ITEM RESPONSE THEORY MODELING WITH NONIGNORABLE MISSING DATA

JONALD L. PIMENTEL
University of Twente, The Netherlands

Samenstelling promotiecommisie

Voorzitter/secretaris	Prof. dr. H.W.A.M. Coonen
Promotor	Prof. dr. C.A.W. Glas
Assistent Promotor	Dr. ir. J.-P. Fox
Leden	Prof. dr. W.J. van der Linden
	Prof. dr. K. Sijtsma
	Prof. dr. H. Kelderman
	Prof. dr. C.W.A.M. Aarts
	Dr. R.R. Meijer

Pimentel, Jonald L.
Item Response Theory Modeling with Nonignorable Missing Data
Proefschrift Universiteit Twente, Enschede. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN: 90-365-2295-1
Druk: PrintPartners Ipskamp B.V., Enschede, The Netherlands

ITEM RESPONSE THEORY MODELING WITH NONIGNORABLE MISSING DATA

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 donderdag 15 december 2005 om 16.45 uur

door

Jonald L. Pimentel

geboren op 24 juni 1969
te Don Carlos, Filippijnen

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

ASSISTENT-PROMOTOR: Dr. ir. J.-P. Fox

Acknowledgement

This doctoral thesis is the result of my four years research work in the Department of Research Methodology, Measurement and Data Analysis at the University of Twente, The Netherlands. I wish to thank individuals and organizations for their direct and indirect support.

First, I would like to express my thanks and deep gratitude to my supervisor Prof. dr. C.A.W. Glas for his teachings, corrections, patience and guidance. Cees, you introduced me and taught me to the world of psychometrics and in particular the item response theory. Without your brilliant ideas, motivation, expertise and help maybe my appreciation towards IRT modeling is still a "missing" phenomenon. Thank you for the opportunity of working with you through four nonignorable years. Our discussions and your coaching I will always treasure. To Dr. Ir. Jean-Paul Fox who served as my daily supervisor for the last two years. Jean, thanks for all the Bayesian ideas. Without your help, writing this thesis would be more difficult for me. I appreciated all your help, guidance and corrections. Our discussions are worth remembering whom I always find not easy but are all challenging.

To Prof. dr. Wim van der Linden for organizing the yearly IRT workshop where our knowledge of IRT are widen. To my colleagues at the department, Bernard, Hans, Wim Tielen, Sebie, Rob, Edith and our

two active secretaries Lorette and Birgit for making my stay in the department so pleasant. The activities we had like the canoeing and horseback riding and the dinners will always be remembered. I say "Dank U wel". To Frans Houweling, Servan Ako, Karen de Vos, Dio Loman, Frances Leusink and Fumiko Yasuno for all their help.

To my former and present fellow PhD students Leonardo, Irene, Adelaide, Anna, Ya-Hui, Anke, Rinke, Oksana and to my former and present officemates Paula, Roberto, Janneke and Joris (for his help in my thesis cover) for the pleasant cultural exchanges.

To the promotion committee members Prof. dr. Wim van der Linden, Prof. dr. Klas Sijtsma, Prof. dr.Henk Kelderman, Prof. dr. Cees Aarts and Dr. Rob Meijer for taking time to read and judge this thesis. To the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS) for organizing courses, as well as IOPS national & international conferences. To the Dutch government and the University of Twente for the financial support during the course of my studies.

I also appreciated the support of Dr. Virgilio G. Oliva, president of USM and the University of Southern Mindanao in Kabacan, Cotabato, Philippines for allowing me to do further studies in the Netherlands.

To my family, relatives, teachers and family friends for their support and prayers. Special thanks to the Sotaridona's for their help during my first year in Enschede. To the Ramaekers (Luc & Nichol), Valere and Marietta and Hauss-Jbil (Veronique, Hamid, Anna and Ilyane) for their hospitality during our visits. To the ITC christian fellowship, of having found a place to worship the Lord with brethren from different countries. To the Prinseschool and the Familie Oeinck (Martin, Wilma, Jesse, Sam), our host family in Enschede for the unforgetable cultural exchanges.

To my wife Florilyn for her love, understanding and prayers and to our wonderful children Jxyf Raymond, Belgie Jan and Ijssel Kyd who are always the source of joy, relief and encouragement and in a way, a good and reliable indicator for patience.

Finally, I thank the Lord for His strength and wisdom. Without Him I can do nothing.

Jonald Pimentel
Enschede, December 2005

Contents

List of Figures

List of Tables

1 Introduction

Psychometrics is a field of study connected to psychology, education and statistics. It deals with the design, administration, analysis and interpretation of tests for the measurement of psychological variables such as intelligence, aptitude, personality traits and abilities. Further, psychometrics has been used in measuring academic achievement and in health related fields, for example, to measure quality of life.

Psychometric methods have several orientations. Pioneers of psychometrics first developed classical test theory (CTT) and then more recent the item response theory (IRT). CTT can be characterized as the theory of measurement errors. The key concepts involve reliability and validity and both can be assessed mathematically. A reliable measure is measuring something consistent while a valid measure is measuring what it is supposed to measure. The major applications of CTT are item and test analyses and observed score equating. On the other hand, IRT can be characterized as a class of probabilistic models for responses of persons to test items. The main focus of this thesis will be on IRT.

1.1 Item Response Theory

Item response theory (IRT) is a class of probabilistic or stochastic models for two-way data, say, the responses of persons or individuals to test items. An important feature of IRT models is parameter separation, which means that the influences of the test items and persons on the responses are modeled by distinct sets of parameters. In IRT, the performance of a person (an examinee) on a test item can be explained by a set of factors called latent traits or abilities. The relationship between a person's item response and the set of traits underlying them can be described an item characteristic curve (ICC) that gives the response probabilities as a function of the latent traits. For dichotomously scored items, this curve is usually monotonically increasing. This curve specifies that as the level of the ability increases, the probability of a correct response to an item increases also.

Under unidimensional IRT models for dichotomous items, the probability of a correct response depends on the persons' unidimensional ability, say θ, and the parameters that characterize the item. Popular models for items with dichotomous responses are the one,two and three normal ogive models and the one,two and three parameter logistic models namely the Rasch model (1PLM; Rasch, 1960), two parameter logistic model (2PLM; Birnbaum, 1968) and the three parameter logistic model (3PLM). For items with polytomous response, models such as the nominal response model (Bock, 1972), graded response model (Samejima, 1969) and partial credit model (Masters, 1982) are used. There are also available models that handle multidimensional cases if items appear to be sensitive to more than one ability (multidimensional IRT; McDonald 1967, Lord & Novick, 1968). Further, IRT models are available for nonmonotone items.

IRT provides a useful framework of solving a wide variety of measurement problems ranging from test construction, to reporting of test scores. Evidence of its importance can be found in the study of differential item functioning (for multiple groups) , person fit analysis, computerized adaptive testing, item banking, structural item response modeling (e.g. Multilevel IRT modeling, see Fox, 2001), test equating and the handling of missing data i.e. modeling and detecting nonignorable missing data processes using IRT models (Holman & Glas, 2005). The latter application is the focus of this thesis.

The first step in applying item response theory to test data is that of estimating the parameters that characterize the chosen item response model. In fact, the successful application of item response theory depends on the availability of satisfactory procedures for estimating the parameters of the model. Estimation procedures that can be employed to obtain parameter estimates using IRT models are available using both likelihood based and Bayesian methods. Likelihood based methods are the joint maximum likelihood estimation (JML), conditional maximum likelihood estimation (CML) and marginal maximum likelihood estimation (MML) employ inferences. These methods are used in software packages as BILOG-MG (Zimowski, Muraki, Mislevy & Bock,1996), MULTILOG, TESTFACT (Wilson, Wood & Gibbons, 1991), MPLUS (Muthén & Muthén ,1998) ,OPLM (Verhelst,Glas & Verstralen,1995) and ConQuest (Wu, Adams & Wilson, 1997). The alternative method, Bayesian estimation methods, employ inferences from posterior distributions. It has been adopted to the estimation of IRT models with multiple raters, multiple item types, missing data (Patz, & Junker; 1999a, 1999b), testlet structures (Bradlow, Wainer & Wang, 1999), and models with multi-level structure on the ability parameters (Fox & Glas, 2001). The unifying theme of these applications is the use of a Markov chain Monte Carlo (MCMC) algorithm for making Bayesian inferences. Most widely used MCMC methods are the Gibbs sampler and the Metropolis-Hasting algorithm. The software packages WinBUGS (Lunn, Thomas, Best, & Spiegelhalter (2000) and MLIRT (available in the web) are some of the available software packages that employ Bayesian estimation methods in IRT.

The second step in applying item response theory to test a data is that of testing the validity of the item response models. A given item response model may or may not be appropriate for a particular set of test data, that is, the model may not adequately predict or explain the data. Hence essentially, in any IRT application, there is in a need to assess the fit of the model to the data. Model fit has two aspects: item fit and person fit. In the first case, the assumptions evaluated are differential item functioning, the form of the item response curve and local stochastic independence. Test statistics have been proposed by such authors as Mokken (1971), Andersen (1973), Yen (1981, 1984), Molenaar (1983), Glas (1999), and Orlando and Thissen (2000). An overview is given by Glas and Suárez-Falcon (2003). Person fit statistics usually focus on the constancy of ability across the test. Examples are the person fit statistics by Smith (1986), and Snijders (2001). An overview is given by Meijer and Sijtsma (1995; 2001).

1.2 Missing data and Ignorability

Statistical analysis of a given data set will be more complicated in the presence of missing data. Standard methods are not directly applicable if these missing data are present. Therefore, there is a growing interest in the statistical methods that properly account for incomplete data (Little & Rubin, 1987). In behavioral sciences and educational measurement settings, for instance, the type of incompleteness that has been studied thoroughly is missing data due to incomplete designs and random missing data such as unit and item nonresponse cases. Individuals for which all responses are missing are called unit nonresponse while individuals for which only the responses to particular items are missing are as known item nonresponse cases. Huisman (1999) studied the occurrence, causes and ways to handle the statistical inferences of item nonresponses. He investigated the nature of missing data patterns and found methods to handle missing data in test items through imputations.

In literature, four common ways to handle missing data are discussed. First, there is the practice of deleting cases with missing data (listwise or pairwise) before doing the actual analyses. Dropping cases with missing values may occasionally be appropriate, but usually this approach has its hazards. The effect of such a practice will reduce the sample size, which leads to inefficient estimation, and it may lead to biased estimates if the missingness is systematic, for instance, if the missing data are correlated with the outcomes of interest. At this point, we do not define the circumstances that lead to bias precisely, this will be done at the end of this section. Besides loss of precision and introduction of bias, deletion of cases may also lower the power of statistical tests and, finally, sometimes the data are too costly to discard. Therefore, most literature discourages this practice

The second way to deal with missing data is the practice of imputation, that is, filling in the missing data with the use of imputation techniques. Examples are mean imputation, regression imputation, hot-deck imputation & multiple imputation (Little & Rubin, 1987). The third way to deal with missing data, is to ignore the missing data and estimate the model using all available observed data. The problem is that the software used must be able to handle the more complicated computations involved. Further, in some situations, which will be discussed below, this approach still leads to biased estimates.

The fourth way to deal with missing data is by explicitly modeling the mechanism that caused the missing data and incorporated this additional model into the model for the observed data. In this thesis we focus on the third and fourth methods for handling missing data in the framework of IRT models.

Above, we used the vague notion of systematic missing data and posited that this form of missingness might lead to bias in estimates. However, a precise analysis of when this actually happens is quite subtle. The problem has been analyzed by Rubin (1976). He discussed the weakest conditions on the process that caused the missing data such that it is always appropriate to ignore this process when making statistical inferences about the distribution of the data of interest. To define this ignorability principle, suppose θ and ζ are the parameters of the observed data and the missing data process, respectively. Further, suppose D is the missing data indicator. In the framework of this thesis, D will be a matrix with elements $d_{ik} = 1$ if for persons i and items k a realization x_{ik} was observed and $d_{ik} = 0$ if x_{ik} was missing. Then, the missing data is said to be missing at random (MAR) if the probability of D given the observed data $x_{(1)}$, missing data $x_{(0)}$, the parameter ζ and, possibly, observed covariates y does not depend on the missing data $x_{(0)}$, that is,

$$P(D|x_{(0)}, x_{(1)}, \zeta, y) = P(D|x_{(1)}, \zeta, y).$$

Furthermore, in a likelihood-based framework, the parameters ζ and θ should be distinct, that is, the joint space of (ζ, θ) should factorize into a (ζ) and (θ) space. If the missing data are MAR and distinctness holds, then the missing data is said to be ignorable. So in likelihood based inferences, if the missing data are MAR then the missing data mechanism or process is ignorable. This means that we do not take into account ζ in the analysis and still the resulting estimates of our parameters are consistent. In the Bayesian framework, the missing data mechanism is said to be ignorable if the missing data are MAR and the priors of ζ and θ are independent.

In educational measurement, it often happens that item nonresponses are nonignorable missing data. An example, for instance, is a test with a time limit condition, where examinees of lower ability do not reach the items at the end. Thus, the pattern of missingness in this case depends on the ability that is measured and hence the missing data are not generally ignorable.

1.3 Objectives and Outline of the Thesis

The topic of this thesis is IRT modeling in the presence of nonignorable missing item responses. The main theme of this thesis is that, apart from the observed item responses, also the variable d_{ik} can be modeled by IRT.

The first part of this thesis (Chapters 2 & 3) will examine the effect in the bias of the model parameter estimates when IRT model for the nonignorable missing data is introduced in the estimation. The purpose of the inclusion of an IRT model for the mechanism that governs the missing data is to reduce the bias in the parameter estimates of the model parameters in the case of violation of Rubin's ignorability principle. Further, the reduction in bias will also be studied when the IRT model for missing data includes observed covariates. The combined model for the observed item responses and the missing data indicator is a multidimensional IRT model with two dimensions for the persons parameters: one for the observed data and one for the missing data. They are assumed to be correlated. The model parameters are estimated using the marginal maximum likelihood method (MML).

In Chapter 2, we investigate through a simulation study for both the dichotomous and polytomous case the effect in the bias of the parameters estimates. Further, the difference in precision of the parameters is investigated between including an IRT model for the missing data process with and without observed covariates.

