender. The outcomes of the tests for differential item functioning are given in Table 4.2. The topics are in the same order as in Table 4.1.

Table 4.2: Lagrange tests for differential item functioning

			Boys		Girls	
Topic	LM	Prob	Obs	Expct	Obs	Expct
Dutch	6.46	.011	6.17	6.25	6.27	6.11
German	0.77	.380	6.55	6.53	6.14	6.19
English	8.86	.003	6.63	6.54	6.05	6.22
History	3.13	.077	6.57	6.51	6.12	6.22
Mathematics	0.01	.910	5.90	5.89	5.67	5.68
Gen. Econ.	4.08	.043	6.31	6.25	5.87	5.99
Bus. Econ.	0.19	.666	6.25	6.26	6.08	6.05

The second and third columns give the values of the LM statistics and the significance probabilities. The statistic has one degree of freedom. The tests are based on the differences between observed and expected average topic scores for boys and girls. The values are given in the last four columns. They can be used to assess the seriousness of a model violation. This is important because the power of the test increases with the sample size, and with a large sample size, a significant result is easily obtained. In the present example, the test for English has the lowest significance probability. This is because the Girls score on average 0.17 score point lower than expected and the Boys score on the average 0.09 score points higher than expected.

Table 4.3 and Table 4.4 give analogous results for the test for the shape and the test for local independence, respectively. For the first test, the sample was divided into two groups of students: a group with 50% of the students that obtained the lowest scores, and a group of 50% of the students that obtained the highest scores. Business Economy,was highly significant: The lower scoring group obtained a lower average score than expected and the higher scoring group obtained a higher average than expected.

For the evaluation of the assumption of local independence, for every topic k in the list the sample was again divided into two groups: students that scores low on Topic $k - 1$. In principle, all combinations of subjects could have been tested for

Table 4.3: Lagrange tests for the response function

			Low		High	
Topic	LM	Prob	Obs	Expct	Obs	Expct
Dutch	1.51	.219	5.98	5.94	6.43	6.47
German	0.98	.322	5.88	5.92	6.94	6.90
English	5.60	.018	5.89	5.97	6.97	6.90
History	1.46	.227	5.87	5.90	6.96	6.92
Mathematics	0.02	.893	5.41	5.40	6.24	6.24
Gen. Econ.	1.95	.163	5.61	5.67	6.71	6.65
Bus. Econ.	14.49	.000	5.67	5.79	6.71	6.59

violation of local independence; the example serves as an illustration of the method. With respect to local independence, it can be seen that the association between German and English Language and between General and Business Economics was not properly modeled. In both cases, the association between the topics was higher than expected. This suggests that a unique dimension for the two forms of economy might result in a better fit.

Table 4.4: Lagrange tests for local independence

				Low		High	
Topic 1	Topic 2	LM	Prob	Obs	Expct	Obs	Expct
German	Dutch	2.98	.084	6.09	6.16	6.74	6.67
English	German	27.41	.000	5.80	5.96	7.07	6.91
History	English	0.04	.838	6.04	6.04	6.79	6.78
Mathematics	History	1.57	.210	5.54	5.49	6.10	6.15
Gen. Econ.	Mathematics	0.08	.774	5.83	5.82	6.49	6.51
Bus. Econ.	Gen. Econ.	22.71	.000	5.68	5.84	6.70	6.53

Finally, Table 4.5 gives the results for the test targeted at the factor structure, that is, targeted at the factor loadings α_{kh} fixed to zero. These are the zero factor loadings marked with an asterisk in Table 4.1. Note that none of the tests was significant at the 5% level. So in this respect the model fitted very well.

Table 4.5: Lagrange test for the factor structure

Topic k	Dim h	LM	Prob
German	2	0.09	.755
German	3	0.28	.591
English	2	0.21	.644
English	3	0.86	.353
Mathematics	1	0.15	.690
Mathematics	3	0.61	.433
Gen. Econ.	1	0.00	.966
Gen. Econ.	2	0.70	.401
Bus. Econ.	1	3.28	.070

4.6. A Simulation Study of Type I Error Rate and Power

The Type I error rate or significance level of a test is the probability of rejecting the null hypothesis of perfect model fit when the null-model is true. In the present study, a significance level of 5% was used. On the other hand, power is the probability of rejecting the null hypothesis when a model violation occurs. One could call this the detection-rate or hit-rate. For all tests described above, both the Type I error rate and the power were studied using simulation studies. In these studies, data were generated according to the model under the null-hypothesis or the model under the alternative hypothesis, that is, under the null-model with an added model violation. In all studies, the sample size was varied as 500, 1000 and 4000.

The simulation studies were carried out in two setups. The first setup pertained to the tests of DIF, the shape of the response function, and local independence, the second setup pertained to the test for the factor structure. We will first outline the first setup and summarize the results. Then the second setup will be treated. In the simulations in the first setup, a unidimensional version of the model was used where the student parameters θ_n were drawn from a standard normal distribution. The number of items was varied as 10, 20 and 40, and the item location parameters β were equally spaced between -1.0 and 1.0. Finally, the item discrimination parameters α were all equal to 1.0.

4.6.1. Type I Error Rate

The study with respect to the Type I error rate was conducted using both the exact expressions for the second order derivatives given in (4.7) and in the Appendix, and

the approximation given by (4.8). The number of replications in the simulation study was 100 for each combination of the sample size and test length. For the test on DIF, the numbers of simulees in each group were equal. For the tests for the item response function and local independence, two score groups were formed (so $S_k = 2$ for all k) and the cut-off score was always equal to zero. As a result, the sizes of the two groups were approximately equal. The Type I error rate was computed as the number of tests significant at the 5% level aggregated over all items. The results are presented in Table 4.6.

Table 4.6: Type I error rate of three test statistics computed using exact and approximated matrices of second order derivatives

N	K	DIF Test		IRF Test		LID Test	
		Exact	Approx.	Exact	Approx.	Exact	Approx.
500	10	.05	.06	.04	.04	.03	.03
	20	.05	.04	.04	.04	.04	.06
	40	.05	.06	.05	.08	.03	.07
1000	10	.05	.04	.07	.04	.04	.04
	20	.06	.05	.05	.04	.05	.05
	40	.05	.06	.05	.06	.05	.07
4000	10	.06	.05	.05	.04	.04	.05
	20	.05	.05	.05	.05	.04	.05
	40	.05	.06	.05	.06	.05	.06

It can be seen that the control of Type I error rate was generally good. There were no main effects of sample size and test length. Further, there were no striking difference between the two versions of the statistic.

4.6.2. Differential Item Functioning

In the simulation study on the power of the tests to detect differential item functioning, three values were chosen for the effect size: $\delta = 0.1$, $\delta = 0.2$ and $\delta = 0.5$. Following the terminology of Cohen (1988), these effect sizes can be labelled as minimal, small and large. The item and person parameters were the same as in the study of the Type I error rate. Within every one of the 100 replications, the model violation was imposed on one randomly chosen item. The results are given in Table 4.7.

The columns labelled “Hits” give the proportion of replications for which the test on the differentially functioning item was significant at the 5% level. So these

columns give an estimate of the power of the test. The columns labelled "False Alarms" give the proportion of significant results for the items conforming to the model, aggregated over replications and all model conform items. These columns give an estimate of the Type I error rate.

Table 4.7: Detection of differential item functioning

			DIF Test		IRF Test		LID Test	
N	K	δ	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms
500	10	.1	.69	.06	.05	.04	.05	.03
		.2	1.00	.07	.05	.04	.05	.03
		.5	1.00	.15	.07	.04	.05	.04
	20	.1	.68	.06	.05	.05	.04	.06
		.2	1.00	.07	.08	.05	.05	.06
		.5	1.00	.09	.05	.05	.07	.06
	40	.1	.74	.07	.10	.06	.08	.07
		.2	1.00	.07	.07	.07	.08	.07
		.5	1.00	.08	.06	.08	.09	.08
1000	10	.1	.90	.06	.05	.04	.05	.04
		.2	1.00	.10	.05	.04	.03	.04
		.5	1.00	.22	.12	.04	.05	.04
	20	.1	.94	.06	.09	.04	.06	.06
		.2	1.00	.07	.07	.05	.07	.06
		.5	1.00	.10	.07	.05	.05	.07
	40	.1	.96	.06	.05	.07	.06	.07
		.2	1.00	.06	.06	.08	.09	.07
		.5	1.00	.07	.07	.08	.06	.08
4000	10	.1	1.00	.08	.07	.05	.07	.06
		.2	1.00	.23	.12	.05	.05	.07
		.5	1.00	.45	.31	.05	.05	.08
	20	.1	1.00	.05	.07	.05	.12	.11
		.2	1.00	.10	.05	.05	.11	.11
		.5	1.00	.23	.16	.06	.08	.12
	40	.1	1.00	.06	.07	.06	.13	.14
		.2	1.00	.06	.06	.06	.13	.14
		.5	1.00	.10	.07	.07	.08	.13

Note that the test on DIF displayed the largest proportion of hits; in most instances, this proportion was equal to 1.00. Note further that the proportion of hits for the test targeted to DIF has main effects of test length and sample size. Finally, the

control of Type I error rate, that is, the proportion of false alarms, remained generally close to the nominal significance level. The main exceptions occurred for the large effect size in combination with a short test. The explanation is that in these cases the imposed model violation was such that every combination led to a global model violation affecting all items. The two other statistics had both the proportion of hits and false alarms at the nominal significance level. Form a diagnostic perspective, it is desirable that tests have power against specific model violations, so this is a positive result.

4.6.3. Item Response Functions

The results of the simulation studies with respect to the power of the three tests to detect violation of the item response function are shown in Table 4.8. The power is reported in the columns labelled "Hits". It can be seen that in the present case the test targeted at DIF had no power. The test on the fit of the items response function had the highest power. But the test targeted at local independence had also power to detect violation, although its power was of course less than the power of the specific test. In both cases, there were clear main effects of the effect size δ, sample size and test length. Further, it can be seen that the Type I error rate was well under control.

Local Independence

The results for the detection of violations of local independence are shown in Table 4.9. It can be seen that the test targeted at violation of local independence now attained the highest power. Again, there were clear main effects of the effect size δ, the sample size and the test length. The test for the shape of the IRFs also had considerable power but the power of the test on DIF hardly exceeded the nominal significance level. For all three tests, the Type I errors were virtually similar to their nominal levels.