In Chapter 3, an approach analogous to the approach of Chapter 2 is applied to data from a test with a time limit (a speeded test). The missing data indicator will be modeled using the so-called sequential or steps model (Tutz, 1990; Verhelst, Glas and de Vries, 1997). Also here the model parameters are estimated using MML. Simulation studies are conducted to test the method, first ignoring the missing data process and second including the step model for the missing data.

In Chapter 4, two methods for deciding whether the missing data are ignorable or nonignorable in the IRT framework are proposed that are based on the splitter item technique (Van den Wollenberg, 1979; Molenaar, 1983). It is tested whether the item parameter estimates differ across subsets of item response data. In the first method, the observed data are split-up according to the values of the splitter item. Then, the estimated marginal distributions of the item parameters corresponding to both data sets are compared for detecting differences. In the second method, an IRT model for the observed data is extended with group specific item parameters. These extra parameters, known as Bayesian modification indices (Fox & Glas, 2005) provide information regarding item parameter differences across groups. They are estimated using MCMC, but these estimates do not interfere with the estimation of the other model parameters. Simulation studies were undertaken to illustrate the methods.

In Chapter 5, we develop a fixed effect IRT model for modeling group specific item parameters. The idea is to extend the class of binary IRT models with fixed effects. We propose a general MCMC method to simultaneously estimate all model parameters. The proposed model is used in two practical applications. First, to detect whether a response mechanism is ignorable or not using the splitter item technique and second to detect differential item functioning. Simulation studies are presented to show how the proposed model can be applied.

2

IRT Models for Nonignorable Missing Data Processes

ABSTRACT: Missing data usually present special problems for statistical analyses, especially when the data are not missing at random, that is, when the ignorability principle defined by Rubin (1976) does not hold. This chapter presents a model-based procedure that handles non-ignorable missing data using item response theory (IRT). The relevant model for the observed data is estimated concurrently with the IRT model for the missing data process. As an example, the generalized partial credit model is used to model the observed data while the Rasch model is used to model the missing data process. Simulation studies for dichotomous and polytomous data are presented that show that the bias in the item parameter estimates obtained ignoring the missing data process can be removed or reduced by using the explicit model for the missing data process. It is shown that the IRT model for missing data can also include observed covariates. Using a simulation study, it is shown that the bias in the parameters can be greatly reduced when observed covariates were included in the estimation.

KEYWORDS: item response theory, latent traits, missing data, non-ignorable missing data, observed covariates

2.1 Introduction

In research, missing data is always a source of concern for people who are doing statistical analyses. It raises the level of complexity of making statistical inference. Many researchers, methodologist, and software developers resort to editing the data, although ad hoc edits may do more harm than good by producing results that are substantially biased, inefficient and unreliable (Schafer & Graham, 2002). One way to alleviate the bias in the item parameter estimates is the identification of the variables that explain the cause of missing data. These explanatory variables are called ”mechanism or process” variables. By including a model for this missing data mechanism in the estimation we can reduce or eliminate the bias (due to missingness) in our parameter estimates. Theoretically, if all the process variables associated with a particular piece of missing data can be identified and modeled accurately as controls, the impact of the missing data can be statistically adjusted to the point where it is ignorable (Little & Rubin, 1987). In practice, it is difficult to identify these process variables for all cases of missing data. However, if the given data set contains missing observations, the mechanism causing this missingness can be characterized by its variety of randomness (Rubin, 1976) as missing at random (MAR) and missing completely at random (MCAR).

Suppose θ and ζ are the parameters of the observed data and the missing data process, respectively, and D is the missing data indicator with elements $d_{ik} = 1$ if a realization x_{ik} was observed and $d_{ik} = 0$ if x_{ik} was missing for persons i and items k. Following Rubin's definition, missing data is said to be MAR if the probability of D given the observed data x_{obs}, missing data x_{mis}, some parameter ζ and observed covariates y does not depend on the missing data x_{mis} that is, if

$$P(D|x_{obs}, x_{mis}, \zeta, y) = P(D|x_{obs}, \zeta, y).$$

Furthermore, the parameters ζ and θ are distinct if there are no functional dependencies, that is, restrictions on the parameter space (frequentist version) or if the prior distributions of ζ and θ are independent (Bayesian case). If these two components (MAR and distinctness) are satisfied then the missing data is said to be ignorable, otherwise the missing data are nonignorable. If MAR and distinctness hold, the missing data process is ignorable for statistical inferences, which means that we do not have take into account the distribution of D and ζ, yet the consistency of the estimates is not threatened by the occurrence of the missing data.

In the framework of IRT, missing data can be split into four types (Lord, 1974). The first consists of missing observations which result from a priori fixed incomplete test and calibration designs. The second consists of classes of response-contingent designs such as two and multistage testing (Lord, 1980) designs and computerized adaptive testing. These designs produce ignorable missing data, because the design variables D are completely determined by the observed responses. The third type is ignorable missing data that results from unscalable responses such as “do not know” or “not applicable”, or items missing from booklets. The fourth and last type of missing data results from a nonignorable missing data mechanism. These will, for instance, occur when low-ability respondents fail to produce a response or responses as a result of discomfort or embarrassment. Bradlow and Thomas (1998) mentioned that ignoring this type of missing data process could produce bias in the parameter estimates.

Statistical inference based on the observed data when the missing data process is not ignorable in most cases leads to biased estimates of the parameters of the model. Some literature suggested remedies. One helpful proposal is to model the process that caused the missing data (Heckman, 1979), and the applications discussed below all fall in this category.

Copas and Farewell (1998) argued that nonignorable nonresponse can be explained by covariates such as a subjective measure of enthusiasm to respond. For example, it is expected that when an issue under study is sensitive, an individual may be embarrassed to give a response. In the same manner, students with a low proficiency may fail to respond to difficult items. In the framework of a medical survey, Holman and Glas (2005) report that patients with a relatively high functional status may boost the estimate of their level by failing to respond to items of a physical disability scale.

Moustaki (1996, see also Bartholomew & Knott, 1999) developed a general latent trait and latent class model for mixed observed variables. Within this framework, three methods for dealing with nonignorable missing data were proposed (O’Muircheartaigh & Moustaki, 1999; Moustaki & O’Muircheartaigh, 2000; Moustaki & Knott, 2000). In the first method for the treatment of nonresponse, the missing value is treated as a separate response category. The method includes the missing values in the analysis of the observed items to obtain information about the missing values based on what has been observed, i.e. they used the interrelationships among the items. This information is related to the attitude dimension or dimensions in which they can connect attitude with the nonresponse.

The second method to deal nonresponse is computing response propensities. The idea is to use the propensity score to weight item responses and respondents to account for item and unit nonresponse and to obtain adjusted estimates. This response propensity method uses a logistic or probit regression which is fitted to a binary item response-nonresponse variable for the survey item of interest with a set of covariates.

The third method is to use a latent variable model with two latent dimensions, one to summarize the response propensity and the other to summarize the individual position on the dimension of interest (such as ability or attitude). As an example, O'Muircheartaigh and Moustaki (1999) used a latent variable model for the treatment of item nonresponse in attitude scales. They combined the idea of latent variable identification with the issues of nonresponse adjustment to surveys. This latent variable approach allows missing values to be included in the analysis and equally important allowed information about attitude to be inferred from nonresponse. Their method handled binary (dichotomous), metric and mixed (binary and metric) manifest items with missing values.

Working within the latter approach, Holman and Glas (2005) proposed an IRT model that allows concurrent estimation of IRT item parameters for both a model for the observed dichotomous responses and the missing data indicators. In this chapter we extend this approach to polytomous item responses. Further, we generalize the model to include covariates for the item responses and the missing data indicators. Using a simulation study, we will investigate to what extent the bias in the parameters of the observed data model can be reduced if the observed covariates are included. Both methods (the method with and without covariates) will be applied in both dichotomous and polytomous cases in order to assess the feasibility of the method.

This chapter consists of four sections and is organized in the following manner. After this introduction, the general IRT model for both the observed data and missing data are presented. The following section describes the MML estimation procedure. Finally, the last section presents the simulation studies that will apply the proposed method.

2.2 IRT Models

2.2.1 General IRT model for missing data

Let X be a two-dimensional data matrix with elements x_{ik}, where persons are indexed as $i = 1, ..., N$ and items are indexed as $k = 1, ..., K$. If a combination of i and k has been observed, then the entry x_{ik} is the observation, otherwise it is equal to some arbitrary constant. We define a design matrix D of the same dimension as X with elements $d_{ik} = 1$ if x_{ik} was observed, otherwise $d_{ik} = 0$

Using the elements of X and D, one of our objectives is to make inferences on the individual person parameter θ_i, which are potentially influenced by a latent person variable ζ representing the missing data process.

To model the missing data process, we use a Q-dimensional IRT model proposed by Reckase (1985, 1997) and Ackerman (1996a & 1996b). This model, which is in logistic form, has the probability of an observation given by

$$p(d_{ik} = 1|\zeta_i, \gamma_k, \delta_k) = \frac{\exp(\sum_q^Q \gamma_{kq}\zeta_{iq} - \delta_{k0})}{1 + \exp(\sum_q^Q \gamma_{kq}\zeta_{iq} - \delta_{k0})} \tag{2.1}$$

where γ_{kq} and δ_{k0} are the item parameters (discrimination and difficulty) of the missing data indicator, which we will also refer to as the missing data process.

The model (2.1) is the Rasch model (Rasch, 1960) for dichotomous items when $Q = 1$, and $\gamma_{kq} = 1$ and the two parameter logistic (2PL) model (Lord & Novick, 1968) when $Q = 1$.

When the amount of missing data is small, the appropriate model must have few parameters (like the Rasch model) to be estimable (Lord, 1983).

Another approach of modeling the missing data process is using a normal-ogive representation (McDonald, 1967 & 1997; Lord & Novick, 1968) which is comparable to the logistic approached we used above but we will not discuss it in this chapter.

2.2.2 Combined IRT models for missing data and observed data

- *Combined IRT models of missing data and observed data without observed covariates*

Suppose θ and ζ are the person's latent variables related to the observed data and missing data and let $g(\theta)$ and $g(\ \zeta)$ be their densities. Let $p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)$ be the measurement model for the observed data. It is the probability of the response (observed) variable conditioned on the latent variable of the observed data, the design variable (missing data indicator) and item parameters. Let $p(d_{ik}|\ \zeta_i, \gamma_k, \delta_k)$ be the measurement model for the missing data indicator. It is the probability of the design variable conditioned on the latent variable and item parameters for missing data process. The general models we are using in our estimation procedure are the models described in Holman and Glas (2005). The first model which, we call the MAR model, is given in likelihood form as

$$\prod_{i,k} p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)p(d_{ik}|\zeta_i, \gamma_k, \delta_k)g(\theta_i)g(\zeta_i). \tag{2.2}$$

It is the model that ignores the missing data process, and we ignore the model for the missing data process $p(d_{ik}|\ \zeta_i, \gamma_k, \delta_k)g(\theta_i)$ in the estimation process. The latent variables for the observed data and the missing data process are not related in the MAR model.

The second model, which we call the $NONMAR$ model, is the model where missing data process is included in the estimation process. In this model, the latent variables for both the observed and missing data process θ and ζ are correlated by Σ. This model is written in likelihood form as

$$\prod_{i,k} p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)p(d_{ik}|\zeta_i, \gamma_k, \delta_k)g(\zeta_i, \theta_i|\Sigma), \tag{2.3}$$

where $g(\cdot)$ is the density of ζ_i and θ_i. It is assumed to follow a Multivariate Normal distribution with mean vector $\mathbf{0}$ and variance-covariance Σ. Expressions (2.2) and (2.3) will be used in our procedure to make inferences on the estimation of the model parameters.

In a Bayesian framework, $g(\zeta_i, \theta_i|\Sigma)$ can be seen as a prior for the latent variables. Therefore, statistical inferences under the ignorability assumption are not justified, because the priors of the parameters modeling the observations and the missing data process are not independent. So both Bayesian estimation based on the full posterior and Bayes modal estimation integrating out part of the parameters are not appropriate.

In a frequentist framework, the argument that ignorability does not apply is more subtle. The fact that the two latent variables θ

and ζ are correlated as such does not imply a functional dependence. Béguin and Glas (2001) and Holman and Glas (2005) give conditions for identification of the model. From their conclusions it follows that the basis of the two-dimensional latent space can always be transformed in such a way that both the model for the observations and the model for the missing data indicators depend on the same two latent variables. Therefore, the latent parameters of the two models are not distinct. In other words, within the framework of the model they are functionally dependent.

- *Combined IRT models for missing data and observed data with observed covariates*

The IRT model for missing data can also include observed covariates y. We present a model in likelihood form as

$$\prod_{i,k} p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)p(d_{ik}|\zeta_i, \gamma_k, \delta_k)g(\zeta_i, \theta_i|\Sigma, \eta, y), \tag{2.4}$$

where its components are similar to (2.3), but with an addition of regression coefficients η. The latent variable for the missing data is expressed as a linear combination of the observed covariates that is,

$$\zeta_i = \sum_{s=0}^{p} \eta_s y_{is} + \varepsilon_i \tag{2.5}$$

where we assume $y_{i0} = 1$ and ε_i is the random error which follows a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance Σ.

So in (2.5) we want to model the components ζ of the missing data process through the observed covariates y with the same assumption that latent variables ζ and θ are correlated.

2.2.3 The Generalized Partial Credit Model (GPCM)

For the observed responses, we consider items with dichotomous and polytomous responses and they will be analyzed in general using the multidimensional generalized partial credit model (GPCM; Muraki, 1992). For persons i $(i = 1, ..., N)$ responding to item k $(k = 1, ..., K)$ in category g $(g = 0, ..., m_k)$. The probability of responding in a category g of item k by person i is given by

$$\psi_{kg}(\theta_i) = p(X_{ikg} = 1|\theta_i, \alpha_k, \beta_k) = \frac{\exp(g\sum_q^Q \alpha_{kq}\theta_{iq} - \beta_{kg})}{1 + \sum_{h=1}^{m_k} \exp(h\sum_q^Q \alpha_{kq}\theta_{iq} - \beta_{kh})} \tag{2.6}$$

where α_k and β_k are the item vectors of discrimination and difficulty parameters. where $\alpha_k = \{\alpha_{k1}, ...\alpha_{kq}, ...\alpha_{kQ}\}$ is a Q-dimensional vector of discrimination parameters, $\theta_i = \{\theta_{i1}, ..., \theta_{iq}, ..., \theta_{iQ}\}$ is a Q-dimensional vector of person's parameters and β_{kg} is a scalar item parameter for difficulty. We assume $\beta_{k0} = 0$ so that estimates of β_{kg} are unique

Model (2.6) will be a specific model depending on the values of some of its parameters. When $m_k = 1$, (2.6) is the multidimensional two-parameter logistic model (2PL; Birnbaum, 1968) and in addition becomes the multidimensional partial credit model (PCM; Masters,1982; Masters & Wright, 1997) when item discrimination $\alpha_k = 1$ and further it is the multidimensional Rasch model for dichotomous items when $m_k = 1$.

2.3 MML Estimation

2.3.1 Estimation method

Suppose $\mathbf{x}_i$ is the response pattern of respondent i, and $\mathbf{X}$ is the data matrix. Under the MML approach, it is assumed that possibly multidimensional ability parameters θ_i are independent and identically distributed with density $g(\theta; \lambda)$. Usually, it is assumed the person's ability is normally distributed with population parameters λ (which are the mean μ and variance σ^2 for the unidimensional case, or the mean vector μ and the covariance matrix $\mathbf{\Sigma}$ for the multidimensional case). Item parameters ϕ consist of discrimination parameters (α_k, or α_q for the unidimensional and the multidimensional cases, respectively) and the item difficulties β_k whose elements are $(\beta_{k1}, \beta_{k2}, ..., \beta_{kg}, ..., \beta_{km_k})$. MML estimation derives its name from maximizing the log-likelihood that is marginalized with respect to θ, rather than maximizing the joint log-likelihood of all person parameters θ and item parameters ϕ. Below we will give a general derivation of MML estimation, and therefore the person parameters θ are assumed to include the parameters of the missing data indicator ζ, and likewise ϕ includes γ and δ. Let υ be a vector of all item and population parameters that is $\upsilon^t = (\phi^t, \lambda^t)$. Then the marginal likelihood of υ is given by

$$L(\upsilon; \mathbf{X}, \mathbf{D}) = \int ... \int \prod_i^N p(\mathbf{x}_i, \mathbf{d}_i | \theta_i, \phi) g(\theta_i, \lambda) d\theta_i$$

$$L(\upsilon; \mathbf{X}, \mathbf{D}) = \prod_i^N \int ... \int p(\mathbf{x}_i, \mathbf{d}_i | \theta_i, \phi) g(\theta_i, \lambda) d\theta_i$$

and hence the marginal log-likelihood of υ is

$$\log L(\upsilon; \mathbf{X}, \mathbf{D}) = \log \prod_i^N \int ... \int p(\mathbf{x}_i, \mathbf{d}_i | \theta_i, \phi) g(\theta_i, \lambda) d\theta_i$$

which is equivalent to the expression

$$\log L(\upsilon; \mathbf{X}, \mathbf{D}) = \sum_i^N \log \int ... \int p(\mathbf{x}_i, \mathbf{d}_i | \theta_i, \phi) g(\theta_i, \lambda) d\theta_i. \tag{2.7}$$

The reason for maximizing the marginal rather than the joint likelihood is that maximizing the latter does not generally lead to consistent estimates. This is related to the fact that the number of person parameters grows proportional with the number of observations, and, in general, this leads to inconsistency (Neyman & Scott, 1948). Results from simulation studies by Wright and Panchapakesan (1969) and Fischer and Scheiblechner (1970) showed that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) have shown that MML estimates of structural parameters, say the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models.

Now, the marginal likelihood equations for υ can then be easily derived using Fisher's identity (Efron, 1977; Louis 1982; also see, Glas, 1992, 1998). The first order derivatives with respect to υ can be written as

$$\mathbf{h}(\upsilon) = \frac{\partial}{\partial \upsilon} \log L(\upsilon | \mathbf{X}, \mathbf{D}) = \sum_i^N E(\omega_i(\upsilon) | \mathbf{x}_i, \mathbf{d}_i, \upsilon) \tag{2.8}$$

where $\omega_i(\upsilon)$ is

$$\begin{aligned}\omega_i(\upsilon) &= \frac{\partial}{\partial \upsilon} \log p(\mathbf{x}_i, \theta_i | \mathbf{d}_i, \upsilon) \\ &= \frac{\partial}{\partial \upsilon} \left[\log p(\mathbf{x}_i | \theta_i, \mathbf{d}_i, \phi) + \log g(\theta_i | \lambda)\right] \end{aligned} \tag{2.9}$$

with

$$p(\mathbf{x}_i | \theta_{i,}, \mathbf{d}_i, \phi) = \prod_k \prod_{g=0}^{m_k} \psi_{kg}(\theta_i)^{d_{ik} x_{ikg}} \tag{2.10}$$

and the expectation is with respect to the posterior distribution $p(\theta_i | \mathbf{x}_i, \mathbf{d}_i, \upsilon)$. The identity in (2.8) is closely related to the EM-algorithm (Dempster, Laird and Rubin, 1977), which is an algorithm for finding the maximum of a likelihood marginalized over unobserved data. The present application fits this framework when the response patterns are viewed as observed data and the ability parameters as unobserved data. Together they are referred to as the complete data. The EM algorithm is applicable in situations where direct inference based on the marginal likelihood is complicated, and the complete data likelihood equations, i.e., equations based on $\omega_i(\upsilon)$ are easily solved. Given some estimate of υ as υ^*, the estimate can be improved by solving $\sum_i^N E(\omega_i(\upsilon) | \mathbf{x}_i, \mathbf{d}_i, \upsilon^*) = \mathbf{0}$ with respect to υ. Then this new estimate becomes υ^* and the process is iterated until convergence.

Application of this framework to deriving the likelihood equations of the structural parameters of the multidimensional GPCM proceeds as follows. The likelihood equations are obtained upon equating (2.8) to zero, so explicit expressions are needed for (2.9). Given the design vector $\mathbf{d}_i$, the ability parameter θ_i and the item parameters of the multidimensional GPCM, the probability of response pattern $\mathbf{x}_i$ is given by (2.10). By taking first order derivatives of the logarithm of this expression, the expressions for (2.9) are found as

$$\omega_i(\alpha_{kq}) = d_{ik} \left[\theta_{iq}(x_{ikg} - \psi_{ikg})\right] \tag{2.11}$$

and

$$\omega_i(\beta_{kg}) = d_{ik}(\psi_{ikg} - x_{ikg}) \tag{2.12}$$

where $\psi_{igk} = \psi_{gk}(\theta_i)$, thus the likelihood equations for the item parameters are found upon inserting these expressions into (2.8) and equate the resulting expressions to zero, hence

$$\sum_{i}^{N} E(\theta_{iq}\psi_{ikg}d_{ik}|\mathbf{x_i},\mathbf{d}_i,\upsilon) = \sum_{i}^{N} E(d_{ik}\theta_{iq}x_{ikg}|\mathbf{x}_i,\mathbf{d}_i,\upsilon)$$

simplifying further

$$\sum_{i}^{N} E(\theta_{iq}\psi_{ikg}d_{ik}|\mathbf{x}_i,\mathbf{d}_i,\upsilon) = \sum_{i}^{N} d_{ik}x_{ikg}E(\theta_{iq}|\mathbf{x}_i,\mathbf{d}_i,\upsilon) \tag{2.13}$$

and similarly

$$\sum_{i}^{N} E(d_{ik}\psi_{ikg}|\mathbf{x}_i,\mathbf{d}_i,\upsilon) = \sum_{i}^{N} E(d_{ik}x_{ikg}|\mathbf{x}_i,\mathbf{d}_i,\upsilon)$$

then

$$\sum_{i}^{N} E(d_{ik}\psi_{ikg}|\mathbf{x}_i,\mathbf{d}_i,\upsilon) = \sum_{i}^{N} d_{ik}x_{ikg} \tag{2.14}$$

To derive the likelihood equations for the population parameters, the first order derivatives of the logarithm of the density of the ability parameters $g(\theta|\lambda)$, where λ is the vector of population parameters which is the mean vector μ and the covariance matrix $\mathbf{\Sigma}$ are needed. In the present case, $g(\theta|\mu,\mathbf{\Sigma})$ is the well-known expression for the q-dimensional multivariate normal distribution with mean vector μ and the covariance matrix $\mathbf{\Sigma}$, whose probability density is

$$g(\theta_i|\lambda) = g(\theta_i|\mu,\mathbf{\Sigma}) = (2\pi)^{-q/2}\,|\mathbf{\Sigma}|^{-1/2}\exp\left(-1/2(\theta-\mu)^t\mathbf{\Sigma}^{-1}(\theta-\mu)\right)$$

where $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix, so it is easily verified that these derivatives are given by

$$\omega_i(\mu) = 1/2(\mathbf{\Sigma}^{-1}(\theta-\mu)) \tag{2.15}$$

and

$$\omega_i(\mathbf{\Sigma}) = 1/2[(\theta-\mu)(\theta-\mu)^t\mathbf{\Sigma}^{-2}-\mathbf{\Sigma}^{-1}] \tag{2.16}$$

where elements considered in $\mathbf{\Sigma}$ are the diagonals.

The likelihood equations to obtain μ are again found upon inserting these expressions in (2.8) and equating the resulting expressions to zero, that is

$$\sum_{i}^{N} E(\mathbf{\Sigma}^{-1}(\theta-\mu)|\mathbf{x}_i,\lambda) = \mathbf{0}.$$

and by simplifying the expression by working on the expectations of the stochastic variable θ and the parameters we solve μ as

$$\mu = \frac{\sum_i^N E(\theta|\mathbf{x_i}, \lambda)}{N}$$

Similarly for $\mathbf{\Sigma}$, the resulting expression is

$$\begin{aligned}\sum_i^N E((\theta-\mu)(\theta-\mu)^t \mathbf{\Sigma}^{-2}|\lambda) &= \sum_i^N E((\mathbf{\Sigma}^{-1})^t|\lambda)\\ \sum_i^N E((\theta-\mu)(\theta-\mu)^t \mathbf{\Sigma}^{-2}|\lambda) &= N(\mathbf{\Sigma}^{-1}) \qquad (2.17)\end{aligned}$$

simplifying leads to

$$\mathbf{\Sigma} = \frac{\sum_i^N E((\theta-\mu)(\theta-\mu)^t|\mathbf{x_i}, \lambda)}{N}.$$

Note that the standard errors are also easily derived in this framework: Mislevy (1986) points out that the information matrix can be approximated as

$$\mathbf{H}(\, \upsilon,\ \upsilon) \approx \sum_i^N E(\, \omega_i(\upsilon) \mid \mathbf{x}_i, \mathbf{d}_i, \upsilon\,) E(\, \omega_i(\upsilon) \mid \mathbf{x}_i, \mathbf{d}_i, \upsilon\,)^t \qquad (2.18)$$

and the standard errors are the diagonal elements of the inverse of this matrix.

The basic approach presented so far can be generalized in two ways. First, the assumption that all respondents are drawn from one population can be replaced by the assumption that there are multiple populations of respondents. Usually, it is assumed that each population has a normal ability distribution indexed by a unique mean and variance parameter. Bock and Zimowski (1997) pointed out that this generalization together with the possibility of analyzing incomplete item-administration designs provides a unified approach to such problems as differential item functioning, item parameter drift, non-equivalent groups equating, vertical equating and matrix-sampled educational assessment. Item calibration for CAT also fits within this framework.

2.3.2 Observed Covariates

We will now derive the MML estimation equations for the regression parameters for a model with observed covariates, such as given in (2.5). The population model is now given by

$$g(\theta_i|\mathbf{y}_i, \mathbf{B}, \mathbf{\Sigma}) = (2\pi)^{-q/2} |\mathbf{\Sigma}|^{-1/2} \exp\left(-1/2(\theta_i - \mathbf{B}^t\mathbf{y}_i)^t\mathbf{\Sigma}^{-1}(\theta_i - \mathbf{B}^t\mathbf{y}_i)\right)$$

where $\mathbf{B}$ is a $p \times q$ matrix of regression parameter coefficients and $\mathbf{\Sigma}$ is a $q \times q$ variance-covariance matrix. Equivalently the general model of the latent variables can be expressed as a linear combination of the observed covariates with parameters regression coefficients and parameter residuals in the matrix form

$$\theta = \mathbf{YB} + \mathbf{E} \tag{2.19}$$

where θ is the $n \times q$ matrix of latent variables,$\mathbf{Y}$ is a $n \times p$ matrix of observed covariates, and $\mathbf{E}$ is the $n \times q$ matrix of residuals. In general, if we let $p \times q$ matrix $\widehat{\mathbf{B}}$ be the of estimate of $\mathbf{B}$. Then the maximum likelihood estimate of $\mathbf{B}$ is given by

$$\widehat{\mathbf{B}} = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\theta. \tag{2.20}$$

Furthermore, application of Fisher's identity leads to the expression of a vector of the first order partial derivative with respect to $\mathbf{B}$, that is

$$\begin{aligned}\mathbf{h}(\mathbf{B}) = \frac{\partial}{\partial\mathbf{B}} \ln L(\mathbf{B}|\mathbf{X}, \mathbf{D}) &= \int ... \int (\mathbf{B} - (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\theta)p(\theta|\mathbf{X})d\theta \\ &= \mathbf{B} - (\mathbf{Y}^t\mathbf{Y})^{-1} \int ... \int \mathbf{Y}^t\theta p(\theta|\mathbf{X})d\theta \\ &= \mathbf{B} - (\mathbf{Y}^t\mathbf{Y})^{-1} \left[\mathbf{Y}^t E(\theta|X)\right].\end{aligned} \tag{2.21}$$

Now, simplifying the expectation of the θ given the data, we have

$$\mathbf{h}(\mathbf{B}) = \mathbf{B} - (\mathbf{Y}^t\mathbf{Y})^{-1} \sum_i^N \left[\mathbf{y}_i^t E(\theta_i|\mathbf{x}_i)\right]. \tag{2.22}$$

Setting $\mathbf{h}(\mathbf{B}) = \mathbf{0}$, we can calculate $\widehat{\mathbf{B}}$ as

$$\widehat{\mathbf{B}}= (\mathbf{Y}^t\mathbf{Y})^{-1} \sum_i^N \left[\mathbf{y}_i^t E(\theta_i|\mathbf{x}_i)\right]. \tag{2.23}$$

We can solve the estimate of the variance-covariance of the residuals $\boldsymbol{\Sigma}_{\varepsilon}$ as

$$\widehat{\boldsymbol{\Sigma}_{\varepsilon}} = \frac{1}{N}\sum_{i}^{N}\int_{\theta_i}\left(\theta_i - \mathbf{B}^t\mathbf{y}_i\right)\left(\theta_i - \mathbf{B}^t\mathbf{y}_i\right)^t g(\theta|\mathbf{y_i},\mathbf{B},\boldsymbol{\Sigma}_{\varepsilon})d\theta. \tag{2.24}$$

2.4 Simulation procedure

A simulation study was undertaken to asses the effect of a missing data process as described in (2.3) and (2.4) on the estimates of item parameters. The simulation study consisted of two parts. The first part extends the study by Holman and Glas (2005) to a situation where the model for the missing data indicators is multidimensional, and studies the effects of including no, part of, or all latent dimensions of this model in the estimation. The second simulation study pertains to the effects of adding observed covariates to the model.