4.6.4. Type I Error Rate and Power of the Test for the Factor Structure

The test for the factor structure can only be meaningfully applied in a multidimensional version of the model, and, therefore, the setup chosen here was somewhat different. The simulations were run in two versions, say Study 1 and Study 2. In Study 1, the generating values of the item parameters and the covariance matrix were chosen equal to the parameter estimates obtained in the empirical example presented above. So the parameters used to generate the data are given in Table 4.1. In Study 2,

Table 4.8: Detection of violation of the item response function

			DIF Test		IRF Test		LID Test	
N	K	δ	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms
500	10	.1	.05	.06	.24	.06	.09	.03
		.2	.06	.06	.71	.07	.12	.04
		.5	.08	.06	1.00	.08	.23	.04
	20	.1	.05	.06	.27	.05	.14	.06
		.2	.06	.06	.86	.05	.18	.05
		.5	.05	.06	1.00	.06	.29	.05
	40	.1	.09	.07	.49	.08	.19	.08
		.2	.08	.07	.96	.07	.20	.08
		.5	.09	.07	1.00	.07	.29	.09
1000	10	.1	.06	.10	.26	.05	.14	.03
		.2	.05	.07	.94	.05	.24	.04
		.5	.07	.05	1.00	.06	.42	.05
	20	.1	.05	.06	.37	.05	.20	.06
		.2	.05	.05	.97	.06	.23	.06
		.5	.05	.06	1.00	.04	.37	.06
	40	.1	.06	.06	.60	.07	.18	.07
		.2	.05	.06	1.00	.07	.29	.08
		.5	.07	.06	1.00	.06	.43	.07
4000	10	.1	.05	.05	.69	.05	.21	.07
		.2	.05	.05	1.00	.09	.53	.04
		.5	.05	.05	1.00	.05	.90	.04
	20	.1	.07	.05	.91	.06	.34	.11
		.2	.08	.05	1.00	.06	.59	.21
		.5	.05	.05	1.00	.07	.88	.10
	40	.1	.07	.05	.97	.06	.44	.12
		.2	.03	.05	1.00	.05	.60	.12
		.5	.05	.05	1.00	.06	.86	.12

the covariance matrix remained the same, but there were nine items. It was assumed that the first three items only loaded on the first dimension, the second three items only loaded on the second dimension and the last three items only loaded on the last dimension. All factor loadings where either equal to one or zero. All β-parameters were equal to -6.0. The sample size was varied as 500, 1000 and 4000. In the studies to assess the power, the second factor loading of the second item, α_{22}, which was equal to zero in the null-model, was varied as $\alpha_{22} = 0.2$ and $\alpha_{22} = 0.5$. Each leg

Table 4.9: Detection of violation of local independence

			DIF Test		IRF Test		LID Test	
N	K	δ	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms
500	10	.1	.06	.05	.09	.04	.11	.04
		.2	.07	.06	.13	.05	.41	.04
		.5	.05	.06	.23	.05	.95	.04
	20	.1	.07	.06	.11	.05	.17	.05
		.2	.05	.06	.12	.06	.40	.06
		.5	.05	.06	.14	.05	.93	.06
	40	.1	.07	.07	.14	.08	.17	.07
		.2	.06	.07	.17	.08	.38	.07
		.5	.09	.07	.18	.08	.90	.07
1000	10	.1	.05	.05	.11	.05	.12	.04
		.2	.06	.05	.12	.04	.69	.04
		.5	.05	.05	.40	.04	1.00	.04
	20	.1	.05	.06	.14	.05	.13	.06
		.2	.06	.06	.12	.05	.64	.06
		.5	.05	.06	.26	.05	.98	.06
	40	.1	.05	.06	.10	.07	.11	.08
		.2	.07	.06	.12	.07	.60	.07
		.5	.06	.06	.14	.07	1.00	.07
4000	10	.1	.05	.05	.19	.05	.38	.06
		.2	.05	.06	.49	.05	1.00	.06
		.5	.06	.05	.91	.07	1.00	.05
	20	.1	.05	.05	.12	.05	.18	.12
		.2	.05	.05	.29	.05	.99	.12
		.5	.07	.05	.57	.05	1.00	.11
	40	.1	.05	.05	.12	.06	.20	.13
		.2	.06	.05	.19	.06	.95	.13
		.5	.06	.05	.27	.06	1.00	.13

of the study had 100 replications. The results are shown in Table 4.10. The rows with an effect size $\alpha_{22} = 0.0$ pertain to the Type I error rate. It can be seen that the Type I error rate was close to its nominal value of 5%. For the power studies, where the model was violated by choosing α_{22} unequal to zero, the situation is more complex than in the previously reported power studies. This has to do with the fact that in a multidimensional model, the model fit can be improved in more than one way. For instance, in Table 4.1, it can be seen that the factor pattern of Topic 2 and

Topic 3 are similar. So if α_{22} is erroneously specified as zero, while it is in fact 0.5, the specification error can not only be compensated by freeing α_{22}, but also by freeing α_{32} which will then move to a negative value. This is in fact what happened in both Study 1 and Study 2. Further, in Study 2, the first three rows of the matrix of factor loadings were the same, so here also the test for α_{12} should be sensitive to the model violation imposed on α_{22}. Therefore, the outcomes in the columns of Table 4.10 labeled Power are the proportions of significant outcomes of the LM tests for analogous elements in identical rows in the factor matrix. So in Study 1, the LM tests for α_{22}, and α_{32}, and in Study 2 the LM tests for α_{12}, α_{22}, and α_{32}. Note that the power has main effects of the effect size and the sample size, as was expected. Further, the Type 1 error rate was well under control. So the conclusion here is that the LM tests give a clear hint regarding the possible directions to obtain model fit, but it remains the choice of the content matter expert which direction to chose.

Table 4.10: Type I error rate and power of the test for the factor structure

	Effect	Study 1		Study 2	
N	Size	Power	Type I Error	Power	Type I Error
500	0.0		.04		.04
	0.2	.37	.06	.31	.04
	0.5	.63	.04	.55	.04
1000	0.0		.04		.04
	0.2	.40	.03	.45	.04
	0.5	.64	.02	.58	.04
4000	0.0		.05		.05
	0.2	.42	.05	.50	.05
	0.5	.70	.05	.60	.06

4.7. Conclusion

An MML framework for estimation and testing of a model for continuous responses was presented and simulation studies were conducted to assess the Type I error rate and power. The simulation studies showed that these tests had good properties. Further, the tests are based on residuals, that is, differences between observed and expected mean scores, that support an appraisal of the seriousness of the model

violation. Finally, the tests give an indication of the source of the lack of model fit, and provide a direction for model modification.

A final remark concerns the likelihood-based framework that was chosen for this study. An advantage of MML framework adopted here is that the item parameters and the covariance matrix can be estimated simultaneously. Therefore, the standard errors of the estimates and the distribution of the test statistics take all uncertainty into account. A disadvantage is the limit on the dimensionality of the model imposed by the computational restrictions. However, considerable progress has been made in broadening these limits (Schilling & Bock, 2005).

A well known alternative approach to estimating the model considered here is a Bayesian procedure using a Markov Chain Monte Carlo (MCMC) algorithm (see, for instance, Gelman et al., 1995). Examples are the procedures outlined by Shi and Lee (1998) and Béguin and Glas (2001). However. Bayesian estimation methods based on the MCMC algorithm are usually combined with data augmentation methods, and this also limits the size of the problems (in terms of number of persons, items and dimensions) that can be handled. Further, the procedures for testing model fit in a Bayesian framework are not yet satisfactory developed. At this moment, two approaches to testing model fit based on a philosophy comparable to the one used above are studies. The first approach is to use likelihood-based statistics as posterior predictive checks (Hoijtink, 2001, Glas & Meijer, 2003). As a general approach this may have problems because, as was pointed out by Maris (2005), the power characteristics of posterior predictive checks are far from optimal. An alternative approach, labeled Bayesian modification indices has been recently proposed by Fox and Glas (2005) but this approach has not yet been tested broadly for a general class of models. So for the time being, the proven robustness of MML estimation and testing methods still justifies their widespread use.

Appendix

4.A. Information Matrix for the Items

The information matrix is the sum over students n of terms

$$-E\left(B_n(\eta,\eta)|\mathbf{x}_n,\eta\right) - E\left(b_n(\eta)b_n(\eta)'|\mathbf{x}_n,\eta\right) + E\left(b_n(\eta)|\mathbf{x}_n,\eta\right)E\left(b_n(\eta)|\mathbf{x}_n,\eta\right)', \quad (4.28)$$

where

$$b_n(\eta) = \frac{\partial}{\partial\eta}\log Pr\left(\mathbf{x}_n,\theta_n;\eta\right) \quad (4.29)$$

and

$$B_n(\eta,\eta) = \frac{\partial^2 \log Pr\left(\mathbf{x}_n,\theta_n;\eta\right)}{\partial\eta\partial\eta'}. \quad (4.30)$$

The last term in (4.28) can be directly inferred from the estimation equations given by (4.12) and (4.13).

The kernel of the log-likelihood per student and item is given by

$$\log L_{nk} = -\frac{1}{2}(x_{nk}-\tau_{nk})^2, \text{ with } \tau_{nk} = \sum_h \alpha_{kh}\theta_{nh} - \beta_k.$$

For the items, the following derivatives are easily checked:

$$\frac{\partial \log L_{nk}}{\partial\alpha_{kh}} = -\theta_{nh}(X_{nk}-\tau_{nk})$$

$$\frac{\partial \log L_{nk}}{\partial\beta_k} = (x_{nk}-\tau_{nk})$$

$$\frac{\partial^2 \log L_{nk}}{\partial\alpha_{kh}^2} = \theta_{nh}^2$$

$$\frac{\partial^2 \log L_{nk}}{\partial\beta_k^2} = -1$$

$$\frac{\partial^2 \log L_{nk}}{\partial\alpha_{kh}\partial\alpha_{kp}} = \theta_{nh}\theta_{np}$$

$$\frac{\partial^2 \log L_{nk}}{\partial\alpha_{kh}\partial\beta_k} = -\theta_{nh}$$

Inserting these identities into (4.28) gives the information matrix for the items.

5

Bayesian Methods for IRT Models for Discrete and Continuous Responses

ABSTRACT: A comprehensive Bayesian estimation method using a Markov chain Monte Carlo (MCMC) computational method was developed that can be used to simultaneously estimate the parameters for models for discrete and continuous responses for a broad class of models. The method is illustrated with examples of the analysis of the grades from Central Examinations in Secondary Education in the Netherlands. A comparison between the grades from these examinations is complicated by the interaction between the students' pattern and level of proficiency on one hand and their choice of examination subjects on the other hand. Since this choice may cause a violation of the ignorability principle underlying most inferences in IRT, the model for the responses was enhanced with a model for the choice of the examination subjects. To illustrate the estimation procedure, estimates of both a model without and with this enhancement are presented. Finally, it will be shown how the proportion of variance in the grades explained by the students' schools and the effect of covariates (in this case Gender) can be estimated.

5.1. Introduction

Most applications of IRT models are to categorical data (Rasch, 1960; Samejima, 1969; Bock, 1972; Lord, 1980; Masters, 1982). However, situations may arise where the responses to the items are continuous. IRT models for continuous responses are outlined in Mellenbergh (1994), Moustaki (1996) and Skrondal and Rabe-Hesketh (2004). The present report focuses on Bayesian estimation methods for multidimensional IRT models for discrete and continuous responses simultaneously. A comprehensive estimation method using a Markov chain Monte Carlo (MCMC) computational method is developed that can simultaneously estimate the parameters for models for discrete (dichotomous and polytomous) responses and continuous responses for a broad class of models. Mostly, the method follows a proposal by Shi and Lee (1998, also see Béguin, & Glas, 2001), but we present several new features of the method as well. An analysis of the scaling of students' scores on a number of examination topics will be given as an example of the proposed methods.

Another problem studied in this chapter concerns the problem of missing data. Usually, it is assumed that missing responses (both this missing by design and randomly during the test administration process) do not depend on the latent variable to be measured. Procedures for analyzing data subject to this kind of missing mechanism were proposed by Lord (1974), who examined the imputation of partially correct item scores, and Bock (1972), who proposed treating omitted responses as an additional response category. However, it has also been shown that this type of missing responses can be ignored in the analysis (Bock & Aitkin, 1981). This is not the case if the missing responses result from a non-ignorable missing data mechanism. This type of data may emanate from low-ability respondents who fail to produce a response, as a result of discomfort or embarrassment, or simply because they have skipped items. Another example are missing responses due to time constraints. Bradlow and Thomas (1998) and Holman and Glas (2005) show that ignoring this kind of missing data process leads to bias in parameter estimates. Therefore, the model for the responses is enhanced with a selection model for the missing data indicators. In the present chapter, the application of such models will be illustrated with an example of the analysis of examination grades from central examinations in secondary education in the Netherlands. Since the comparison between the examination grades is expected to be complicated by the interaction between the students' pattern of proficiencies on the one hand and their choice of examination subjects on the other, the IRT model for

the grades is enhanced with a model for the choice of the examination subjects.

5.2. The Model

5.2.1. A Model for Continuous Responses

Let students be indexed $n = 1, ..., N$, and let items be indexed $k = 1, ..., K$. It is assumed that the observation z_{nk}, on student n and item k is normally distributed, that is

$$P(Z_{nk} = z_{nk} \mid \theta_n, \mathbf{a}_k, b_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left(-\frac{(z_{nk} - \eta_{nk})^2}{2\sigma_k^2}\right). \tag{5.1}$$

The expectation of the item response is a linear function of the explanatory latent variables, that is,

$$\begin{aligned} \eta_{nk} &= \sum_{q=1}^{Q} a_{kq}\theta_{nq} - b_k \\ &= \mathbf{a}_k'\theta_n - b_k, \end{aligned} \tag{5.2}$$

where $\mathbf{a}_k$ is a vector of the parameters $(a_{k1}, ..., a_{kq}, ..., a_{kQ})$ which are usually called factor loadings and b_k is a location parameter. Further, $\theta_n = \left(\theta_{n1}, ..., \theta_{nq}, ..., \theta_{nQ}\right)$ is the Q-dimensional proficiency parameter of student n. We assume that $\sigma_k^2 = 1$, for all k. That is, we assume that all the observed responses have the same scale.

5.2.2. Models for Discrete Responses

Graded Response Model

Samejima (1969) proposed a model for polytomously scored items, where the probability of a response in category j ($j = 1, ..., m$) of item k, is given by

$$P(Y_{nkj} = 1; \eta_{nkj}) = \begin{cases} 1 - \Phi(\eta_{nk1}) & \text{if } j = 0 \\ \Phi\left(\eta_{nkj}\right) - \Phi\left(\eta_{nk(j+1)}\right) & \text{if } 0 < j < m \\ \Phi(\eta_{nkm}) & \text{if } j = m\,, \end{cases} \tag{5.3}$$

where Φ denotes the standard normal cumulative distribution function, and $\eta_{nkj} = \alpha_k'\theta_n - \beta_{kj}$. To ensure that the probabilities $P(Y_{nkj} = 1; \eta_{nkj})$ are positive, the restriction $b_{k(j+1)} > b_{kj}$, for $0 < j < m$ is imposed.

A Model for a Singly Peaked Response Function

For dichotomously scored items, response probabilities that are monotonically increasing in θ may not always be appropriate. Examples are often encountered in attitude assessment. For instance, the question "Should public use of marijuana be fined?" may be disaffirmed by respondents with a liberal attitude towards drugs, but also by respondents with a strict attitude towards drugs, the latter because they take the view that a fine is far too lenient. Below, an application of singly peaked response functions in the framework of educational assessment will be presented. The model that will be considered is closely related to models by Andrich and Luo (1993, also see Andrich, 1997) and Verhelst and Verstralen (1993). Both models have a singly-peaked response probability, only the functional form of the probability is chosen differently. The model by Andrich and Luo (1993) has a hyperbolic cosine function probability function. The model by Verhelst and Verstralen (1993) is derived from the partial credit model with three response categories, where the highest and lowest categories are collapsed. This approach will also be used here, only the graded response model will be used as the starting point because this leads to a much simpler functional form. So, we assume that the probability of a positive response is equal to

$$P(Y_{nk} = 1; \eta_{nk}) = \Phi(\eta_{nk1}) - \Phi(\eta_{nk2}) \tag{5.4}$$

with $\eta_{nkj} = \mathbf{a}_k'\theta_n - b_{kj}$ $(j = 1, 2)$ and $b_{k1} < b_{k2}$ to guarantee that $P(Y_{nk} = 1; \eta_{nk})$ is positive. Note that

$$\begin{aligned} P(Y_{nk} = 0; \eta_{nk}) &= 1 - P(Y_{nk} = 1; \eta_{nk}) \\ &= 1 - \Phi(\eta_{nk1}) + \Phi(\eta_{nk2}). \end{aligned}$$

The model in (5.4) is related to a graded response model for responses that takes the values 0, 1 or 2, where the responses 0 and 2 are collapsed to $Y_{nk} = 0$. This conceptualization will also play a role in the estimation procedure for the model.

5.2.3. Higher-Level Models for Person Parameters

On a second level, it can be assumed that all first-level person parameters are i.i.d. samples from a multivariate normal distribution, that is,

$$\theta_n \sim N(\mu_{\mathcal{P}}, \Sigma_{\mathcal{P}}). \tag{5.5}$$

We may also assume that the students may be nested under some higher-level units. For instance, students may be nested in classes. The higher-level units will be indexed $p = 1, ..., P$. We imposed a two-level regression model on the latent variables θ_{npq}, that is,

$$\theta_{npq} = \sum_{s=1}^{S} \beta_{psq} x_{nps} + \varepsilon_{npq}$$

and

$$\beta_{psq} = \sum_{t=1}^{T} \gamma_{sqt} w_{ptsq} + \upsilon_{psq}.$$

It will be assumed that x_{np1} and w_{p1sq} are equal to one. The error terms have distributions

$$\varepsilon_{np} \sim N(0, \Sigma),$$

where Σ is a $Q \times Q$ covariance matrix and

$$\upsilon_p \sim N(\mathbf{0}, \mathbf{T}),$$

where $\mathbf{T}$ is a $SQ \times SQ$ covariance matrix. Both $\mathbf{T}$ and Σ are not restricted to be diagonal. An alternative formulation is that person parameters θ_{np} are predicted with a linear regression model, where $\mathbf{X}_{np}$ are observed covariates, β_p are the regression parameters in unit p, and $\Sigma_{\mathcal{P}}$ is the covariance-matrix of the residuals. Then the density of θ_{np} is given by

$$\theta_{np} \sim N(\mathbf{X}_{np}\beta_p, \Sigma_{\mathcal{P}}),$$

and the regression parameters are themselves also random variables with the regression model

$$\beta_p \sim N(\mathbf{W}_p\gamma, \mathbf{T}),$$

where $\mathbf{W}_p$ are observed covariates, γ, are regression parameters and $\mathbf{T}$ is the covariance-matrix of the residuals. The priors of all covariance matrices are non-informative inverse-Wishart distributions (see, for instance, Box & Tiao, 1973).

5.2.4. Combined IRT Models for the Responses and the Missing Data Indicator

In most analysis, it is assumed that the process causing the missing data can be ignored (see Rubin, 1976). However, if there are unobserved factors that influence

the realization of the missingness, ignorability does not hold and then the inferences made using an IRT model ignoring the missing data can be severely biased (Bradlow & Thomas, 1998; Holman & Glas, 2005). However, this bias can be removed when the model for the responses is enhanced with an IRT model that serves a selection model (see, for instance, O'Muircheartaigh, & Moustaki, 1999; Moustaki & O'Muircheartaigh, 2000; Moustaki & Knott, 1999; Holman & Glas, 2005).

The combination of the model for the responses **X** (which can either or both be discrete responses denoted by **Y** or continuous responses denoted by **Z**) and the missing data indicators **D** proceeds analogously to the approach by Holman and Glas (2005) adopted for modeling skipped items in a test. They consider two classes of models: the *MAR* and *NONMAR* models. Define a $N \times K$ matrix **D** of missing data indicators

$$d_{nk} = \begin{cases} 1 & \text{if a person } n \text{ responds to an item } k \\ 0 & \text{if otherwise.} \end{cases} . \tag{5.6}$$

Let $p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1,}\alpha_1,\beta_1)$ be some model for the observed response pattern $\mathbf{x}_n = (x_{n1}, ..., x_{nk}, ..., x_{nK})$, where $\theta_{n1,}$ is a latent proficiency parameter, and α_1 and β_1 are item parameters. Further, let $p(\mathbf{d}_n|\theta_{n0}, \alpha_0,\beta_0)$ be a model for the missing data indicator, for which we take one of the IRT models for dichotomous responses above, The model has latent person parameters θ_{n0} and item parameters α_0 and β_0. Finally, $g_0(\theta_{n0})$ and $g_1(\theta_{n1})$ are the prior densities of the latent person parameters. In the sequel, we assume these densities to be standard normal.

Then, the posterior of the person parameters of respondent n is proportional to

$$p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1,}\alpha,\beta)p(\mathbf{d}_n|\theta_{n0}, \alpha_0,\beta_0)g_0(\theta_{n0})g_1(\theta_{n1}). \tag{5.7}$$

In (5.7), the latent variables θ_{n1} for the observed data and θ_{n0} for the missing data process are independent, so the posterior factors into two independent components: one for $\mathbf{x}_n$ and one for $\mathbf{d}_n$. Hence we can ignore the model for the likelihood of the missing data $p(\mathbf{d}_n|\theta_{n0}, \alpha_0,\beta_0)g_0(\theta_{n0})$ and obtain estimates using

$$p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1,}\alpha_1,\beta_1)g_1(\theta_{n1}) \tag{5.8}$$

only. The model given by (5.8) will be called the *MAR* model.

A violation of ignorability is created if the latent variables for the observed data and the missing data indicators, θ_{n1} and θ_{n0} are dependent. Hence the name *NONMAR* model. In the sequel, it will be assumed that θ_{n1} and θ_{n0} have a multivariate normal distribution with a covariance matrix Σ. To identify the latent scale, the

mean of this distribution is set equal to zero. The posterior is proportional to

$$p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1,}\alpha_1, \beta_1)p(\mathbf{d}_n|\theta_{n0}, \alpha_0, \beta_0)g(\theta_{n0}, \theta_{n1}|\Sigma), \tag{5.9}$$

where $g(\theta_{n0}, \theta_{n1}|\Sigma)$ is the prior density of θ_{n0} and θ_{n1}. If the off-diagonal elements of Σ are non-zero, the complete model for $\mathbf{x}_n$ and $\mathbf{d}_n$ to obtain unbiased estimates of the parameters has to be considered. Using simulation studies, Holman and Glas (2005) showed that the bias in the estimates of the item parameters increases as a function of the correlation between θ_{n0} and θ_{n1}.

5.3. Bayesian Estimation

The procedure that will be presented is both an extension of the procedure for Bayesian MCMC estimation for factor analysis models with continuous and polytomous data by Shi and Lee (1998) and Béguin and Glas (2001) and of the Bayesian MCMC procedure for the multilevel IRT model presented by Fox and Glas (2001, 2002, 2003).

5.3.1. Prior Distributions

The conjugate prior distribution for $(\mu_\mathcal{P}, \Sigma_\mathcal{P})$ is a normal-inverse-Wishart distribution (see, for instance, Box & Tiao, 1973). With respect to the choice of $\kappa_0, \upsilon_0, \mu_0$ and Λ_o, a non-informative prior distribution is obtained if $\kappa_0 \to 0, \upsilon_0 \to -1$ and $|\Lambda_o| \to 0$. These parameter values result in the multivariate version of Jeffrey's prior density. The item parameters are collected in a vector ξ with sub-vectors ξ_k, $k = 1, ..., K$. A multivariate normal prior $\xi_k \sim N(\mu_\mathcal{I}, \Sigma_\mathcal{I})$ will be assumed.