2.4.1 Data generation and parameter estimation

To study the effects of including no, part of, or all latent dimensions of the model for the missing data process in the estimation procedure, latent person parameters were drawn from three-variate normal distribution. The sample size was $N = 500$ persons. The variances of the latent variables was always equal to one. The correlation between the latent trait variables θ_i and ζ_i, $\rho(\theta, \zeta)$, varied as $0.0, 0.4$ and 0.8. Also the correlations between the two dimensions of the missing data process $\rho(\zeta_1, \zeta_2)$, varied as $0.0, 0.4$ and 0.8. The items were either dichotomously and polytomously scored. The test consisted of $K = 10$ items. The values d_{ik} and x_{ik} were drawn from $p(d_{ik}|\ \zeta_i, \gamma_k, \delta_k)$ and $p(x_{ik}|d_{ik}, \theta_i, \alpha_k, \beta_k)$, respectively. The data were used to compute MML estimates of the item parameters under various assumptions. Then the values of item parameters estimates over replications r $(r = 1, ..., R,\ R = 100)$, say $\widehat{\phi_r}$ were compared with the values of the parameters used to generate the data using the mean absolute error (MAE) and mean squared error (MSE). There is no index k because all item parameters were equal. The formula to obtained MAE for item parameters is given by

$$MAE(\phi) = \frac{1}{R}\sum_{r=1}^{R}\left|\widehat{\phi_r} - \phi\right| \tag{2.25}$$

where R denote the number of replications of the simulation procedure and $\widehat{\beta_i}$ is the estimate of the item parameter β_k. On the other hand to obtained the MSE for β_k,it is given by

$$MSE(\phi) = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\phi_r} - \phi\right)^2 . \tag{2.26}$$

For the dichotomous case in the simulation, two conditions were used: in the first, the item parameters for all k were $\alpha_k = 1, \gamma_k = 1, \delta_k = -1$ and $\beta_k = 0$, these initial entries give us about 25% missing data and, in the second, we considered $\alpha_k = 1, \gamma_k = 1$, $\delta_k = 0$ and $\beta_k = 0$ which results to about 50% missing data. The MAE and MSE results of the item parameters estimates for the combination $K = 10$ and $N = 500$ are given in Table 2.1 and Table 2.2.

For the polytomous case, items with three response categories were used in the simulation. The item parameters for all k were $\alpha_k = 1, \gamma_k = 1, \delta_k = -1$ and $\beta_k = -1, 1$, and $\alpha_k = 1, \gamma_k = 1, \delta_k = 0$ and $\beta_k = -1, 1$. The MAE and MSE results of the item parameters estimates for the combination $K = 10$, $N = 500$ are given in Tables 2.3 and 2.4.

TABLE 2.1. MAE of item parameter estimates under MAR and NON-MAR models(dichotomous Case); Estimation Model: (Observed data: 2PL, missing data: 1PL); Dimension of missing data process=2; N=500; K=10; $\alpha = 1$; $\beta = 0$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

				Mean Absolute Error			
δ	ρ_1	ρ_2	DMis	α	β	δ	γ
-1	.0	-	-	.168	.102		
	.4	.0	0	.169	.118		
			1	.163	.113	.126	.467
			2	.163	.113	.106	.162
		.4	0	.169	.107		
			1	.164	.102	.109	.205
			2	.165	.102	.108	.149
		.8	0	.170	.110		
			1	.165	.104	.100	.137
			2	.165	.104	.104	.140
	.8	.4	0	.176	.140		
			1	.156	.105	.101	.151
			2	.160	.104	.099	.135
		.8	0	.170	.137		
			1	.153	.103	.099	.136
			2	.156	.103	.103	.137
0	.0	-	-	.225	.120		
	.4	.0	0	.245	.148		
			1	.228	.133	.081	.568
			2	.223	.128	.089	.154
		.4	0	.229	.142		
			1	.209	.124	.079	.194
			2	.209	.125	.084	.147
		.8	0	.222	.144		
			1	.214	.121	.086	.126
			2	.214	.122	.088	.126
	.8	.4	0	.257	.210		
			1	.187	.128	.078	.158
			2	.186	.129	.083	.136
		.8	0	.245	.220		
			1	.192	.133	.083	.121
			2	.194	.133	.083	.121

TABLE 2.2. MSE of item parameter estimates under MAR and NON-MAR model (dichotomous Case); Estimation Model: (Observed data: 2PL, missing data: 1PL); Dimension of missing data process=2; N=500; K=10; $\alpha = 1$; $\beta = 0$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

				Mean Squared Error			
δ	ρ_1	ρ_2	DMis	α	β	δ	γ
-1	.0	-	-	.046	.016		
	.4	.0	0	.046	.021		
			1	.043	.020	.023	.312
			2	.043	.020	.018	.043
		.4	0	.047	.018		
			1	.044	.016	.019	.065
			2	.044	.016	.019	.039
		.8	0	.047	.019		
			1	.044	.017	.016	.031
			2	.044	.017	.017	.032
	.8	.4	0	.051	.029		
			1	.039	.017	.016	.035
			2	.042	.017	.016	.030
		.8	0	.047	.028		
			1	.038	.017	.015	.029
			2	.040	.017	.017	.030
0	.0	-	-	.089	.023		
	.4	.0	0	.110	.034		
			1	.088	.028	.010	.423
			2	.085	.026	.013	.040
		.4	0	.088	.031		
			1	.072	.024	.010	.062
			2	.072	.024	.012	.036
		.8	0	.087	.032		
			1	.077	.023	.012	.025
			2	.076	.023	.013	.027
	.8	.4	0	.023	.062		
			1	.056	.026	.010	.038
			2	.056	.026	.011	.029
		.8	0	.104	.067		
			1	.062	.028	.011	.023
			2	.064	.028	.012	.024

TABLE 2.3. MAE of item parameter estimates under MAR and NONMAR model (Polytomous Case); Estimation Model: (Observed data: PCM, missing data: 1PL); Dimension of missing data process=2; N=500; K=10; $\alpha = 1$; $\beta = -1, 1$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

				Mean Absolute Error				
δ	ρ_1	ρ_2	DMis	α	$\beta1$	$\beta2$	δ	γ
-1	.0	-	-	.137	.135	.194		
	.4	.0	0	.142	.129	.194		
			1	.139	.126	.192	.127	.470
			2	.138	.125	.193	.103	.167
		.4	0	.138	.139	.198		
			1	.136	.129	.194	.109	.192
			2	.135	.129	.194	.107	.162
		.8	0	.136	.137	.206		
			1	.133	.126	.193	.098	.138
			2	.132	.127	.192	.100	.139
	.8	.4	0	.152	.153	.247		
			1	.140	.129	.200	.100	.154
			2	.138	.126	.200	.103	.138
		.8	0	.138	.150	.241		
			1	.129	.125	.196	.098	.130
			2	.128	.125	.197	.102	.130
0	.0	-	-	.187	.156	.239		
	.4	.0	0	.182	.173	.259		
			1	.175	.148	.250	.080	.548
			2	.174	.145	.242	.088	.155
		.4	0	.189	.173	.257		
			1	.182	.150	.241	.078	.182
			2	.182	.150	.243	.084	.143
		.8	0	.188	.182	.274		
			1	.183	.151	.246	.087	.131
			2	.183	.151	.246	.090	.129
	.8	.4	0	.197	.228	.367		
			1	.165	.148	.245	.081	.152
			2	.167	.143	.247	.086	.137
		.8	0	.195	.241	.411		
			1	.170	.154	.250	.088	.120
			2	.171	.153	.249	.090	.123

TABLE 2.4. MSE of item parameter estimates under MAR and NON-MAR model (Polytomous Case); Estimation Model: (Observed data: PCM, missing data: 1PL); Dimension of missing data process=2; N=500; K=10; $\alpha = 1$; $\beta = -1, 1$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

				Mean Squared Error				
δ	ρ_1	ρ_2	DMis	α	$\beta1$	$\beta2$	δ	γ
-1	.0	-	-	.031	.029	.062		
	.4	.0	0	.033	.027	.058		
			1	.031	.025	.059	.023	.322
			2	.031	.024	.061	.016	.046
		.4	0	.031	.032	.061		
			1	.030	.027	.060	.018	.058
			2	.030	.027	.060	.019	.044
		.8	0	.030	.031	.065		
			1	.029	.025	.059	.015	.030
			2	.029	.025	.059	.016	.032
	.8	.4	0	.037	.037	.090		
			1	.029	.026	.064	.015	.036
			2	.030	.024	.063	.017	.030
		.8	0	.031	.037	.087		
			1	.026	.025	.061	.016	.027
			2	.026	.025	.061	.017	.026
0	.0	-	-	.062	.039	.091		
	.4	.0	0	.057	.047	.101		
			1	.053	.035	.100	.010	.418
			2	.052	.033	.095	.012	.039
		.4	0	.062	.049	.100		
			1	.057	.036	.097	.010	.054
			2	.057	.036	.098	.011	.035
		.8	0	.060	.055	.115		
			1	.057	.038	.098	.012	.028
			2	.056	.038	.098	.012	.027
	.8	.4	0	.064	.079	.187		
			1	.041	.035	.092	.010	.034
			2	.043	.033	.095	.012	.030
		.8	0	.064	.088	.225		
			1	.046	.037	.101	.012	.023
			2	.047	.036	.101	.013	.024

All four tables have the same format. In all tables, δ refers to the difficulty parameter for the missing data process used in generating data. Consider Table 2.1. The first row pertains to a base-line condition where $\rho(\theta, \zeta) = 0.0$. So ignorability holds, and there are 25% missing data. The values of the MAE(α) and MAE(β) given in the two columns labeled α and β; they are the mean absolute errors over the 100 replications, and they serve as a baseline. The next three rows pertain to data generated using $\rho(\theta, \zeta) = 0.4$ and $\rho(\zeta_1, \zeta_2) = 0.0$. These data were analyzed using no, one and two dimensions for the missing data indicator. The column DMis refers to the number of dimension of the latent variable for the missing data process. The columns denoted by α, β, δ and γ are the estimated values of the mean absolute errors of the item parameters for the observed data α (discrimination), β (difficulty) and the missing data process δ (difficulty), γ (discrimination). For polytomous case, reported in the Table 2.3 there are two columns for the mean absolute error of the location parameters referred as β_1 and β_2. The analogous mean squared errors are given in Table 2.2 and Table 2.4.

The simulations (please refer to Table 2.1 until Table 2.4) showed that both MAE and MSE values of the item parameters in the parameter estimates were inflated when the model for missing data process was excluded in the parameter estimation. The effect increased as correlation between the latent variables for both observed data and the missing data process increased. For instance if we consider Table 2.1, when $\delta = 0$, (that is, when there are 50% missing data) the baseline, which refers to MAR data, shows that MAE(α) = 0.225 and MAE(β) = 0.120. When $\rho(\theta,\, \zeta) = 0.4$ and $\rho(\zeta_1, \zeta_2) = 0.0$, and the missing data is ignored (DMis = 0), the MAE for α and β have values 0.245 and 0.148 respectively. So the first conclusion is that ignoring the missing data process leads to inflated estimation errors.

When the model for the missing data process was included in the analysis, that is, when the $NONMAR$ model was used, the MAE values dropped to 0.228 for α and 0.133 for β when DMis=1 and MAE(α) = 0.223 and MAE(β) = 0.128, when DMis=2. In general, a decrease in the values of the MAE and the MSE of the item parameters was observed and this decrease was positively related to the number of dimensions included. Similar results are also observed for the values of MAE and MSE of the item parameters δ and γ for missing data process. So the second conclusion is that invoking the missing data process leads to a reduction of estimation errors, even if it is not completely invoked.

The third conclusion that can be drawn from the tables is that when the missing data process is completely modeled, the estimation errors can even fall below the errors of the baseline. For instance, in Table 2.1 we see that for $\delta = 0.0$, the MAE$(\alpha) = 0.225$ for the baseline and MAE$(\alpha) = 0.194$ for $\rho(\theta, \zeta) = \rho(\zeta_1, \zeta_2) = 0.8$ and DMis $=$ 2. Obviously, invoking a model for the missing data indicator results in the exploitation of collateral information.

The fourth conclusion pertains to a main effect of the extent to which MAR is violated. Inspection of the tables shows that if we ignore the missing data process (DMis $=$ 0), the magnitude of the estimation error for $\rho(\theta, \zeta) = 0.8$ is greater than the magnitude for $\rho(\theta, \zeta) = 0.4$. For instance, in Table 2.1 we see that conditionally on $\rho(\zeta_1, \zeta_2) = 0.4$, the MAEs for α are .229 and .257, respectively. Finally, if we consider all results, there is no clear effect of $\rho(\zeta_1, \zeta_2)$

2.4.2 Data generation and parameter estimation with observed covariates

The simulation procedure used was analogous to the simulation procedure in the previous section, but with added feature of including observed covariates. To achieve comparability with the previous section, the regression coefficients were chosen as follows. Let Σ_θ be the covariance matrix of both the latent abilities for the observed responses and the missing data indicator. As in the previous section, there was one dimension for the observed responses and there were two dimensions for the missing data process. Only the case $\rho(\zeta_1, \zeta_2) = 0.8$ was considered here. Further, either $\rho(\theta,\, \zeta) = 0.4$ or $\rho(\theta,\, \zeta) = 0.8$. Let Σ_ε be the diagonal matrix of the variances of the error terms. These variances were all equal to 0.15. The regression coefficients $\mathbf{B}$ were chosen such that

$$\Sigma_\theta = \mathbf{B}\mathbf{B}^t + \Sigma_\varepsilon.$$

Note that the matrix $\mathbf{B}$ is now the Cholesky-decomposition of the matrix $\Sigma_\theta - \Sigma_\varepsilon$, so the upper off-diagonal elements are equal to zero. The latent variables were ordered in such a way that the regression model for the latent variable for the observations only depended on the first covariate, and the two latent variables for the missing data indicator depended on the first two and all three covariates, respectively.