5.3.2. Data Augmentation

A continuous response z_{nk} needs no data augmentation step. Discrete responses are mapped to a latent continuous response z_{nk} in a number of data augmentation steps. After this mapping, the MCMC algorithm does not distinguish between continuous observed and latent responses. First, if a response is not observed, that is, if $d_{nk} = 0$ for a combination of n and k, z_{nk} is randomly drawn from a normal distribution $\phi(z_{nk}; \eta_{nk}, 1)$, where $\phi(\,.\,; \eta_{nk}, 1)$ stands for the normal density with mean η_{nk} and standard deviation equal to one.

The Graded Response Model

The data augmentation scheme for the graded response model, developed by Johnson and Albert (1999), is a generalization of the scheme for dichotomous items proposed by Albert (1992). We first broaden the definition of the item parameters with $b_{k0} = -\infty$ and $b_{km} = \infty$, so we have $\eta_{nk0} = -\infty$ and $\eta_{km} = \infty$. Then simulation is based on the posterior

$$p(z_{nk}\,|y_{nk},\eta_{nk}) \propto \prod_{j=1}^{m} \phi(z_{nk};\eta_{nkj},1) y_{nkj} \left[\mathbf{I}(\eta_{nk(j-1)} < z_{nk} \leqslant \eta_{nkj})\right]. \tag{5.10}$$

Note that the factor $y_{nkj}\left[\mathbf{I}(\eta_{nk(j-1)} < z_{nk} \leqslant \eta_{nkj})\right]$ is positive only if $y_{nkj} = 1$ and $\eta_{nk(j-1)} < z_{nk} \leqslant \eta_{nkj}$.

Singly Peaked Response Model

Since the singly peaked model can be viewed as a collapsed version of the graded response model, we map the observed dichotomous response unto (0,1,2) using

$$\begin{aligned} P(U_{nk} = 0\,|Y_{nk} = 0,\eta_{nk},\gamma_k) &\propto 1 - \Phi(\eta_{nk1}) \\ P(U_{nk} = 1\,|Y_{nk} = 1,\eta_{nk},\gamma_k) &= 1 \\ P(U_{nk} = 2\,|Y_{nk} = 0,\eta_{nk},\gamma_k) &\propto \Phi(\eta_{nk2}). \end{aligned} \tag{5.11}$$

5.3.3. *Posterior Simulation*

The aim of the procedure is to simulate samples from the joint posterior distribution of the parameters and the augmented data given the observed continuous data $\mathbf{z}$ and the discrete data $\mathbf{y}$. This posterior is given by

$$\begin{aligned} p(\xi,\theta,\widetilde{\mathbf{z}},\mathbf{u},\mu,\Sigma\,|\mathbf{y},\mathbf{z}) \quad = \quad & p(\mathbf{z},\mathbf{u}\,|\mathbf{y}\,;\xi,\theta,)p(\theta \mid \mathbf{X}_{\mathcal{P}},\beta_{\mathcal{P}},\Sigma_{\mathcal{P}})p(\xi\,\big|\mathbf{X}_I,\beta_I,\Sigma_I\,) \\ & p(\beta\,|\mathbf{W}_{\mathcal{P}},\gamma,\mathbf{T},\Sigma_{\mathcal{P}}\,)p(\gamma|\mathbf{T})p(\Sigma_{\mathcal{P}})p(\mathbf{T}_{\mathcal{P}}) \end{aligned}$$

where $\mathbf{z}$ are the observed responses and $\widetilde{\mathbf{z}}$ are the augmented latent responses. Samples from this posterior distribution are generated using the Gibbs sampler (Gelfand & Smiths, 1990). The Gibbs sampler requires that the parameter vector is divided in

a number of components, and each successive component is sampled from its conditional distribution given sampled values for all other components. This sampling scheme is repeated until the sampled values form stable posterior distributions.

1. Draw $\mathbf{u}$ and $\mathbf{z}$ conditional on θ, ξ,and $\mathbf{y}$,
2. Draw θ conditional on $\mathbf{z}, \xi, \Sigma_{\mathcal{P}}$, $\mathbf{X}, \beta$,
3. Draw ξ conditional on $\mathbf{z}$ and $\theta, \Sigma_{\mathcal{I}}$, $\mu_{\mathcal{I}}, \mathbf{u}$ and $\mathbf{y}$,
4. Draw β_p conditional on $\theta_p, \Sigma, \mathbf{T}, \gamma, \mathbf{W}_p, \mathbf{X}_p$,
5. Draw Σ conditional on $\lambda, \mathbf{X}, \beta$
6. Draw γ conditional on $\mathbf{W}, \mathbf{T}, \beta$
7. Draw $\mathbf{T}$ conditional on $\mathbf{B}, \mathbf{W}, \gamma$.

The procedure thus amounts to iterative generation of parameter values using the above steps. The details of the steps are given in the Appendix. To evaluate the convergence of the procedure, multiple chains can be started from different points to evaluate convergence by comparing the between- and within-sequence variance. Another approach is to generate a single MCMC chain and to evaluate convergence by dividing the chain into subchains and comparing between- and within-subchain variance. For these and other technical details, see Gelman, Carlin, Stearn and Hall (1995).

5.4. An Empirical Example

5.4.1. The Data

The methods presented above are illustrated with an analysis of a data set of Dutch Central examinations. The data used in this study were collected by the Dutch Inspection of Education. The data are a subset of the data of approximately 18-year old students of pre-university schools that took their final examination in the school year 1994/1995. The students chose an examination package that consisted of 7 or 8 subjects. The analysis was restricted to 60 fairly common combinations of examination subjects. The resulting data set consisted of the examination results of 6142 students. The examination scores were on a scale of 0 to 10, with two significant digits after the decimal point.

5.4.2. Impact of the Selection Model

First, the differences between the model without and with the selection model were evaluated. For both models, the Gibbs sampler was run using 40,000 iterations. To ensure convergence, multiple MCMC chains from different starting points were generated and the between- and within-sequence variance were compared (see, for instance, Gelman, Carlin, Stearn & Hall, 1995). Because the number of augmentation variables z_{nk} ($n = 1, ..., 6142$, $k = 1, ..., 16$) proved to cause storage problems, for every iteration cycle of the MCMC algorithm a new sample of 2000 respondents was uniformly drawn from the available 6142 respondents. Normal priors were used for the *a* and *b* parameters. All prior means for non-fixed *a* parameters were set equal to one; the prior mean for the *b* parameters was set equal to the negative of the grand mean of all examination grades. The variances were set equal to 5.0, so the resulting prior was quite vague. The covariance matrix had a non-informative prior.

Table 5.1 gives the estimates obtained without the selection model. The point estimates reported are posterior expectations and posterior standard deviations. A number of factor loadings is fixed to zero; they are the same loadings as fixed in the previous chapters. The first two dimensions were identified by fixing the rows of the topics German and Mathematics. Therefore, they can be labeled as language and mathematics ability. The third dimension was identified as a dimension that related to Economy but Dutch Language and History also had significant loadings on this dimension. The estimated *b* parameters are shown in the last column. Since the mean of the ability distribution was scaled to zero, the opposite of the *b* parameters reflect the average grades on the topics. Note that the average score on Mathematics was lowest and the average score on Latin was highest. The estimated covariance matrix and the associated correlation matrix are given at the bottom of the table. Note that the Economics dimension correlated highly with the Mathematics dimension. The other two correlations were only moderate.

In the next analysis, a model for the choice variables d_{nk} was invoked. A latent variable θ_{Q+1} was assumed to govern the choice of the examination subjects, with the realizations of the choice variable defined by (5.4). If the students' proficiency level is highly correlated with the choice of examination subjects, then θ_{Q+1} will be highly correlated with $\theta_1, ..., \theta_Q$ also. The dependence between the latent variables is modeled by assuming the $\theta_1, ..., \theta_{Q+1}$ has a multivariate normal distribution. The correlations between θ_{Q+1} and the proficiency dimensions $\theta_1, ..., \theta_Q$ describe the extent to which the choice of an examination subject depends on the proficiency level. So if, for example, the correlation between θ_1 and θ_{Q+1} is positive, a high

Table 5.1: Bayesian estimates of the parameters of the factor model for the examination scores (Starred entries are fixed)

Topic	a_{k1}	a_{k2}	a_{k3}	b_k	$Se(a_{k1})$	$Se(a_{k2})$	$Se(a_{k3})$	$Se(b_k)$
Dutch	0.19	0.08	0.48	-6.21	.052	.099	.078	.020
Latin	0.24	-0.02	0.10	-7.33	.104	.122	.123	.059
Greek	0.17	0.03	0.19	-6.90	.099	.127	.103	.055
French	1.11	0.00*	0.00*	-6.83	.077			.039
German	1.00*	0.00*	0.00*	-6.41				.034
English	1.05	0.00*	0.00*	-6.43	.048			.021
History	0.44	0.18	0.58	-6.42	.058	.126	.128	.033
Geography	0.00*	1.10	0.00*	-6.23		.129		.035
Appl.Math	0.00*	1.00*	0.00*	-5.81				.022
Adv.Math	-0.09	1.23	0.10	-5.99	.111	.119	.128	.036
Physics	0.00*	1.26	0.00*	-6.11		.122		.033
Chemistry	0.00*	1.27	0.00*	-6.69		.109		.038
Biology	0.00*	1.13	0.00*	-6.51		.100		.035
Gen. Econ.	0.00*	0.00*	1.00*	-6.11				.032
Bus. Econ.	0.00*	0.33	0.83	-6.19		.098	.096	.019
Arts	0.16	-0.01	0.11	-6.61	.178	.199	.188	.101
Covariance Matrix					$Se(\sigma_{*1})$	$Se(\sigma_{*2})$	$Se(\sigma_{*3})$	
Language	0.806				.014			
Mathematics	0.329	0.589			.018	.011		
Economy	0.439	0.581	0.611		.018	.019	.012	
Correlation Matrix								
Language	1.000							
Mathematics	0.477	1.000						
Economy	0.626	0.968	1.000					

level on proficiency dimension θ_1 is positively related with subjects that load high on dimension θ_{Q+1}. Further, the magnitude of the correlations between $\theta_1, ..., \theta_Q$ and θ_{Q+1} gives an indication of the extent to which the assumption of ignorability is violated. If these correlations are close to zero, the choice behavior is not related to proficiency, and the missing data are ignorable. If, on the other hand, these correlations are substantial, the choice variable is highly related to the proficiencies for the students. Since the students can only chose a limited number of subjects, it is reasonable to assume that the probability of choosing a subject as a function of the proficiency dimension θ_{Q+1} is singly peaked: Students will probably chose subjects within a certain region of the proficiency dimension θ_{Q+1} and avoid subjects that are too difficult or too easy. The too difficult subjects are avoided because of the risk of failing the examination, and the too easy subjects are avoided because they do not

Table 5.2: Bayesian estimates of the parameters of the factor model for the examination scores enhanced with a selection model (Starred entries are fixed)

Topic	a_{k1}	a_{k2}	a_{k3}	b_k	$\bar{b}_k$	$se(a_{k1})$	$se(a_{k2})$	$se(a_{k3})$	$se(b_k)$	$se(\bar{b}_k)$
Dutch	0.22	0.14	0.58	-6.20	–	.051	.095	.085	.030	
Latin	0.25	0.00	0.09	-7.01	-0.77	.108	.123	.123	.060	.071
Greek	0.20	0.04	0.20	-6.89	-1.09	.103	.128	.103	.057	.080
French	1.22	0.00*	0.00*	-6.83	-0.77	.077			.039	.043
German	1.00*	0.00*	0.00*	-6.42	-0.62				.034	.031
English	1.08	0.00*	0.00*	-6.42	–	.042			.026	
History	0.55	0.18	0.57	-6.41	-0.19	.059	.127	.127	.033	.029
Geography	0.00*	1.11	0.00*	-6.22	0.11		.129		.037	.039
Appl.Math	0.00*	1.00*	0.00*	-5.82	0.01				.022	.029
Adv.Math	-0.09	1.27	0.11	-6.06	0.50	.118	.126	.129	.039	.042
Physics	0.00*	1.28	0.00*	-6.11	0.61		.122		.033	.039
Chemistry	0.00*	1.28	0.00*	-6.70	0.80		.132		.043	.045
Biology	0.00*	1.14	0.00*	-6.53	0.99		.111		.036	.039
Gen. Econ.	0.00*	0.00*	1.00*	-6.11	-0.31				.032	.034
Bus. Econ.	0.00*	0.35	0.85	-6.22	-0.13		.097	.097	.019	.029
Arts	0.16	0.01	0.09	-6.66	0.50	.178	.202	.199	.103	.099
Covariance Matrix						$se(\sigma_{*1})$	$se(\sigma_{*2})$	$se(\sigma_{*3})$	$se(\sigma_{*4})$	
Language	0.808					.015				
Mathematics	0.333	0.600				.018	.014			
Economy	0.445	0.584	0.619			.018	.019	.013		
Choice	0.128	0.759	0.565	1.101		.011	.011	.011	.009	
Correlation Matrix										
Language	1.000									
Mathematics	0.478	1.000								
Economy	0.629	0.958	1.000							
Choice	0.136	0.934	0.684	1.000						

contribute to a package suited for the desired level of university study. An IRT choice model that reflects this feature is given in (5.4).

The results of the analysis are displayed in Table 5.2. Note that the choice-dimension had significant positive correlations with all proficiency dimensions. The correlation with the Mathematics dimension was highest. For the choice dimension, displaying the factor loadings is not very informative, since they were all equal to one. Therefore, the average of the two subject parameters, that is, $\bar{b}_k = (b_{k1} + b_{k2})/2$ is displayed for all subjects in the last column labelled $\bar{b}_k$. The parameters $\bar{b}_k$ can be seen as estimates of the location of the subject on this fourth proficiency dimension. Note that the parameters for Dutch and English cannot be estimated, because these two examination subjects are obligatory, and so all the choice variables d_{nk} for these examination subjects are structurally equal to one and the parameters related to these

subjects cannot be estimated.

The interpretation of the mean parameters $\overline{b}_k$ is as follows. The choice dimension correlates positively with the three proficiency dimensions, and highest with the Mathematics dimension. This dimension can be viewed as an overall proficiency dimension, and the choice of subjects is assumed governed by this proficiency. Since the "difficulty parameters" $\overline{b}_k$ are estimates of the location of the subjects on the fourth proficiency dimension, they represent the ordering of the examination subjects on this dimension. That is, "difficult subjects" as Advanced Mathematics ($\overline{b}_k = 0.50$), Physics ($\overline{b}_k = 0.61$), Chemistry ($\overline{b}_k = 0.80$), and Biology ($\overline{b}_k = 0.99$) are chosen by the more proficient students.

5.4.3. Variance Attributable to Schools

The next research question tackled was how much of the variance in the latent person parameters is attributable to the schools. Therefore, the MCMC analysis of the previous report was redone with a two-level model (without covariates) for the ability parameters. That is, the overall covariance matrix was partitioned into a within schools covariance matrix $\Sigma_{\mathcal{P}}$ and a between schools covariance matrix $\mathbf{T}_{\mathcal{P}}$. The Gibbs sampler was run using 40,000 iterations. The results are shown in Table 5.3.

The point estimates reported are posterior expectations (EAP) and posterior standard deviations (PSD). Note that the choice-dimension has significant positive correlations with all proficiency dimensions. The correlation with the Science-dimension is highest. For the choice-dimension, displaying the factor loadings is little informative, since they are all equal to one. Therefore, the average of the two subject parameters, that is, $\overline{b}_k = (b_{k1} + b_{k2})/2$ are displayed for all subjects in the last column labelled $\overline{b}_k$. The parameters $\overline{b}_k$ can be seen as an estimate of the location of the subject on this fourth proficiency dimension. Note that the parameters for Dutch and English cannot be estimated, because these two examination subjects are obligatory and so all the choice variables d_{nk} for these examination subjects are structurally equal to one and the parameters related to these subjects cannot be estimated.

The within schools covariance matrix $\Sigma_{\mathcal{P}}$ and the between schools covariance matrix $\mathbf{T}_{\mathcal{P}}$, and the associated correlation matrices $\mathcal{R}(\Sigma_{\mathcal{P}})$ and $\mathcal{R}(\mathbf{T}_{\mathcal{P}})$ are given in the last panels of Table 5.3. The question regarding the proportion of variance in the latent person parameters attributable to the schools could, in principle, be addressed by using these variance estimates to compute intra class correlation coefficients (ICCs, see, for instance Bryk and Raudenbush, 1992). However, to obtain some measure of the credibility of the ICCs, it is more convenient to sample their values during the

Table 5.3: Bayesian estimates of parameters of examination topics (Starred entries are fixed)

Topic	Estimates					Standard errors				
	a_{k1}	a_{k2}	a_{k3}	b_k	$\bar{b}_k$	a_{k1}	a_{k2}	a_{k3}	b_k	$\bar{b}_k$
Dutch	0.20	0.16	0.60	-6.21	–	.050	.095	.099	.031	
Latin	0.24	0.00	0.09	-7.02	-0.77	.109	.123	.127	.060	.072
Greek	0.20	0.04	0.21	-6.89	-1.00	.101	.128	.111	.057	.080
French	1.20	0.00*	0.00*	-6.83	-0.76	.077			.039	.044
German	1.00*	0.00*	0.00*	-6.52	-0.69				.032	.029
English	1.08	0.00*	0.00*	-6.33	–	.042			.026	
History	0.45	0.12	0.60	-6.81	-0.22	.059	.127	.128	.033	.029
Geography	0.00*	1.09	0.00*	-6.23	0.12		.129		.037	.039
Appl.Math.	0.00*	1.00*	0.00*	-5.91	0.05				.022	.029
Adv.Math	-0.11	1.33	0.11	-6.06	0.55	.118	.126	.131	.039	.042
Physics	0.00*	1.38	0.00*	-6.14	0.63		.122		.033	.039
Chemistry	0.00*	1.36	0.00*	-6.71	0.79		.130		.042	.045
Biology	0.00*	1.29	0.00*	-6.53	0.98		.109		.036	.039
Gen. Econ.	0.00*	0.00*	1.00*	-6.22	-0.44				.032	.034
Bus. Econ.	0.00*	0.35	0.99	-6.19	-0.22		.097	.097	.019	.029
Arts	0.15	0.00	0.10	-6.66	0.47	.180	.209	.201	.103	.099
Covariance Matrix $\Sigma_{\mathcal{P}}$						σ_{*1}	σ_{*2}	σ_{*3}	σ_{*4}	
Language	0.711					.013				
Science	0.226	0.610				.018	.015			
Economy	0.297	0.456	0.530			.018	.018	.013		
Choice	0.026	0.634	0.425	1.044		.013	.013	.014	.011	
Covariance Matrix $\mathbf{T}_{\mathcal{P}}$						τ_{*1}	τ_{*2}	τ_{*3}	τ_{*4}	
Language	0.098					.025				
Science	0.008	0.006				.027	.034			
Economy	0.049	0.020	0.102			.018	.041	.036		
Choice	0.003	0.018	0.049	0.077		.040	.036	.038	.038	
Correlation Matrix $\mathcal{R}(\Sigma_{\mathcal{P}})$						Correlation Matrix $\mathcal{R}(\mathbf{T}_{\mathcal{P}})$				
Language	1.000					1.000				
Science	0.343	1.000				0.339	1.000			
Economy	0.484	0.803	1.000			0.496	0.816	1.000		
Choice	0.031	0.795	0.571	1.000		0.030	0.826	0.555	1.000	

MCMC procedure and to compute their EAPs and PSDs. The ICCs were computed as the variance ratio

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

where σ^2 and τ^2 are the appropriate diagonal elements from $\Sigma_{\mathcal{P}}$ and $\mathbf{T}_{\mathcal{P}}$, respectively. The posterior means and standard deviations of the six ICCs are shown in Table 5.4. Note that the ICCs for the Language and Economy dimensions exceed 10%. On the other hand, the ICC for the Science dimension is very close to zero.

Table 5.4: Bayesian estimates of intraclass correlations ρ

Topic	EAP	PSD
Language	.124	.010
Science	.008	.009
Economy	.159	.011
Choice	.071	.015

5.4.4. Variance Attributable to Gender

The second research question concerned the proportion of variance attributable to gender. The answer this question, a second analysis was carried out with gender as a predictor for each of the four ability dimensions. As above, the question was addressed for all dimensions. The proportion explained variance was computed as

$$\delta = \frac{\sigma^2_{\text{Model 0}} - \sigma^2_{\text{Model 1}}}{\sigma^2_{\text{Model 0}}},$$

where $\sigma^2_{\text{Model 0}}$ and $\sigma^2_{\text{Model 1}}$ are the EAPs of the appropriate diagonal elements of $\Sigma_{\mathcal{P}}$ obtained in the analysis without and with gender as a covariate, respectively. The results are shown in Table 5.3. Note that an estimate of the reliability of the indices is now lacking. The reason is that the indices are now computed from two separate analysis and not sampled in a single analysis. The computation of a measure for the reliability of the estimate of δ remains a point for further research. The results are displayed in the second column of Table 5.5.

Table 5.5: Bayesian estimates of gender effect β and proportion variance explained δ

Topic	EAP δ	EAP β	PSD β
Language	.081	.121	.014
Science	.071	-.007	.014
Economy	.006	.001	.014
Choice	.032	-.040	.013

Note that the proportions of variance explained by gender are highest for the Language and Science dimensions. Further, the EAP estimates and the PSDs of the regression coefficient for gender are displayed in the last two columns. Male gender

was coded zero, female gender was coded one. So a positive value of β reflects a higher ability level for the females, while a negative value is indicates the opposite. Note the average for the Language dimension is higher for the females, while their average on the overall ability dimension (the choice dimension) is lower.

5.5. Discussion

The problems addressed in this article were related to Bayesian estimation methods for multidimensional IRT models for combined discrete and continuous data. The methods are illustrated with an analysis of a data set of Dutch Central Examinations. The interaction between the students' pattern and level of proficiency on one hand and their choice of examination subjects on the other hand complicates the comparison between the grades obtained on these examinations. The choice of subjects causes a violation of the ignorability principle for missing data underlying most inferences in IRT. The multidimensional IRT model for the responses was therefore expanded with a model for the choice of the examination subjects.