As before, the sample size was $N = 500$. Again the test length was $K = 10$ and the item parameters were also as used above. One hundred replications were made for every combination of δ, $\rho(\theta, \zeta)$ and DMis, where DMis is again the number of dimensions included for the missing data process.

The results are given in the Tables 2.5, 2.6, 2.7 and 2.8. The format of the tables is analogous to the previous four tables, except for an added column *ncov*, which refers to the number of covariates that were included in the parameter estimation. Note that also the baseline model where $\rho(\theta, \zeta) = 0.0$ (the MAR model) includes a covariate. This was done to enable the comparison with the NONMAR models.

TABLE 2.5. MAE of item parameter estimates under MAR and NONMAR model (dichotomous Case); Estimation Model: (Observed data: 2PL, missing data: 1PL); variance=0.15; N=500; K=10; $\alpha = 1$; $\beta = 0$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

					Mean Absolute Error			
δ	ρ_1	ρ_2	DMis	ncov	α	β	δ	γ
-1	.0	-	-	1	.121	.095		
	.4	.8	0	1	.157	.117		
			1	2	.131	.097	.102	.150
			2	3	.120	.097	.095	.104
	.8	.8	0	1	.188	.142		
			1	2	.159	.110	.092	.127
			2	3	.126	.101	.089	.100
0	.0	-	-	1	.145	.117		
	.4	.8	0	1	.170	.164		
			1	2	.150	.116	.083	.134
			2	3	.140	.114	.078	.094
	.8	.8	0	1	.313	.228		
			1	2	.204	.134	.092	.125
			2	3	.155	.126	.084	.097

TABLE 2.6. MSE of item parameter estimates under MAR and NONMAR model (dichotomous Case); Estimation Model: (Observed data: 2PL, missing data: 1PL); variance=0.15; N=500; K=10; $\alpha = 1$; $\beta = 0$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

					Mean Squared Error			
δ	ρ_1	ρ_2	DMis	NCOV	α	β	δ	γ
-1	.0	-	-	1	.023	.014		
	.4	.8	0	1	.042	.020		
			1	2	.028	.015	.017	.036
			2	3	.022	.015	.015	.016
	.8	.8	0	1	.059	.031		
			1	2	.039	.019	.013	.024
			2	3	.024	.016	.012	.015
0	.0	-	-	1	.034	.022		
	.4	.8	0	1	.051	.040		
			1	2	.038	.022	.011	.029
			2	3	.031	.021	.010	.014
	.8	.8	0	1	.190	.070		
			1	2	.075	.028	.013	.025
			2	3	.037	.025	.011	.015

TABLE 2.7. MAE of item parameter estimates under MAR and NONMAR model (Polytomous Case); Estimation Model: (Observed data: PCM, missing data: 1PL); variance=0.15; N=500; K=10; $\alpha = 1$; $\beta = -1, 1$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

					Mean Absolute Error				
δ	ρ_1	ρ_2	DMis	ncov	α	β1	β2	δ	γ
-1	.0	-	-	1	.109	.118	.175		
	.4	.8	0	1	.139	.133	.212		
			1	2	.115	.120	.179	.099	.149
			2	3	.108	.119	.177	.093	.103
	.8	.8	0	1	.143	.159	.254		
			1	2	.130	.125	.198	.103	.132
			2	3	.107	.119	.180	.096	.104
0	.0	-	-	1	.126	.146	.217		
	.4	.8	0	1	.166	.180	.309		
			1	2	.138	.143	.211	.090	.143
			2	3	.128	.140	.211	.084	.101
	.8	.8	0	1	.191	.241	.417		
			1	2	.159	.148	.238	.089	.126
			2	3	.127	.140	.219	.084	.097

TABLE 2.8. MSE of item parameter estimates under MAR and NONMAR model (Polytomous Case); Estimation Model: (Observed data: PCM, missing data: 1PL); variance=0.15 ; N=500; K=10; $\alpha = 1$; $\beta = -1, 1$; $\gamma = 1$; $\rho_1 = \rho(\theta, \zeta)$; $\rho_2 = \rho(\zeta_1, \zeta_2)$.

					Mean squared Error				
δ	ρ_1	ρ_2	DMis	ncov	α	β1	β2	δ	γ
-1	.0	-	-	1	.019	.022	.049		
	.4	.8	0	1	.032	.028	.069		
			1	2	.021	.023	.050	.016	.036
			2	3	.018	.022	.049	.014	.016
	.8	.8	0	1	.033	.039	.093		
			1	2	.026	.025	.060	.017	.027
			2	3	.018	.022	.051	.014	.017
0	.0	-	-	1	.026	.034	.074		
	.4	.8	0	1	.045	.052	.135		
			1	2	.031	.033	.069	.013	.034
			2	3	.026	.031	.069	.011	.015
	.8	.8	0	1	.063	.089	.227		
			1	2	.041	.034	.093	.012	.025
			2	3	.024	.031	.076	.011	.015

Referring to Table 2.5, when we have $\delta = 0$, i.e., 50% missing data. The baseline showed entries for MAE(α) = 0.145 and MAE(β) = 0.117 (as compared to the first simulation in Table 2.1, MAE(α) = 0.225 and MAE(β) = 0.120). This increase in precision is due to the inclusion of a covariate. When we increased the correlation to $\rho(\theta,\, \zeta) = 0.4$ and $\rho(\, \zeta_1,\, \zeta_2) = 0.8$, results showed that when the missing data process was ignored and only the covariate for θ was included in the estimation, the values MAE(α) = 0.170 and MAE(β) = 0.164 were obtained. When one dimension for the missing data process using the $NONMAR$ model which include two covariates were considered i.e., $ncov = 2$, results showed MAE(α) = 0.150 and MAE(β) = 0.116. Further, when two dimensions for the missing data process in the $NONMAR$ model were considered i.e., when three covariates were included in the model for the missing data, results obtained were MAE(α) = 0.140 and MAE(β) = 0.114

It can be seen that increasing the correlation of the latent variables θ and ζ that is increasing the violation of ignorability, resulted in a more bias in the parameter estimates when the covariates are ignored. Including them reduced the bias to a value close to the baseline.

3

Modeling Nonignorable Missing Data in Speeded Tests

ABSTRACT: If a test is administered under a limited-time condition, items at the end of the test are often not endorsed. In most instances, the pattern of missing responses depends on the ability that is measured and, therefore, the missing data are not ignorable in statistical inference. In the present paper, the data are modeled using a combination of two item response theory (IRT) models: one IRT model for the observed response data and one IRT model for the missing data indicator. The missing data indicator is modeled using the sequential model by Tutz (1990, also see, Verhelst, Glas & de Vries, 1997). The two IRT models are connected by invoking the assumption that their latent person parameters have a joint multivariate normal distribution. The model parameters are estimated using marginal maximum likelihood. Simulation studies showed that treating the missing data as ignorable leads to considerable bias in the parameter estimates. Further, it was found that including an IRT model for the missing data removes this bias in the parameter estimates. The impact of the method in practical situations is illustrated with data from the calibration of a time-limit test for measuring intelligence.

KEYWORDS: ignorability, item response theory, marginal maximum likelihood, nonignorable missing data, sequential model, step model

3.1 Introduction

Missing data can be organized into two categories: ignorable and nonignorable missing data. If the missing data are missing at random (MAR) and the parameter of interest and the parameters of the missing data process are distinct, the missing data are ignorable.

That is, with these assumptions, inferences based on the likelihood function and likelihood ratios that ignore the missing data process are valid and consistent (Rubin, 1976; Little & Rubin, 1987; Heitjan, 1994).

When the missing data are nonignorable, the likelihood function and likelihood ratios that ignored the missing data process give rise to biased item parameter estimates (Holman and Glas, 2005). An appropriate method to deal with these problems is to model the missing data process (Heckman, 1979). The idea is to identify and model the explanatory variables in the missing data mechanism or process that caused the missing data. Basing inferences concurrently on this model and the relevant model for the observed data, reduces bias caused by ignoring nonignorable missing data (see for instance O'Muircheartaigh & Moustaki, 1999); Moustaki & O'Muircheartaigh, 2000; Bartholomew & Knott, 1999; Moustaki & Knott, 2000; Holman & Glas, 2005).

In this chapter, we are dealing with item nonresponses in tests and examinations where responses are missing consecutively on items at the end of the test, that is, the respondent has not reached the end of the test because of a time limit. It must be expected that the number of items endorsed is correlated with the respondent's ability level and therefore, the missingness is nonignorable. This form of missingness is closely related to missingness caused by skipping of items by respondents with a low ability. Holman & Glas (2005) show that ignoring this missing data process can lead to bias in the estimates of the item parameters. Bradlow and Thomas (1998) also mentioned that ignoring this type of missing data process could produce bias in the parameter estimates.

In this chapter, it is shown that the missing data indicator of a test with a time limit can be modeled by the sequential model by Tutz (1990), also known as the steps model (Verhelst, Glas, & De Vries, 1997). The observed responses will be modeled by the 2PL model, but this choice is not essential. The step model for the missing data indicator could be combined with any parametric IRT model.

This chapter is made up of six sections and is organized as follows. After this section, a general notation is presented for IRT models for the missing data process and the model for observed data will be discussed. Then a presentation about the estimation procedure using the marginal maximum likelihood method follows. In the next section the results of a number of simulation studies will be presented. An application of the method using an intelligence test will be undertaken in the fifth section. Finally, the last section gives a discussions of the results, and some conclusions and recommendations for further research.

3.2 A General IRT Model

3.2.1 General IRT model for missing data

Let $\mathbf{X}$ be a two-dimensional data matrix with elements x_{ik}, where the persons are indexed $i = 1, ..., N$ and items $k = 1, ..., K$. When the combination of i and k is observed, the entry x_{ik} is the observation, otherwise it is equal to some arbitrary constant. We define a design matrix $\mathbf{D}$ of the same dimension as X with elements

$$d_{ik} = \begin{cases} 0 & \text{if } x_{ik} \text{ is missing} \\ 1 & \text{if } x_{ik} \text{ is observed.} \end{cases}$$

To model the responses missing as a result of the speededness of the test, we will focus on the unobserved responses at the end of the response pattern. So the focus will be on a string of missing data indicators $d_{ik} = 0$, for $k = k', ..., K$. Intermediate missing responses will be considered ignorable missing data.

The table below presents a $N \times 2K$ data matrix , where N is the total number of respondents and K is the total numbers of items. The matrix contains the observed data $\mathbf{X}$ (with missing data indicated by 9 as a dummy) and the missing data indicator $\mathbf{D}$.

	Observed data **X**						Missing data indicator **D**					
Persons	1	2	3	k	.	K	1	2	3	k	.	K
1	0	1	9	9	9	9	1	1	0	-	-	-
2	1	1	9	9	9	9	1	1	0	-	-	-
3	1	0	1	9	9	9	1	1	1	0	-	-
.	.	.	.	.	.	.	.	.	.	.	.	.
i	1	1	0	1	9	9	1	1	1	1	0	-
.	.	.	.	.	.	.	.	.	.	.	.	.
	1	0	1	0	1	9	1	1	1	1	1	0
N	1	1	0	1	0	0	1	1	1	1	1	1

The item nonresponse occurred due to time limit condition and interacts with the level ability of the respondent. In a case of dichotomously scored items, respondent i answering to item k can get a correct 1 or incorrect 0 response. Then he stumbled on an item which has 0 entry in **D** and 9 in **X**,and the succeeding items are skipped as well. To model the missing data process of this case, we use the steps model (Verhelst, Glas, & De Vries;1997) given by

$$\begin{aligned} &p(d_{i1} = d_{i2} = ... = d_{ik-1} = 1 \ \& \ d_{ik} = 0) = \\ &\qquad \left[\prod_{h=1}^{k} p_k\left(\zeta_i\right)\right]\left[1 - p_{k+1}(\zeta_i)\right] \quad \text{if } 0 \leq k < K \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} &p(d_{i1} = d_{i2} = ... = d_{ik-1} = d_{ik} = 1) = \\ &\qquad \prod_{k=1}^{K} p_k\left(\zeta_i\right) \quad \text{if } k = K, \end{aligned} \tag{3.2}$$

where we assumed that

$$p_k\left(\zeta_i\right) = \frac{\exp(\gamma_k \zeta_i - \delta_k)}{1 + \exp(\gamma_k \zeta_i - \delta_k)}. \tag{3.3}$$

So $p_k\left(\zeta_i\right)$ is equivalent to the 2PL model for dichotomously scored items. The model entails that the respondent makes item-steps until the first wrong response, and then stops taking item-steps. Usually the data show too little variation to estimate the slope parameter γ_k, so usually we assume that $\gamma_k = 1$. Note that ζ_i refers to the latent variable for the missing data process. Further, we impose a restriction on the difficulty of the item -steps:

$$\delta_k = \tau + (k - K)\eta,$$

where τ is the baseline (overall) level and η models a monotone change in the probability of an observation as a function of the position of the item in the test. The reason for this restriction on δ_k is that the first item-steps are usually taken by all respondents, so the difficulty of these steps cannot be estimated. Further, the restriction supports a monotonously decreasing probability of observing a response.