A three-factor model and a four factor model where the fourth factor pertained to a choice model were evaluated and compared. These models produced very similar results with respect to the factor structure, the difficulty parameters of the subjects and the covariance matrix of the proficiency dimensions. Only the difficulty parameter of the subject Latin went down substantially. The choice-dimension had significant positive correlations with all proficiency dimensions. So the choice of the subjects by the students is clearly related to their proficiency level. This also means that the choice-dimension can be viewed as an overall proficiency dimension. The correlation of the choice dimension with the Science dimension was highest. So this overall proficiency dimensions depends mostly on the Science-dimension.

MCMC analysis of two-level factor models (with and without covariates) were presented to show how to evaluate the amount of variance in the latent person parameters attributable to the schools. This proportion was evaluated by intra class correlation coefficients (ICC). The ICCs for the Language and Economy dimensions were significant and exceeded 0.10. Also the ICC for the choice-dimension was significant. However, the ICC for the Science dimension is very close to zero. The conclusion is that differences between schools are important for Language and Economy subjects, but not for Science.

The analysis with gender as a predictor for each of the four latent dimensions was carried out to estimate the effect of Gender. The effect of Gender was highest for the

Language and Science dimensions. The Gender effect for the Language dimension is positive for the females. This is in accordance with the common opinion that girls are better in language. On the other hand, the Gender effect for females was negative for the choice-dimension. This implies on one hand that girls choose slightly easier subjects, and that the overall proficiency of the girls is slightly lower.

Overall, serious implications of violation of ignorability for comparing the difficulty of subjects did not occur. However the fact that the model for proficiency should at least be multidimensional implies that one should be cautious when publishing school performance results to provide parents and their children with information for their school choice.

Appendix

5.A. The MCMC Algorithm in Detail

Step 1

This step is the data augmentation step which boils down to sampling from the distributions (5.10) and (5.11).

Step 2

To draw from the conditional distribution of θ, an orthogonally standardized ability variable θ^o is defined. So the elements of $\theta_n^o = (\theta_{n1}^o, ...\theta_{nQ}^o)'$ have independent standard normal distributions. Let $\mathbf{L}$ be the Cholesky decomposition of $\Sigma_{\mathcal{P}}$, that is, $\Sigma_{\mathcal{P}} = \mathbf{L}\mathbf{L}'$. Define $\theta_n^o = \mathbf{L}^{-1}(\theta_n - \mu)$. Now η_{nk} can be written as

$$\begin{aligned}\eta_{nk} &= \sum_{q=1}^{Q}(a_{kq}\theta_{nq}) - b_k \\ &= \sum_{q=1}^{Q}(a_{kq}\sum_{h=1}^{Q} L_{hq}\theta_{nq}^o + \mu_q) - b_k,\end{aligned}$$

or, in matrix notation,

$$\eta_n = \mathbf{A}\mathbf{L}\mathbf{L}^{-1}(\theta_n - \mu + \mu) - \mathbf{b} = \mathbf{A}(\mathbf{L}\theta_n^o + \mu) - \mathbf{b},$$

with η_n and $\mathbf{b}$ vectors of length k, θ_n, θ_n^o and μ vectors of length Q and $\mathbf{A}$ a $k \times Q$ matrix with entries a_{kq}. The ability parameters θ_n^o have a posterior density given by

$$p(\theta_n \,|\xi, \mathbf{z}, \mathbf{y}, \mathbf{w}) \propto \phi(\theta_n^o; \mathbf{0}, \mathbf{I}) \prod_{k=1}^{k} \phi(z_{nk}; \eta_{nk}, 1).$$

This entails that $\mathbf{z}_n + \mathbf{b} - \mathbf{A}\mu = \mathbf{B}\theta_n^o + \varepsilon_n$, where $\mathbf{B} = \mathbf{A}\mathbf{L}$ and ε_n is a vector of error terms ε_{nk}, which are i.i.d. $N(0, 1)$. It then follows that

$$\theta_n^o \sim N\left((\mathbf{I} + \Sigma^{-1})^{-1}\Sigma^{-1}\hat{\theta}_n^o, \quad (\mathbf{I} + \Sigma^{-1})^{-1}\right),$$

with $\hat{\theta}_n^o$ the common least squares estimate $\hat{\theta}_n^o = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'(\mathbf{z}_n + \mathbf{b} - \mathbf{A}\mu)$ and $\Sigma = (\mathbf{B}'\mathbf{B})^{-1}$. Now θ_n can be obtained by the transformation $\mathbf{L}\theta_n^o + \mu = \theta_n$.

Step 3a: Sample a and/or b

Draw the item parameters ξ_k= ($\mathbf{a}_k$,b_k), or, in case of the graded response model, ξ_k= ($\mathbf{a}_k$). In the latter case, the b parameters are drawn in Step 3c. Consider a multivariate normal prior for the item parameters ξ_k with mean, $\mu_{\xi 0} = (\mu_{a1}, ...\mu_{aQ}, \mu_b)'$ and variance, $\Sigma_{\xi 0}$. Define $\mathbf{X}$ as a $n \times (Q + 1)$ matrix with rows $(\theta_{n1}, ..., \theta_{nq}, ..., \theta_{nQ}, -1)$. Conditional on θ, $\mathbf{z}_k = (z_{1k}, ..., z_{nk})'$ satisfies the linear model

$$\mathbf{z}_k = \mathbf{X}\xi_k + \varepsilon_k,$$

where $\varepsilon_k = (\varepsilon_{1k}, ..., \varepsilon_{nk}, ..., \varepsilon_{nk})'$ and the ε_{nk} are i.i.d. $N(0, 1)$. The likelihood function of ξ is of normal form with mean $\widehat{\xi_k} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}_k$ and variance $v = (\mathbf{X}'\mathbf{X})^{-1}$. Combining this with the normal distributed priors, one obtains

$$\xi_k \,|\theta, \mathbf{z}, \mathbf{y} \;\sim\; N(\mu_{\xi_k}, (\Sigma_{\xi 0}^{-1} + \mathbf{X}'\mathbf{X})^{-1}),$$

where $\mu_{\xi_k} = (\Sigma_{\xi 0}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\Sigma_{\xi 0}^{-1}\mu_{\xi 0} + \mathbf{X}'\mathbf{z}_k)$.

Step 3b: Modification for Fixed a Parameters

Define the $n \times (Q + 1)$ matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. $\mathbf{X}_1$ has entries equal to the analogous entries of $\mathbf{X}$, except for the columns associated with the fixed item parameters. The latter entries are equal to zero. In the same manner, $\mathbf{X}_2$ has entries equal to the corresponding entries of $\mathbf{X}$, except for the columns associated with the free item parameters. Also, here, the latter are equal to zero. Note that $\mathbf{X} = \mathbf{X}_1$+$\mathbf{X}_2$. Conditional on θ, $\mathbf{z}_k = (z_{1k}, ..., z_{nk})'$ satisfies the linear model

$$\mathbf{z}_k^* = \mathbf{z}_k - \mathbf{X}_1\xi_k = \mathbf{X}_2\xi_k + \varepsilon_k,$$

where $\varepsilon_k = (\varepsilon_{1k}, ..., \varepsilon_{nk}, ..., \varepsilon_{nk})'$ and the ε_{nk} are i.i.d. $N(0, 1)$. Combining this with the normal distributed priors, one obtains

$$\xi_k \,|\theta, \mathbf{z}, \mathbf{y} \;\sim\; N(\mu_{\xi_k}, (\Sigma_{\xi 0}^{-1} + \mathbf{X}_2'\mathbf{X}_2)^{-1}),$$

where $\mu_{\xi_k} = (\Sigma_{\xi 0}^{-1} + \mathbf{X}_2'\mathbf{X}_2)^{-1}(\Sigma_{\xi 0}^{-1}\mu_{\xi 0} + \mathbf{X}_2^T\mathbf{z}_k^*)$. Note that the prior covariance $\Sigma_{\xi 0}^{-1}$ assures that $\Sigma_{\xi 0}^{-1} + \mathbf{X}_2'\mathbf{X}_2$ is invertible.

Step 3c: Graded Response Model

For the graded response model, the b parameters are drawn using a hybrid Metropolis-Hastings sampler outlined by Johnson and Albert (1999).

- Set $\sigma_{MH} = 0.05/m$. This value can be adjusted if the acceptance rate is too low.
- For $j = 1, ..., m-1$, sample candidates δ_{kj} from $N(b_{kj}, \sigma^2_{MH})$ truncated to the interval $(\delta_{k(j-1)}, b_{k(j+1)})$.
- Compute the acceptance ratio

$$R = \prod_n \prod_j \left(\frac{\Phi\left(\eta_{nkj}\right) - \Phi\left(\eta_{nk(j+1)}\right)}{\Phi\left(\mathbf{a}'_k\theta_n - \delta_{kj}\right) - \Phi\left(\mathbf{a}'_k\theta_n - \delta_{k(j-1)}\right)} \right)^{y_{nkj}} \otimes \prod_j \frac{\Phi\left((b_{k(j+1)} - b_{kj})/\sigma_{MH}\right) - \Phi\left((b_{k(j-1)} - b_{kj})/\sigma_{MH}\right)}{\Phi\left((b_{k(j+1)} - b_{kj})/\sigma_{MH}\right) - \Phi\left((b_{k(j-1)} - b_{kj})/\sigma_{MH}\right)}.$$

- Set $b = \delta$ with probability R, otherwise, keep the previous draw of b.

Step 4

Sampling of β_p, $p = 1, ..., P$. Define

- β_p as a SQ-dimensional vector of the elements β_{spq},
- θ_p as a N_pQ-dimensional vector of the elements θ_{npq}, where N_p is the number of observations in the sample from the p-th population,
- γ as a SQT-dimensional vector of the elements γ_{sqt},
- $\mathbf{W}_{psq}$ as a T-dimensional vector of the elements w_{ptsp} and $\mathbf{W}_p = \{\mathbf{W}_{psq}\} \otimes \mathbf{I}_{SQ}$. Note that $\mathbf{W}_p$ is a $SQ \times SQT$ matrix,
- $\mathbf{X}^*_p = \{\mathbf{X}_p\} \otimes \mathbf{I}_Q$, with $\mathbf{X}_p$ a matrix of the elements $\{\mathbf{x}_{nps}\}$. Note that $\mathbf{X}_p$ is a $N_pQ \times SQ$ matrix.

Given all other parameters, the conditional distribution of β_j is normal, that is

$$\beta_p \mid \theta_p, \Sigma, \mathbf{T}, \gamma, \mathbf{W}_p, \mathbf{X}_p \sim N\Big(\quad \Phi\left[\mathbf{X}^{*\prime}_p\theta_p + \mathbf{T}^{-1}\mathbf{W}_p\gamma\right] \quad , \quad \Phi \quad \Big),$$

with $\Phi = (\mathbf{X}^{*\prime}_p\mathbf{X}^*_p + \mathbf{T}^{-1})^{-1}$.

Step 5

Sampling of Σ. Define the matrix of residuals $\mathbf{S} = \sum_p (\theta_{np} - \mathbf{B}_p \mathbf{X}_{np})(\theta_{np} - \mathbf{B}_p \mathbf{X}_{np})'$, with

θ_{np} a Q-vector of the elements θ_{npq},

$\mathbf{X}_{pq}$ a S-dimensional vector of the elements x_{nps},

$\mathbf{B}_p$ a $Q \times S$ matrix of the elements β_{psq}.