3.2.2 Combined IRT models for Observed data and missing data

The models that we will use for the comparison are analogous to the MAR and $NONMAR$ models described in Holman and Glas (2005) and in the previous chapter (refer to equation (2.2) and (2.3)). The likelihood of the MAR model is given by

$$\prod_{i,k} p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)p(d_{ik}|\zeta_i, \gamma_k, \delta_k)g(\zeta_i)g(\theta_i) \qquad (3.4)$$

where $p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)$ is the measurement model, $\theta_{i,}$ is the latent person ability parameter, and α_k, β_k are item parameters of the observed data. Further, $p(d_{ik}|\zeta_i, \gamma_k, \delta_k)$ is the model for the missing data indicator and γ_k, δ_k are item parameters of the missing data process. Finally, $g(\zeta_i)$ and $g(\theta_i)$ are the densities of the latent parameters. We assume these densities to be standard normal. In (3.4), the latent variables θ for the observed data and ζ for the missing data process are not correlated and hence we can ignore the model for the missing data i.e. for maximum likelihood estimation we can ignore $p(d_{ik}|\zeta_i, \gamma_k, \delta_k)g(\zeta_i)$. On the other hand, the $NONMAR$ model also described in Chapter 2 in this thesis, is the model where the missing data process is included in the estimation. In that case, the latent variables for both the observed data and the missing data process, θ and ζ, respectively, are correlated by with a correlation parameter $\mathbf{\Sigma}$. The likelihood of the $NONMAR$ model is written as

$$\prod_{i,k} p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)p(d_{ik}|\zeta_i, \gamma_k, \delta_k)g(\zeta_i, \theta_i|\mathbf{\Sigma})), \qquad (3.5)$$

where $g(\cdot)$ is the density of ζ_i and θ_i which is assumed to follow a multivariate normal distribution with mean vector 0 and variance-covariance $\mathbf{\Sigma}$. Expressions (3.5) will be used in the procedure to make inferences when all latent variables for observed data and missing data process are considered and then compared to the results on values of item parameters estimates when the (3.4) model was used.

3.2.3 The Generalized Partial Credit Model (GPCM)

In general, the observed responses will be modeled by a multidimensional version of the generalized partial credit model (Muraki, 1992). In the unidimensional case, the person's ability or proficiency is represented by a scalar parameter. However, in many cases it is a priori clear that multiple abilities are involved in producing the observed responses or the dimensionality of the ability structure might not be clear at all. In these cases, multidimensional ability parameters are needed to describe the ability or proficiency level of a person. Béguin & Glas (2001) state that multidimensional IRT models can serve confirmatory and exploratory purposes.

For persons i $(i = 1, ..., N)$ responding to item k $(k = 1, ..., K)$ the probability of responding in a category g $(g = 0, ..., m_k)$ is given by

$$\psi_{kg}(\theta_i) = p(X_{ikg} = 1|\theta_i, \alpha_k, \beta_k) = \frac{\exp(g \sum_q^Q \alpha_{kq}\theta_{iq} - \beta_{kg})}{1 + \sum_{h=1}^{m_k} \exp(h \sum_q^Q \alpha_{kq}\theta_{iq} - \beta_{kh})} \tag{3.6}$$

where $\alpha_k = \{\alpha_{k1}, ...\alpha_{kq}, ...\alpha_{kQ}\}$ is a Q-dimensional vector of discrimination parameters or factor loadings, $\theta_i = \{\theta_{i1}, ..., \theta_{iq}, ..., \theta_{iQ}\}$ is a Q-dimensional vector of person's parameters and β_{kg} is a scalar location parameter.

Model (3.6) will be a specific model depending on the values of some of its parameters. When $m_k = 1$, (3.6) is the multidimensional two-parameter logistic model (2PL; Birnbaum, 1968) which is the one we use in the simulation studies reported in this chapter and, further, (3.6) becomes the multidimensional partial credit model (PCM; Masters, 1982; Masters & Wright, 1997) when $\alpha_k = 1$ and additionally, the multidimensional Rasch model for dichotomous items when $m_k = 1$ and $\alpha_k = 1$.

Note that the model for the missing data indicator (3.3) is a special case of the GPCM given by (3.6). Therefore, both models can be combined into one concurrent model, for instance a model of the form of (3.5), by assuming a Q-dimensional model where the model for the item responses only loads on the first $Q-1$ dimensions, while the model for the missing data indicator only loads on the Q-th dimension. The ensemble of the latent parameters of a respondent then has a Q-variate normal distribution with a mean equal to zero and a covariance matrix $\mathbf{\Sigma}$.

3.3 MML Estimation

Suppose $\mathbf{x}_i$ is the response pattern of respondent i, and let $\mathbf{X}$ be the data matrix. Under MML approach, it is assumed that possibly multidimensional ability parameters θ_i are independent and identically distributed with density $g(\theta; \lambda)$. Usually, it is assumed that person's ability is normally distributed with population parameters λ (which are the mean μ and the variance σ^2 for the unidimensional case, or the mean vector μ and the covariance matrix $\mathbf{\Sigma}$ for the multidimensional case). Item parameters ϕ consist of discrimination parameters (α_k, or α_k for the unidimensional and the multidimensional cases, respectively) and the item difficulties β_k whose elements are $(\beta_{k1}, \beta_{k2}, ..., \beta_{kg}, ..., \beta_{km_k})$. Given the remark in the previous section that the models for the item responses and the missing data indicator can be brought together in one concurrent model, MML estimation will be described without explicitly distinguishing between item parameters and person parameters associated with the observations or the indicators.

MML estimation derives its name from maximizing the log-likelihood that is marginalized with respect to θ, rather than maximizing the joint log-likelihood of all person parameters θ and item parameters ϕ. Let υ be a vector of all item and population parameters that is $\upsilon^t = (\phi^t, \lambda^t)$. Then the marginal likelihood of υ is given by

$$L(\upsilon; \mathbf{X}, \mathbf{D}) = \int ... \int \prod_i^N p(\mathbf{x}_i, \mathbf{d}_i | \theta_i, \phi) g(\theta_i; \lambda) d\theta_i$$

that is

$$L(\upsilon; \mathbf{X}, \mathbf{D}) = \prod_i^N \int ... \int p(\mathbf{x}_i, \mathbf{d}_i | \theta_i, \phi) g(\theta_i; \lambda) d\theta_i$$

and hence the marginal log-likelihood of υ is

$$\log L(\upsilon; \mathbf{X}, \mathbf{D}) = \log \prod_i^N \int ... \int p(\mathbf{x}_i, \mathbf{d}_i | \theta_i, \phi) g(\theta_i; \lambda) d\theta_i$$

which is equivalent to the expression

$$\log L(\upsilon; \mathbf{X}, \mathbf{D}) = \sum_i^N \log \int ... \int p(\mathbf{x}_i, \mathbf{d}_i|\theta_i, \phi) g(\theta_i; \lambda) d\theta_i. \tag{3.7}$$

We maximized the marginal likelihood since it gives us consistent estimates as compared to the ones obtained using the joint likelihood which can be inconsistent. Neyman & Scott (1948) stated that if the number of person parameters grows proportional with the number of observations, then in general this leads to inconsistency when using joint likelihood. Simulation studies of Wright and Panchapakesan (1969) and Fischer and Scheiblechner (1970) showed that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) have shown that marginal maximum likelihood estimates of structural parameters, say the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models.

To derive MML equations, we will introduce the vector of derivatives

$$\begin{aligned} \omega_i(\upsilon) &= \frac{\partial}{\partial \upsilon} \log p(\mathbf{x}_i, \mathbf{d}_i, \theta_i|\upsilon) \qquad (3.8) \\ &= \frac{\partial}{\partial \upsilon} \left[\log p(\mathbf{x}_i, \mathbf{d}_i|\theta_i, \phi) + \log g(\theta_i|\lambda)\right]. \end{aligned}$$

Using Fisher's identity (Efron, 1977; Louis 1982; also see, Glas, 1992, 1998), then the marginal likelihood equations for υ can then be easily derived. The first order derivatives with respect to υ is written as

$$\mathbf{h}(\upsilon) = \frac{\partial}{\partial \upsilon} \log L(\upsilon|\mathbf{X}, \mathbf{D}) = \sum_i^N E(\omega_i(\upsilon)|\mathbf{x}_i, \mathbf{d}_i, \upsilon) \tag{3.9}$$

where $\omega_i(\upsilon)$ is the expression in (3.8) and the expectation is with respect to the posterior distribution $p(\theta_i|\mathbf{x}_i, \mathbf{d}_i, \upsilon)$. The identity in (3.9) is closely related to the EM-algorithm (Dempster, Laird and Rubin, 1977), which is a very useful algorithm for finding the maximum of a likelihood marginalized over unobserved data. This framework fits the present application when the response patterns are viewed as observed data and the ability parameters as unobserved data. Together they are referred to as the complete data. When direct inference based on the marginal likelihood is complicated, the

EM algorithm is applicable in this situations. The complete data likelihood equations, i.e., equations based on $\omega_i(\upsilon)$ are easily found. Given some estimate of υ as υ^*, the estimate can be improved by solving $\sum_i^N E(\omega_i(\upsilon)|\mathbf{x}_i, \mathbf{d}_i, \upsilon^*) = \mathbf{0}$ with respect to υ. Then this new estimate becomes υ^* and the process is iterated until convergence.

Applications of this framework in deriving the likelihood equations of the structural parameters of the multidimensional GPCM proceeds as follows. We will only consider finding the item parameter estimates for the item responses, because the item parameter estimates for the missing data indicators is completely analogous. The complete likelihood is given by

$$p(\mathbf{x}_i|\theta_{i,}, \mathbf{d}_i, \phi) = \prod_k \prod_{g=0}^{m_k} \psi_{kg}(\theta_i)^{d_{ik}x_{ikg}}. \tag{3.10}$$

The likelihood equations are obtained upon equating (3.9) to zero, so explicit expressions are needed for (3.8). Given the design vector $\mathbf{d}_i$, the ability parameter θ_i and the item parameters of the multidimensional GPCM, the probability of response pattern $\mathbf{x}_i$ is given by (3.10). By taking first order derivatives of the logarithm of this expression, the expressions for (3.8) are found as

$$\omega_i(\alpha_{kq}) = d_{ik}\left[\theta_{iq}(x_{ikg} - \psi_{ikg})\right] \tag{3.11}$$

and

$$\omega_i(\beta_{kg}) = d_{ik}(\psi_{ikg} - x_{ikg}), \tag{3.12}$$

where $\psi_{igk} = \psi_{gk}(\theta_i)$, thus the likelihood equations for the item parameters are found upon inserting these expressions into (3.9) and equate the resulting expressions to zero, hence

$$\sum_i^N E(\theta_{iq}\psi_{ikg}d_{ik}|\mathbf{x_i}, \mathbf{d}_i, \upsilon) = \sum_i^N E(d_{ik}\theta_{iq}x_{ikg}|\mathbf{x}_i, \mathbf{d}_i, \upsilon)$$

simplifying further

$$\sum_i^N E(\theta_{iq}\psi_{ikg}d_{ik}|\mathbf{x}_i, \mathbf{d}_i, \upsilon) = \sum_i^N d_{ik}x_{ikg}E(\theta_{iq}|\mathbf{x}_i, \mathbf{d}_i, \upsilon) \tag{3.13}$$

and similarly

$$\sum_i^N E(d_{ik}\psi_{ikg}|\mathbf{x}_i, \mathbf{d}_i, \upsilon) = \sum_i^N E(d_{ik}x_{ikg}|\mathbf{x}_i, \mathbf{d}_i, \upsilon)$$

then

$$\sum_{i}^{N} d_{ik} E(\psi_{ikg}|\mathbf{x}_i, \mathbf{d}_i, \upsilon) = \sum_{i}^{N} d_{ik} x_{ikg} \qquad (3.14)$$

To derive the likelihood equations for the population parameters, the first order derivatives of the logarithm of the density of the ability parameters $g(\theta; \lambda)$, where λ is the vector of population parameters which is the mean vector μ and the covariance matrix $\mathbf{\Sigma}$ are needed. In the present case, $g(\theta; \mu, \mathbf{\Sigma})$ is the well-known expression for the Q-dimensional multivariate normal distribution with mean vector μ and the covariance matrix $\mathbf{\Sigma}$, whose probability density is

$$g(\theta_i; \lambda) = g(\theta_i|\mu, \mathbf{\Sigma}) = (2\pi)^{-q/2} \left|\mathbf{\Sigma}\right|^{-1/2} \exp\left(-1/2(\theta - \mu)^t \mathbf{\Sigma}^{-1}(\theta - \mu)\right)$$

where $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix, so it is easily verified that these derivatives are given by

$$\omega_i(\mu) = 1/2(\mathbf{\Sigma}^{-1}(\theta - \mu)) \qquad (3.15)$$

and

$$\omega_i(\mathbf{\Sigma}) = 1/2[(\theta - \mu)(\theta - \mu)^t \mathbf{\Sigma}^{-2} - (\mathbf{\Sigma}^{-1})^t] \qquad (3.16)$$

where elements considered in $\mathbf{\Sigma}$ are the diagonals.

The likelihood equations to obtain μ are again found upon inserting these expressions in (3.9) and equating the resulting expressions to zero, that is

$$\sum_{i}^{N} E(\mathbf{\Sigma}^{-1}(\theta - \mu)|\mathbf{x}_i, \lambda) = \mathbf{0}$$

and by simplifying the expression by working on the expectations of the stochastic variable θ and the parameters we solve μ as

$$\mu = \frac{\sum_i^N E(\theta|\mathbf{x_i}, \lambda)}{N}$$

Similarly for $\mathbf{\Sigma}$, the resulting expression is

$$\sum_{i}^{N} E((\theta - \mu)(\theta - \mu)^t \mathbf{\Sigma}^{-2}|\lambda) = \sum_{i}^{N} E((\mathbf{\Sigma}^{-1})^t|\lambda)$$

$$\sum_{i}^{N} E((\theta - \mu)(\theta - \mu)^t \mathbf{\Sigma}^{-2}|\lambda) = N(\mathbf{\Sigma}^{-1}) \qquad (3.17)$$

and simplifying leads to

$$\mathbf{\Sigma} = \frac{\sum_i^N E((\theta-\mu)(\theta-\mu)^t|\mathbf{x_i},\lambda)}{N}$$

Note that the standard errors are also easily derived in this framework: Mislevy (1986) pointed out that the information matrix can be approximated as

$$\mathbf{H}(\ \upsilon,\ \upsilon) \approx \sum_i^N E(\,\omega_i(\upsilon) \mid \mathbf{x}_i, \mathbf{d}_i, \upsilon\,)E(\,\omega_i(\upsilon) \mid \mathbf{x}_i, \mathbf{d}_i, \upsilon\,)^t \qquad (3.18)$$

and the standard errors are the diagonal elements of the inverse of this matrix.