Then the conditional distribution of Σ is Inverse-Wishart:

$$\Sigma \mid \lambda, \mathbf{X}, \beta \sim \text{Inv-W}(N, \mathbf{S}^{-1}).$$

Step 6

Sampling of γ.

$$\gamma \mid \mathbf{W}, \mathbf{T}, \beta \sim N\left(\quad \Psi \sum_p \mathbf{W}_p' \mathbf{T}^{-1} \beta_p \quad , \quad \Psi \quad \right)$$

where $\Psi = \sum_p \mathbf{W}_p' \mathbf{T}^{-1} \mathbf{W}_p{}^{-1}$.

Step 7

Sampling of $\mathbf{T}$. Define the matrix of residuals $\mathbf{S} = \frac{1}{P} \sum_p (\beta_p - \mathbf{W}_p \gamma)(\beta_p - \mathbf{W}_p \gamma)'$. Then the conditional distribution of $\mathbf{T}$ is Inverse-Wishart

$$\mathbf{T} \mid \mathbf{B}, \mathbf{W}, \gamma \sim \text{Inv-W}(P, \mathbf{S}^{-1}).$$

Summary

In item response theory (IRT), mathematical models are applied to analyze data from tests and questionnaires used to measure abilities, proficiency, personality traits and attitudes. This thesis is concerned with comparison of subjects, students and schools based on average examination grades using IRT. The difficulty of such comparisons is caused by student's free choice of examination subjects. That means, students only sit examinations in subjects they have chosen themselves. However, if the students have different examination packages, their grades are probably not comparable. The main problem with using examination grades, or grade point averages (GPAs) is the incorrect assumption that all course grades mean essentially the same thing. However, there is always substantial variation among topics, courses, teachers, instructors and grading standards, so GPAs are not automatically comparable.

In Chapter 2 the standardization over subjects was achieved by using both the well known Kelly method (Kelly, 1976) and IRT models. Grades may either be represented as continuous data or as discrete data. Both representations were used and compared. IRT models are mathematical functions that specify the probability of a response of a student to an item in terms of person and item parameters. Therefore, using IRT models with grades as observations allows for separating the effect of the level of the students and the difficulty of the examination subjects. The generalized partial credit model (Muraki, 1992) was used to scale the discrete categorical data. A special IRT model for continuous responses (Mellenbergh,1994) was used for the continuous case. In general this model is equivalent with a factor analysis model. First, a unidimensional representation for proficiency was used. The results obtained using unidimensional IRT models were compared with the results obtained using Kelly's algorithm. Kelly's method and unidimensional IRT methods showed very similar results, both for continuous and discrete grades. The correlation of the rank

order of the estimates of the difficulty of the examination subjects was very high. However, it is not a-priori plausible that the proficiency structure assessed by the examinations is unidimensional. Therefore, as an alternative three-dimensional IRT models with a simple structure where each subject loads on one dimension only, were considered. The results of the three factor models for categorical and continuous grades were again very similar. Finally, a Multilevel model was used to estimate the variance in grades attributable to the schools. The overall conclusion was that the impact of the schools on the outcomes was not very large.

In Chapter 3, we return to the discrete response format and apply the uni- and multi-dimensional generalized partial credit further. For the analysis using the unidimensional model, the sample of students was partitioned into a group with a language-oriented package, a group with a science-oriented package and a group of the other students. Using this partition, it could be shown that the results in the groups were highly implausible. For instance, the implausible result of the high expected grades in Mathematics and Science for the language oriented students. Therefore, it was concluded that the unidimensional model did not fit the data. Next, a model was considered with a multidimensional representation of the proficiency of the students, where, contrary to the previous chapter, each subject could load on more than one dimension. The three-dimensional IRT model had a substantially better fit than the unidimensional IRT model and the implausible result of the high expected grades in Mathematics and Science for the language oriented students vanished. However, also this model was not accepted unconditionally, because it is not a-priori plausible that the examination subjects that were not chosen (the missing data) are missing at random. It was considered that the proficiency of students affected the pattern of the missing data in such a way that this might bias the estimates of the parameters of the examination subjects. To remove this bias, a four-dimensional IRT model was introduced, where the first three dimensions are related to the observed grades, while the fourth dimension is related to the choice variables. This model fitted the data significantly better than the three-dimensional model. Still, the expected grades computed using the two models were very close.

A testing procedure for IRT models for continuous responses (Mellenbergh, 1994; Moustaki, 1996; Skrondal and Rabe-Hesketh, 2004) was developed in the Chapter 4. A method for testing model fit was proposed in the framework of marginal maximum likelihood estimation. The fit to the model is evaluated using the Lagrange multiplier tests. The tests are based on residuals, that is, differences between observed and expected mean scores, that support an appraisal of the seriousness of the model

violation. The tests focus on the assumed form of the response functions, differential item functioning, local stochastic independence and the factor structure underlying the responses. The tests are illustrated with an example of the analysis of data from Central Examinations in Secondary Education, but also the simulation studies for the Type I error rate and power were presented.

In all these previous chapters, the studies have been done in the framework of marginal maximum likelihood. As an alternative, a Bayesian framework is considered in Chapter 5. A Bayesian estimation method using a Markov chain Monte Carlo method was developed that can be used to estimate the parameters for models for combined discrete and continuous data. To illustrate these methods, again the data set of Dutch Central examinations was used. The Multidimensional IRT models with and without a selection model for the choice variables were presented, evaluated and compared. These models produced results that were very similar to the results obtained in previous chapters. This is, the choice-dimension had significant positive correlations with all proficiency dimensions, and highest with Mathematics dimension. So the choice of subjects is governed by an overall proficiency dimension. Next, an MCMC analysis for two-level models (with and without covariates) for the proficiency parameters was performed to show how to estimate the amount of variance in the latent person parameters is attributable to the schools. That was done using the intra-class correlation (ICC) coefficients. The results of this analysis show that ICCs for the Language and Economy dimensions were significant and exceed 0.10. So the differences between schools are more important for the Language and Economy subjects. As another example of the use of the model, an analysis with Gender as a predictor for each of the four ability dimensions was carried out to estimate the proportion of variance attributable to Gender. The impact of Gender was the highest for the Language and Science dimensions. The effect on the Language dimension was positive for the females, while the effect of female Gender on the choice dimension was negative.

The final conclusions are the following: (1) proficiency on examinations is multidimensional. The implication is that one should be rather cautious when publishing school performance results to provide parents and their children with information for their school choice and when defining variables for school-effectiveness research. (2) Tools were developed for detecting violations of ignorability. Though no serious violations were found in the present study, the tools developed here may prove their importance in future research in educational science and social science in general.

Samenvatting

Item response theorie (IRT) modellen zijn statistische modellen die worden toegepast voor het analyseren van data van toetsen en vragenlijsten die gebruikt worden voor het meten van vaardigheden, persoonlijkheidskenmerken en attituden. Dit proefschrift houdt zich bezig met het vergelijken van examenvakken, leerlingen en scholen met behulp van IRT. Het probleem bij die vergelijking is dat leerlingen hun examenvakken vrij kunnen kiezen. Leerlingen doen dus alleen examen in vakken die ze zelf gekozen hebben. Hun gemiddelde examenscores zijn waarschijnlijk niet vergelijkbaar omdat ze een verschillend vakkenpakket hebben. Het belangrijkste probleem bij zo'n vergelijking van gemiddelde scores is de onjuiste veronderstelling dat alle examenvakken even moeilijk zijn. In zijn algemeenheid is er echter altijd een substantiele variatie in de moeilijkheid van onderwerpen en vakken, en de strengheid van beoordelingen en normering waardoor gemiddelden niet onder meer vergelijkbaar zijn.

Voor het evalueren van methoden voor de vergelijking van scores werden in alle hoofdstukken van dit proefschrift dezelfde data gebruikt. De data waren van het Centraal Schriftelijk examen voor het VWO uit 1995.

In Hoofdstuk 2 wordt een standaardisatie van scores beschreven met de bekende methode van Kelly (Kelly, 1976) en met behulp van IRT modellen. Beide benaderingen werden vergeleken. De scores kunnen opgevat worden als scores op een continue schaal en als discrete scores. IRT modellen zijn wiskundige functies die de kans op een score van een leerling op een vak formuleren als een functie van persoonsparameters en itemparameters (c.q. parameters geassocieerd met de vakken). Daardoor zijn de effecten van de moeilijkheid van de vakken en de vaardigheid van de leerlingen op de geobserveerde scores te scheiden. Het gegeneraliseerde partial credit model (Muraki, 1992) werd gebruikt voor het analyseren van de discrete scores. Een IRT

model dat equivalent is aan een factor analyse model (Mellenbergh, 1994) werd gebruikt voor het analyseren van de continue scores. Als eerste werd een model met een een-dimensioneel vaardigheidscontinuum gebruikt. De resultaten van deze een-dimensionele IRT analyses werden vergeleken met de resultaten van de methode van Kelly. De resultaten van beide benaderingen waren vergelijkbaar, zowel voor de discrete als voor de continue interpretatie van de scores. De correlatie tussen de volgordes van de schattingen van de moeilijkheidsgraad van de examenvakken was zeer hoog. Het is echter niet a-priori voor de hand liggend dat de vaardigheden die nodig zijn voor de verschillende examenvakken een-dimensioneel zijn. Daarom werd ook een drie-dimensioneel IRT model gebruikt, waarbij drie clusters van vakken betrekking hadden op drie specifieke vaardigheidsdimensies. Opnieuw waren de resultaten voor zowel de discrete als voor de continue interpretatie van de scores vergelijkbaar.

In Hoofdstuk 3 houden we ons alleen bezig met de discrete interpretatie van de scores en passen we opnieuw het een-dimensonele en het multi-dimensionele gegeneraliseerde partial credit model toe. Voor de analyses met het uni-dimensionele model werd de steekproef van examendata opgesplitst in leerlingen met een taalgeorienteerd pakket, leerlingen met een op wis- en natuurkunde georienteerd pakket en een groep van de overige leerlingen. Door gebruik te maken van deze opsplitsing kon worden aangetoond dat de resultaten van het uni-dimensionele IRT model bijzonder onwaarschijnlijk waren. Als voorbeeld noemen we de hoge schatting van de scores op wiskunde en natuurkunde voor de leerlingen met een taalgeorienteerd pakket. Daarom werd geconcludeerd dat een een-dimensioneel model niet bij de data paste. Daarna werd een multidimensioneel IRT model geschat. In tegenstelling tot het vorige hoofdstuk werd hier een model gebruikt waarbij vakken op meerdere dimensies betrekking konden hebben. Dit drie-dimensionele IRT model paste veel beter bij de data en de onwaarschijnlijke resultaten zoals de hoge schatting van de scores op wiskunde en natuurkunde voor de leerlingen met een taalgeorienteerd pakket verdwenen. Dit model kan echter ook niet zonder meer geaccepteerd worden, omdat het niet a-priori vaststaat dat de scores op de niet-gekozen vakken, die opgevat kunnen worden als ontbrekende gegevens, "missing-at-random" zijn. Daarom werd de veronderstelling geanalyseerd dat de vaardigheid van de leerlingen het keuzeproces op een zodanige manier beinvloedde dat de schatting van de moeilijkheidsgraad van de examens onzuiver was. Om deze onzuiverheid te corrigeren werd een vier-dimensioneel IRT model gebruikt, waarbij de eerste drie dimensies vaardigheidsdimensies waren en waarbij de vierde dimensie het keuzeproces representeerde. Dit model paste

significant beter bij de data dan het drie-dimensionele model. De schattingen van de verwachte scores onder beide modellen waren echter bijna gelijk.