The basic approach presented so far can be generalized in two ways. First, the assumption that all respondents are drawn from one population can be replaced by the assumption that there are multiple populations of respondents. Usually, it is assumed that each population has a normal ability distribution indexed by a unique mean and covariance matrix. This generalization together with the possibility of analyzing incomplete item-administration designs provides a unified approach to such problems as differential item functioning, item parameter drift, non-equivalent groups equating, vertical equating and matrix-sampled educational assessment as pointed out by Bock and Zimowski (1997). Further, item calibration for CAT also fits within this framework.

3.4 Simulation Studies

Simulation studies were conducted to asses the effect in the bias of the item parameter estimates when a model for missing data is ignored or included in a model for estimation as described in (3.4) and (3.5). We divided the simulation study into two parts. The first part consists of the data generation using the Rasch model (RM or $1PL$) and the second part was the estimation of the parameters. Two models were used in the estimation of item parameters. The RM model was used for the estimation of item difficulty parameter $\widetilde{\beta}_{MAR}$ when the model for missing data process was ignored (MAR model). The RM version of the sequential (steps) model for the missing data model was used for the model of the missing data when this model was included in the estimation that was concurrent with the estimation of the RM model for the observed data. The item difficulty δ of the conceptual items for the missing data indicator has τ (overall level) and η (increment) as components. Their main purposed was described in the previous section.

For a sample size $N = 500$ persons, latent trait values (θ_i, ζ_i) were drawn from a bivariate normal distribution with means 0, and a covariance matrix $\mathbf{\Sigma}$ with diagonal elements equal to one and correlation ρ. This correlation between the latent variables for the observed data and the missing data process where chosen to vary from 0.0, 0.2, 0.4, 0.6 and 0.8. The test was made of $K = 10$ items and the items were dichotomously scored. The values d_{ik} and x_{ik} were drawn from $p(d_{ik}| \ \zeta_i, \theta_i, \delta_k)$ and $p(x_{ik}|d_{ik}, \theta_{i,}\alpha_k, \beta_k)$ respectively. The generated data were used to compute estimates of the item parameters $\widetilde{\beta}_{MAR}$ when the MAR model was used and item parameter estimates $\widehat{\beta}_{obs}, \widehat{\delta}_{obs}$, $\widehat{\tau}_{obs}$ and $\widehat{\rho}$ when $NONMAR$ model was used, respectively.

The values of $\widetilde{\beta}_{MAR}, \widehat{\beta}_{obs}, \widehat{\delta}_{obs}$, $\widehat{\tau}_{obs}$ and $\widehat{\rho}$ were compared with the values of the parameters used to generate the data (true values) using the mean absolute error (MAE) and mean squared error (MSE). For the formulas to obtained MAE and MSE for the model parameters refer to the equations (2.25) and (2.26) in Chapter 2.

One hundred replications were made for the combination of $K =$ 10 and $N = 500$ and $\rho = 0.0$, 0.2, 0.4, 0.6 and 0.8. The same replications for the combination $n = 1000$ and same k and ρ was also done. The difficulty parameters for the observed data is $\beta_k = 0$ while difficulty parameters for the conceptual items were δ_k =-8, -7, -6, -5, -4, -3, -2, -1, 0, and 1, respectively. These values were chosen such that the item parameters will go from easy to difficult, that is, the probability of observing an item response decreases. So most respondents can respond the first items until they run out of time and then they omit the succeeding items. This is the situation of missingness we are dealing with, where missingness can not be ignored since the response mechanism depends on the ability of the respondents. The result of the simulations are given on Table 3.1 and Table 3.2 for $N = 500$ and $N = 1000$.

TABLE 3.1. MAE of item parameter estimates under MAR and NONMAR model (dichotomous Case); Estimation Model: (Observed data: 2PL; missing data: Steps model); Dimension of missing data process: 1; N=500, 1000; K=10; $\alpha = 1$; $\beta = 0$; $\gamma = 1$; $\rho = \rho(\theta, \zeta)$.

		Mean Absolute Error(MAE)				
N	ρ	$\widehat{\beta}_{obs}$	η	τ	ρ	$\widetilde{\beta}_{MAR}$
500	.0	.107	.121	.288	.094	.107
	.2	.112	.116	.294	.090	.117
	.4	.112	.122	.299	.088	.121
	.6	.114	.104	.253	.129	.126
	.8	.110	.070	.182	.138	.139
1000	.0	.083	.077	.192	.061	.083
	.2	.077	.082	.206	.067	.081
	.4	.079	.078	.193	.075	.091
	.6	.078	.065	.157	.075	.103
	.8	.079	.049	.128	.087	.124

TABLE 3.2. MSE of item parameter estimates under MAR and NONMAR model (dichotomous Case); Estimation Model: (Observed data: 2PL; missing data: Steps model); Dimension of missing data process: 1; N=500, 1000; K=10; $\alpha = 1$; $\beta = 0$; $\gamma = 1$; $\rho = \rho(\theta, \zeta)$.

		Mean Squared Error(MSE)				
N	ρ	$\widehat{\beta}_{obs}$	η	τ	ρ	$\widetilde{\beta}_{MAR}$
500	.0	.020	.031	.170	.015	.020
	.2	.022	.023	.137	.011	.024
	.4	.022	.024	.147	.014	.028
	.6	.024	.023	.134	.016	.030
	.8	.024	.010	.057	.009	.044
1000	.0	.012	.010	.059	.005	.012
	.2	.010	.010	.060	.006	.012
	.4	.011	.010	.059	.006	.016
	.6	.011	.008	.050	.007	.022
	.8	.011	.004	.030	.004	.032

We start the discussion of our results by introducing the notations in the tables. Notation $\rho(\theta,\, \zeta)$ refers to the correlation of the latent variables of the observed data θ and the missing data process ζ. The MAE($\widehat{\beta}_{obs}$) and MSE($\widehat{\beta}_{obs}$) refers to the mean absolute error and mean squared error, respectively, of the estimates of the difficulty parameter $\widehat{\beta}_{obs}$ when the $NONMAR$ model was used. MAE($\widetilde{\beta}_{MAR}$) and MSE($\widetilde{\beta}_{MAR}$) refers to the mean absolute error and mean squared error, respectively, of the estimates of the difficulty parameter $\widetilde{\beta}_{MAR}$ that ignored the model for the missing data (MAR model). The other notations for the MAE's and MSE's of η, τ and ρ refers to the increment of the 'conceptual' items, the difficulty parameter of the last 'conceptual' item and the correlation between θ and ζ, respectively.

The first row of each table are baselines, that is when there is no correlation between the two latent variables, so when ignorability holds. The results showed that when the correlation increases, the MAE and the MSE of the estimates under the assumption of MAR increased considerably. For instance, if we look specifically at Table 3.1, where ρ increases from 0.0 to 0.8 with intervals of 0.2, the MAE($\widetilde{\beta}_{MAR}$) had values of 0.107, 0.117, 0.121, 0.126 and 0.139. These values of the MAE (or the bias) for the $\widetilde{\beta}_{MAR}$ were inflated as expected since the missing data process was excluded. It was also true for MSE of $\widetilde{\beta}_{MAR}$. These results were analogous to the results of the simulations of Holman & Glas (2005) and the results reported in the previous chapter in this thesis.

The MAE and MSE values for $\widehat{\beta}_{obs}$ only showed random fluctuation. Looking again in the Table 3.1, MAE($\widehat{\beta}_{obs}$) had values of 0.107, 0.112, 0.112, 0.114 and 0.110. On the other hand, MAE's of $\widehat{\eta}_{obs}$ showed 0.121, 0.116, 0.122, 0.104, 0.070 and $\widehat{\tau}_{obs}$ showed 0.192, 0.206, 0.193, 0.157, 0.128. The errors of these parameters estimates showed a decreasing trend. Since the marginal distribution of **D** did not change as a function of the correlation, this trend cannot be explained as a result of having more observations in **D** or more information in the responses in **D**. A possible explanation for such a trend is that an increase in the correlation between θ and ζ results in more collateral information on ζ through θ, and therefore more information on τ and η. This collateral informations resulted on a decrease in the standard error of the item parameter estimates for the missing data indicators.

The MAE values for estimated correlation ρ were 0.094, 0.090, 0.088, 0.129 and 0.138, which is an increasing trend. We have no clear explanation of this phenomenon at the moment of this writing.

3.5 Real Data Application

To get an idea of the impact of the approach in a real data situation, data from a calibration sample of an intelligence test for children in primary education was analyzed (van Dijk & Tellegen, 2004). The data set was made up 3145 children responding to of 30 items of a speeded form of the test. The percentage of missing data was equal to 27%. The first 5 items were responded to by all children, the last five items were endorsed by 1004, 855, 786, 622 and 508 children, respectively. The data ware analyzed using both the MAR (ignorable) and NONMAR (nonignorable) models. The estimated correlation between the latent parameters of the observed data and missing data is 0.429, signifying that the missing data process cannot be ignored in the estimation. For both approaches, the values of the items parameters estimates were compared. Results of the parameters estimates are given in Table 3.3. The notation *diff* means the difference between the estimates of the parameters. $\widehat{\alpha}$ and $\widehat{\beta}$ were the estimates of the item discrimination and difficulty respectively of the model that includes the model for the missing data while $\tilde{\alpha}$ and $\tilde{\beta}$ were the estimates of the item discrimination and difficulty respectively of the model that ignored the missing data. The differences between the estimates obtained using the two methods is plotted in Figure 3.1. It can be seen that there is no trend on the difference between the discrimination parameters, but the differences in the item difficulty parameters clearly increase after the 20th item.

TABLE 3.3. Item parameter estimates under MAR and NONMAR model Real data (speeded test),N=3145 examinees K=30 items

item	$\widehat{\alpha}$	$\tilde{\alpha}$	*diff*	$\widehat{\beta}$	$\tilde{\beta}$	*diff*
1	0.876	0.847	0.029	-4.303	-4.277	-0.026
2	1.254	1.281	-0.027	-3.529	-3.550	0.021
3	1.208	1.210	-0.002	-3.132	-3.130	-0.002
4	0.777	0.773	0.004	-1.692	-1.690	-0.002
5	1.154	1.140	0.014	-0.687	-0.684	-0.003
6	0.340	0.351	-0.011	-1.398	-1.400	0.002
7	1.002	0.986	0.016	0.1930	0.1920	0.001
8	1.403	1.400	0.003	0.2680	0.2660	0.002
9	0.701	0.701	0.000	-0.377	-0.378	0.001
10	0.298	0.292	0.006	0.477	0.4760	0.001
11	1.583	1.593	-0.010	0.197	0.1890	0.008
12	1.563	1.538	0.025	0.300	0.2890	0.011
13	0.342	0.347	-0.005	-0.078	-0.081	0.003
14	0.811	0.819	-0.008	-0.479	-0.490	0.011
15	0.893	0.885	0.008	0.907	0.8930	0.014
16	0.468	0.464	0.004	0.562	0.5520	0.010
17	0.402	0.394	0.008	0.997	0.9850	0.012
18	0.446	0.422	0.024	-0.957	-0.969	0.012
19	0.530	0.529	0.001	0.858	0.8380	0.020
20	0.681	0.666	0.015	1.473	1.4370	0.036
21	0.696	0.685	0.011	1.279	1.2380	0.041
22	0.920	0.908	0.012	-0.052	-0.111	0.059
23	0.689	0.689	0.000	1.396	1.3490	0.047
24	0.658	0.649	0.009	1.135	1.0800	0.055
25	1.071	1.052	0.019	1.971	1.8710	0.100
26	1.209	1.192	0.017	2.591	2.4690	0.122
27	0.485	0.473	0.012	0.961	0.9060	0.055
28	0.276	0.265	0.011	2.083	2.0480	0.035
29	0.698	0.689	0.009	2.514	2.4260	0.088
30	0.467	0.457	0.010	1.916	1.8500	0.066

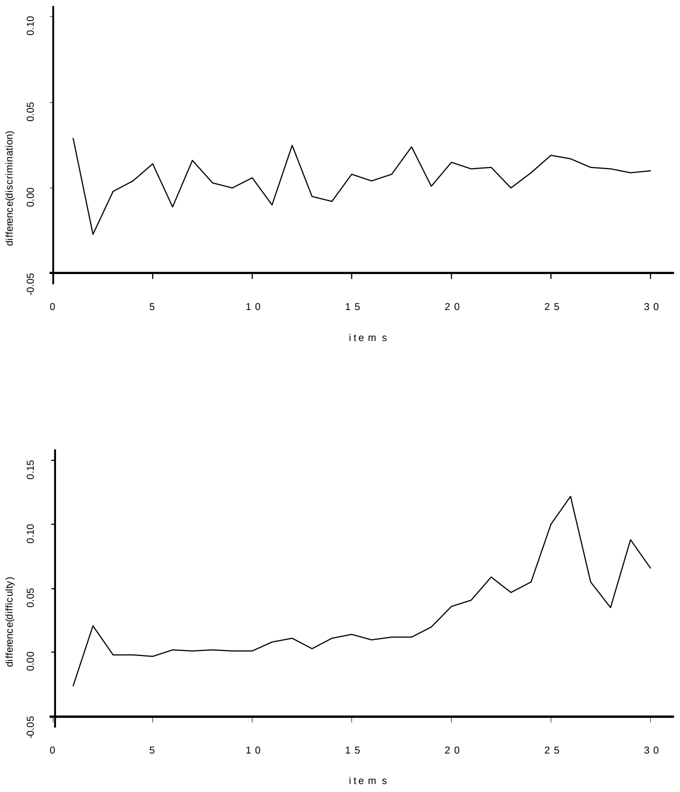

FIGURE 3.1. Plot of the differences of item discrimination and difficulty estimates between MAR and NONMAR models.