In Hoofdstuk 4 werd een methode voor de evaluatie van modelpassing voor het IRT model voor continue responsies (Mellenbergh, 1994; Moustaki, 1996; Skrondal and Rabe-Hesketh, 2004) ontwikkeld. De methode werd ontwikkeld in de context van marginale grootste-aannemelijkheidsschatting. De modelpassing werd geevalueerd met Lagrange multiplier toetsen. De toetsen zijn gebaseerd op residuen, i.e., de verschillen tussen geobserveerde en verwachte gemiddelde scores. Deze verschillen maken het mogelijk de ernst van een modelovertreding te beoordelen. De toetsen zijn gericht op de veronderstelde vorm van de response functie, niet gemodelleerde systematische verschillen in de scores tussen groepen (Engels: differential item functioning), de veronderstelling van locale stochastische onafhankelijkheid, en de structuur van de multidimensionele vaardigheden, c.q. de factorstructuur. De methoden worden geillustreerd met een analyse van de VWO examendata, maar ook met simulatiestudies naar het onderscheidend vermogen en de kans op Type I fouten van de toetsen.

Alle analyses in de tot nu toe beschreven hoofdstukken werden uitgevoerd met marginale grootste-aannemelijkheids-schattingsmethoden. Het alternatief van Bayesiaanse statistiek wordt onderzocht in Hoofdstuk 5. Er wordt een Bayesiaanse schattingsmethode ontwikkeld die gebaseerd is op een Markov chain Monte Carlo (MCMC) rekenmethode. De methode kan gebruikt worden voor een combinatie van discrete en continue scores. De methode wordt opnieuw geillustreerd met een analyse van de VWO examendata. Multidimensionale modellen zonder en met een additioneel selectiemodel voor de keuzevariabelen werden vergeleken. De resultaten waren vergelijkbaar met de resultaten in de vorige hoofdstukken. Opnieuw had de keuzedimensie significante positieve correlaties met de vaardigheidsdimensie, en de correlatie was het hoogst met de vaardigheidsdimensie die gerelateerd was aan wiskunde en natuurkunde. Daaruit kan geconcludeerd worden dat de keuzen gestuurd worden door de vaardigheid van de leerlingen. Hierna werd een MCMC analyse met een tweeniveau model uitgevoerd om de hoeveelheid variantie in de vaardigheidsparameters, die toe te schrijven was aan de scholen waarin de leerlingen zaten te schatten. De proportie variantie van de scholen werd uitgedrukt in een intra-klasse correlatie coefficient (ICC). De resultaten lieten zien dat de ICC voor de taaldimensie en voor de economiedimensie significant van nul verschilden; ze waren groter dan 0.10. Een tweede voorbeeld van toepassing van de methode was een schatting van het effect van het geslacht van de leerlingen op de vier vaardigheidsdimensies. Het effect van

geslacht was het hoogste voor de taaldimensie en voor de dimensie gerelateerd aan wiskunde en natuurkunde. Het effect op de taaldimensie was positief voor de meisjes en het effect op de dimensie gerelateerd aan wiskunde en natuurkunde was positief voor de jongens.

De eindconclusies van dit onderzoek zijn de volgende. (1) De vaardigheidsstructuur waarop de examens een beroep doen is multidimensioneel. De implicatie hiervan is dat men erg terughoudend moet zijn met het publiceren van gemiddelde schoolresultaten ter informatie van scholen, leerlingen en ouders en met het definieren van variabelen in schooleffectiviteitsonderzoek. (2) In dit proefschrift werden statistische methoden ontwikkeld voor het evalueren van schending van de aanname dat onvolledige data niet tot vertekeningen in analyses leiden. Hoewel in de huidige studie geen ernstige schendingen werden gevonden, kunnen de modellen, die voorgesteld zijn, in de toekomst ongetwijfeld hun nut hebben voor onderwijskundig onderzoek en sociaal wetenschappelijk onderzoek in zijn algemeenheid.

Bibliography

Albert, J.H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

Aitchison, J., Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restrains. *Annals of Mathematical Statistics, 29,* 813 – 828.

Andrich, D. (1997). A hyperbolic cosine IRT model. In W.J.van der Linden and R.K.Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 399-414). New York, NJ: Springer.

Andrich, D., & Luo, G.Z. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17,* 253-276.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541-562.

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

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

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm.. *Psychometrika, 46,* 443-459.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 433-448). New York, NJ: Springer.

Box, G. and Tiao, G. (1973). *Bayesian inference in statistical analysis*. Addison-

Wesley Publishing Company, Reading, Massachusetts.

Bradlow, E.T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics, 23,* 236-243.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical Linear Models.* Newbury Park/London/New Delphi: Sage Publications.

Caulkins, J.P., Larkey, P.D., and Wei, J. (1996). Adjusting GPA to Reflect Course Difficulty. *The Heinz School of Public Policy and management, Carnegie Mellon University.*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ., Erlbaum.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B. 39,* 1-38.

Elliot, R., & Strena, A.Ch. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement, 25,* 333-347.

Fitz-Gibbon, C.T. (1994). Monitoring education: indicators, quality and effectiveness. *London: Cassell.*

Fox, J.-P., & Glas, C.A.W. (2001). Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling. *Psychometrika 66,* 271-288.

Fox, J.-P., & Glas, C.A.W. (2002). Modelling measurement error in structural multilevel models. In G.A. Marcoulides and I. Moustaki (Eds.). *Latent Variable and Latent Structure models.* (pp. 245-269). Mahwah, NJ: Laurence Erlbaum.

Fox, J.-P. and C.A.W. Glas (2003), Bayesian modeling of measurement error in predictor variables using Item Response Theory, *Psychometrika* **68**, 169–191.

Fox, J.-P., & Glas, C.A.W. (2005). Bayesian modification indices for IRT models. *Statistica Neerlandica, 59,* 95-106.

Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85,* 398-409.

Gelman, A, Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of proficiency. In M. Wilson, (Ed.), *Objective measurement: Theory into practice, Vol. 1,*

(pp.236-258), New Jersey, NJ: Ablex Publishing Corporation.

Glas, C.A.W. (1998) Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8, vol. 1.* 647-667.

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64,* 273-294.

Glas, C.A.W. & Meijer, R.R. (2003). A Bayesian Approach to Person Fit Analysis in Item Response Theory Models. *Applied Psychological Measurement, 27,* 217-233.

Glas, C.A.W., & Suarez-Falcon, J.C. (2003). A Comparison of Item-Fit Statistics for the Three-Parameter Logistic Model. *Applied Psychological Measurement, 27,* 87-106.

Goldstein, H. (1995). Multilevel statistical models. *London: Edward Arnold.*

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrika, 46,* 153-161.

Hoijtink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory* (pp.109-130). New York, NJ: Springer.

Holman, R. & Glas, C.A.W. (2005). Modelling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58,* 1-17.

Johnson, V.E. (1997). An Alternative to Traditional GPA for Evaluating Student Performance. *Statistical Science, 12(4),* 251-278.

Johnson, V.E., & Albert, J.H. (1999). *Ordinal data modelling.* New York, NJ: Springer.

Johnson, V.E. (2003). *Grade Inflation: A Crisis in College Education*. New York, NJ: Springer.

Kelly, A. (1976). A study on the comparability of external examinations in different subjects. *Research in Education, 1,* 37-63.

Kolen, M.J., & Brennan, R.L. (1995). *Test Equating.* New York, NJ: Springer.

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

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226-233.

Linn, R.L. (1966). Grade adjustments for prediction of academic performance: A review. *Journal of Educational Measurement, 3(4), 313-329.*

Maris, G. (2005). Posterior predictive p-values for classical null hypotheses. *Statistica Neerlandica, 59,* 70-81.

Masters, G.N. (1982) A Rasch model for partial credit scoring. *Psychometrika, 47*, 149 – 174

Mellenbergh, G.J. (1994). A Unidimensional Latent Trait Model For Continuous Item Responses. *Multivariate Behavioral Research, 29(3),* 223-236.

Mislevy, R.J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology, 49,* 313-334.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A, 163,* 445-459.

Moustaki, I., & C. O'Muircheartaigh. (2000). A one dimensional latent trait model to infer attitude from nonresponse for nominal data, *Statistica, 2000,* 259-276.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159- 176.

Muraki, E., & Bock, R.D. (2002). PARSCALE: *Parameter scaling of rating data*. Chicago, IL: Scientific Software, Inc.

Muthén, L. K., & Muthén, B. O.(2003). *Mplus version 2.14.* Los Angeles, CA: Muthén & Muthén.

Newton, P.E. (1997). Measuring comparability of standards between subjects: why our statistical techniques do not make the grade. *British Educational Research Journal, 23,* 4, 433-449.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, A, 162),* 177-194.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64.

Rao, C.R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50-57.

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

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Samejima, F. (1969). Estimation of latent proficiency using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17.*

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.

Schafer, J.L, & Olsen, M.K. (1998) Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research, 33 (4),* 545-571.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. . *Psychometrika, 70,* 533-555.

Shi, J. Q., & Lee, S. Y. (1998). Bayesian sampling based approach for factor analysis models with continous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 51,* 233-252.

Skrondal, A. & Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. *Charman & Hall/CRC.*

Smits, N., Mellenbergh, G.J., & Vorst, H.C.M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement, 39,* 187-206.

Snijders, T.A.B., & Bosker, R.J. (1999). Multilevel analysis. An introduction to basic and advance multilevel modeling. *London / Thousand Oaks / New Delhi: Sage.*

Sörbom, D. (1989). Model modification. *Psychometrika, 54,* 371-384.

Strena, A. Ch., & Elliot R. (1987). Differential grading standards revisited. *Journal of Educational Measurment, 24,* 281-291.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

te Marvelde, J. M., Glas, C. A. W., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional IRT models to longitudinal data. *Educational and Psychological Measurement, 66,* 5-34.

Thissen, D., Chen, W.-H., & Bock, R.D. (2002). *Multilog*. Lincolnwood, IL: Scientific Software International.

Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago, IL: University of Chicago Press.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). OPLM: computer program and manual. *Arnhem: Cito, the National Institute for Educational Measure-*

ment, the Netherlands.

Verhelst, N.D., & Verstralen, H.H.F.M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitatieve Methoden, 42,* 73-92.

Willms, J.D. (1992). Monitoring school performance: A guide for educators. *Washington/London: Falmer Press.*

Wood, R., Wilson, D.T., Gibbons, R.D., Schilling, S.G., Muraki, E., & Bock, R.D. (2002). TESTFACT: *Test scoring, Item statistics, and Item factor analysis.* Chicago, IL: Scientific Software International, Inc.

Wu, M.L., Adams, R.J., & Wilson, M.R. (1997). *ConQuest: Generalized Item Response Modeling Software.* Australian Council for Educational Research.

Young, J.W. (1990). Adjusting the Cumulative GPA Using Iten Response Theory. *Journal of Educational Measurement, 27(2),* 175-186.

Young, J.W. (1991). Gender Bias in Predicting College Academic Performance: A New Approach Using Item Response Theory. *Journal of Educational Measurement, 28(1),* 37-47.