The impact of the difference between the MAR and NONMAR models will be studied further by computing the global reliability of the test. Usually, this is done by classical test theory. However, if missing data are present, it is more convenient to compute the global reliability via IRT. In an IRT framework, an index of reliability is based on the identity

$$Var(\theta) = E(Var(\theta|x)) + Var(E(\theta|x))$$

(Bechger, Maris, Verstralen & Béguin, 2003). This identity entails that the total variance of the ability parameters is a sum of two components. The first component, $E(Var(\theta|x))$, relates to the uncertainty about the ability parameter. The posterior variance of ability, $Var(\theta|x)$, gives an indication of the uncertainty with respect to the ability parameter, once we have observed the response pattern x. By considering its expectation over the distribution of x, we obtain an estimate of the average uncertainty over the respondents' ability parameters. The second term, $Var(E(\theta|x))$, is related to the systematic measurement component. The expectation serves as an estimate of ability, and by considering the variance of these expectations over the distribution of x, we get an indication of the extent to which the respondents can be distinguished on the basis of their observed responses. Therefore, a reliability index taking values between zero and one can be computed as the ratio of the systematic variance and the total variance, that is

$$\rho = \frac{Var(E(\theta|x))}{Var(\theta)}. \tag{3.19}$$

In the present application, we can compute $Var(E(\theta|x))$ in two ways: under the MAR assumption where we only condition of the observed responses, and under the NONMAR assumption, where we condition on both the observed responses x and the missing data indicator d. In the latter case, we integrate over both latent variables involved, so the expectation in the numerator of (3.19) is computed as

$$\begin{aligned} E(\theta|x,d) &= \int\int \theta \; p(\theta,\zeta|x,d)d\theta d\zeta \\ &= \int\int \theta \; \frac{p(x|d,\theta)p(d|\zeta)g(\theta,\zeta|\Sigma)}{p(x,d)} d\theta d\zeta. \end{aligned}$$

Under the MAR assumption the global reliability was computed as $\rho = 0.658$, under the NONMAR assumption it was computed as $\rho = 0.738$. So in the present case, taking the missing data process into account leads to a substantial increase in the estimate of global reliability.

3.6 Discussion

The results of the simulation study showed that when an IRT model for the missing data process was included in the estimation together with an IRT model for the observed data, even how much we increased the correlation between latent variables θ and ζ that is, we want to make the missing data mechanism more nonignorable, the bias in the item parameters remained constant and lower compared to the case when the model for the missing data was ignored in the estimation. We conclude that the bias in the IRT parameter estimates is reduce when an IRT model for the missing data process is included in the estimation.

The method as applied in the real speeded test data indicated that it is possible to model the missing data with an IRT model. The results showed that the difference in the item parameters, especially the difficulty parameters (refer to Figure 3.1), gets bigger. This is expected since the respondents were under time limit conditions and the items were getting difficult to endorse. So they skipped items more in the end where item nonresponse were incurred. It was shown that the estimate of the global reliability was larger when the missing data process was taken into account.

For further research, it is of great interest to investigate the effect in the bias of the model parameters estimates when observed covariates are included in the model for the missing data given that the data came from a speeded test. From the results of the previous chapter we can expect that inclusion of the observed covariates in the model for the missing data will support the reduction of the bias in the parameters estimates. It is also further recommended to investigate the effect in the bias of the model parameter estimates when more complex IRT models for the observed scores are used. Further, the concept of using the step model to model speededness needs not be confined to a likelihood based framework. It can also be applied to the complex IRT models that are usually estimated in a Bayesian framework. Examples are models with multiple raters, multiple item types, missing data (Patz & Junker, 1999a,b), models for testlet structures (Bradlow, Wainer & Wang, 1999, Wainer, Bradlow & Du, 2000), and models with a multi-level structure on the ability parameters (Fox & Glas, 2001, 2002, 2003). Implementation of NONMAR models in a Bayesian framework will be the topic of the next two chapters.

4

Detecting Nonignorable Missing Data using the Splitter Item Technique

ABSTRACT: Researchers are often confronted with missing data. Direct statistical inference is appropriate if the missing data are ignorable. In a framework of item response theory, two methods based on the splitter item technique are proposed for deciding whether the missing data are ignorable or nonignorable. In the first method, the observed item response data are split according to the values of the splitter item. Then, the estimated marginal distributions of the item parameters corresponding to both data sets are compared for detecting differences. In the second method, an IRT model for the observed data is extended with group specific item parameters. These extra parameters provide information regarding item parameter differences across groups. They are estimated using MCMC and they do not interfere with the estimation of the other model parameters. In a simulation study concerning item-selection designs, both methods are illustrated and compared using probit IRT models.

KEYWORDS: Ignorability, Item response theory, Markov chain Monte Carlo, Missing data, Splitter item technique.

4.1 Introduction

When data are collected using questionnaire or proficiency items (usually in a sample survey), it is possible that there will be missing observations. For making meaningful inferences it is necessary to find out if it is appropriate to ignore the process that causes the missing data. The missing data process or response mechanism is nonignorable when it depends on a respondent's unobserved response and ignorable when the probability of a nonresponse is independent of the respondent's unobserved response. Bayesian (likelihood) inferences based on the observed data are equivalent to the inferences based on the complete posterior (likelihood) reflecting both the observed data and the response mechanism when the response mechanism is ignorable.

Most of the literature on missing categorical data assumes an ignorable response mechanism. However, handling nonignorable nonresponse is getting more attention. Fay (1986), Baker and Laird (1988), and Green and Park (2003) proposed a class of log-linear models for categorical responses subject to nonignorable nonresponse. In a simulation study, Park and Brown (1994, 1997) showed that it is important to decide whether the underlying response mechanism is ignorable or nonignorable. Lord (1983) was one of the first to develop a mathematical model for omitting behavior when the usual item response theory (IRT) models for dichotomously scored multiple choice items cannot handle appropriately omitted responses. O'Muircheartaigh and Moustaki (1999), and Moustaki and Knott (2000) proposed so-called symmetric pattern models for handling item nonresponse in attitude scales. A set of questions is used to measure some underlying latent attitude or ability but the observed item responses contain missing values. They developed a nonignorable nonresponse model based on a latent basic response propensity that describes the tendency of respondents to respond. The probability of an item response depends on the response propensity value. An individual's response to an item does not depend on its propensity value but only on the value of the individual's latent attitude. Bradlow and Zaslavsky (1999) proposed an IRT model for ordinal customer satisfaction data. The item nonresponse, that might be due to either lack of a strong opinion or indifference about the question, was modeled by a logistic regression model. Bradlow and Thomas (1998) showed in a simulation study that common IRT models cannot be used for likelihood or Bayesian inference when the missing data mechanism cannot be ignored. In this particular case, assuming an ignorable response mechanism leads to bias in parameter estimates. Holman and Glas (2005) proposed several IRT models for modeling nonignorable nonresponse.

A relevant issue when analyzing item response data with missing data concerns the process that causes the missing data. When the response mechanism is ignorable, a statistical analysis based on the observed data always leads to correct inference of the data. When the response mechanism is nonignorable, one can eliminate bias only by constructing a model that correctly represents the response mechanism (Little, 1982). However, such models are highly sensitive to misspecification error and they substantially complicate the statistical inference. Therefore, it is recommendable that first the necessity of such a complex model, that is, a model for the observed data extended with a missing-data model, is verified. On the other hand, most of the literature on missing data for categorical problems assumes that the process that caused the missing data can be ignored. In these cases, the assumptions for ignorability should be checked.

In the present paper, two methods will be proposed to verify whether the missing data mechanism can be ignored or not in case of item nonresponse. It will be assumed that the probability of the observed pattern of missing data may be depending on possible values of the missing data and/or the parameters of the data and the parameters of the missing data may not be distinct. No other variables relate to the item score missingness. Both cases lead to a nonignorable response mechanism (Rubin, 1976). In the first method, the splitter item technique (Molenaar 1983; Van den Wollenberg, 1979) is used for splitting the data in two groups depending whether the response item was observed or missing. Then, the marginal posterior distributions of the item parameters corresponding to both groups are compared. In the second method, an IRT model for the observed data is extended such that item parameters may fluctuate across groups. The extra parameters in this more general model are called Bayesian modification indices (Fox and Glas, 2005) and provide information about the relevance of the model extension. In this particular case, they are used to test whether the response mechanism is nonignorable. The parameters of the IRT model for the observed data are estimated using MCMC (Gelfand and Smith, 1990). The BMI values are sampled given the sampled values of the IRT parameters. However, these extra draws do not influence the Markov chain and the chain remains restricted to the manifold of the posterior of the IRT model. It will be shown that the estimated marginal posterior distribution of the BMI values are closely related to their true marginal posterior distribution. As a result, BMI values are sampled as by-products of the MCMC procedure for estimating the parameters of the IRT model for the observed data. The MCMC estimation procedure can be time-consuming and it is, therefore, preferable to compute certain fit statistics during the estimation of the model parameters.

In the next section, a general notation is given for IRT models for the observed data and models for the missing data process. Models for the missing data process are introduced to illustrate the aspect of distinctness. Then, the splitter item method will be described and details will be given of both methods for testing whether the missing data process can be safely ignored. Next, both methods will be applied in three experiments concerning item selection with artificial data. Finally, the last section contains a discussion and suggestion for further research.

4.2 Model and Notation

4.2.1 IRT model for the observed data

The categorical outcome, y_{ik}, represents the item response of person i $(i = 1, \ldots, N)$ on item k $(k = 1, \ldots, K)$. These item responses may be dichotomous or polytomous. Let θ_i denote the latent abilities or attitudes of the respondents responding to the K items. They are collected in the latent vector $\boldsymbol{\theta}$. For dichotomous item responses a two-parameter IRT model is used for specifying the relation between the examinee level on a latent variable and the probability of a particular item response. That is

$$P\big(y_{ik} = 1 \mid \theta_i, a_k, b_k\big) = \Phi\big(a_k\theta_i - b_k\big), \tag{4.1}$$

where a_k is the item discrimination parameter, and b_k is the item difficulty parameter. The item parameters will also be denoted by $\boldsymbol{\xi}_k$, with $\boldsymbol{\xi}_k = (a_k, b_k)$. The function Φ is the cumulative standard normal distribution. For polytomous item responses, the probability that an individual obtains a grade c $(c = 1, \ldots, C)$ on item k is defined by a graded response model (GRM)

$$P\big(y_{ik} = c \mid \theta_i, a_k, \boldsymbol{\kappa}_k\big) = \Phi\big(a_k\theta_i - \kappa_{kc-1}\big) - \Phi\big(a_k\theta_i - \kappa_{kc}\big) \tag{4.2}$$

where the boundaries between the response categories are represented by an ordered vector of thresholds $\boldsymbol{\kappa}$ such that $\kappa_{kr} > \kappa_{ks}$ whenever $r > s$, with $\kappa_{k0} = -\infty$ and $\kappa_{kC} = \infty$. In this case let $\boldsymbol{\xi}_k = (a_k, \boldsymbol{\kappa}_k)$. Consequently, there are a total of $C - 1$ threshold parameters and one discrimination parameter for each item.

4.2.2 A latent variable model for the missing data process

The data matrix of the observed data is partitioned into two parts, the observed part $\mathbf{y}_{obs}$ and the missing part $\mathbf{y}_{mis}$. The pattern of the missing data is given by a matrix $\mathbf{d}$, of the same dimension as $\mathbf{y}$ and equals one when an item is observed and zero otherwise. Although the proposed techniques can be applied to any missing data process, it is assumed that the binary responses (response, nonresponse) are indicators of an underlying latent variable $\boldsymbol{\zeta}$, which represents the tendency to respond (see, for example, Holman & Glas, 2005; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999). The actual response y_{ik} itself depends on the individual's attitude level but the probability of a response depends on the individual's response propensity. The nonresponse or missing observations may include unit non-response, where a respondent does not respond to any of the items, and item non-response where the respondent does respond to some but not all of the items. Attention will be focused on item nonresponse although unit nonresponse is a more serious problem.

The occurrence of missing data is viewed as a random phenomenon. That is, the occurrence of missing data, in terms of item nonresponse of persons responding to an item, is governed by a random process that caused the missingness. Let $\boldsymbol{\zeta}, \boldsymbol{\phi}$ denote the parameters of the missing-data process, where $\boldsymbol{\zeta}$ are person parameters and $\boldsymbol{\phi}$ item parameters. Then, $p(\mathbf{d} \mid \boldsymbol{\zeta}, \boldsymbol{\phi})$ represents the latent variable model for the missing data mechanism. This latent variable model may exists of one or more factors, and can be defined as a confirmatory factor model or an item response theory model. That is, it will be assumed that the pattern of missing data are represented as a function of one or more latent variables. In the present paper, attention is focused on a nonignorable response model by allowing the attitude parameter $\boldsymbol{\theta}$ to affect the probability of responding. So, $p(\mathbf{d} \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\phi})$ is a nonignorable response model. An individual response depends on both the individual's ability or attitude and propensity to respond. For example, examinees with high math abilities may have a higher probability of responding to a math item than examinees with low math abilities. Further, the probability of a missing response depends on an attitude as well as a personality trait (Holman & Glas, 2005), or when measuring customer satisfaction, nonresponse is related to the latent opinion, since a nonresponse indicates a lack of knowledge or interest (Bradlow & Zaslavsky, 1999).

4.3 Detecting a Nonignorable Missing Data Mechanism

The complete data-likelihood of $(\mathbf{y}_{obs}, \mathbf{d})$, given the model parameters can be factorized as

$$p\big(\mathbf{y}_{obs}, \mathbf{d} \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\phi}\big) = p\big(\mathbf{y}_{obs} \mid \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\xi}\big)p\big(\mathbf{d} \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\phi}\big), \tag{4.3}$$

where it is assumed that the missing data are missing at random. Inferences for $(\boldsymbol{\xi}, \boldsymbol{\phi})$ are based on this joint distribution combined with the priors for the model parameters. It is a priori assumed that the attitudes or abilities underlying the observed responses are independent of the propensity to respond. Therefore, the prior for each latent variable is a standard normal distribution. This way the mixture of models, the model for the observed responses and the missing data mechanism, is identified. Both parameters $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ have proper noninformative priors. It follows that the marginal posterior distribution of the item parameters $\boldsymbol{\xi}$ can be specified as,

$$\begin{aligned} p\big(\boldsymbol{\xi} \mid \mathbf{y}_{obs}, \mathbf{d}\big) \propto \int\int\int p\big(\mathbf{y}_{obs} \mid \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\xi}\big)p\big(\mathbf{d} \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\phi}\big) \times \\ p(\boldsymbol{\theta})p(\boldsymbol{\zeta})p(\boldsymbol{\xi})p(\boldsymbol{\phi})\, d\boldsymbol{\phi}\, d\boldsymbol{\zeta}\, d\boldsymbol{\theta} \end{aligned} \tag{4.4}$$

When the $\boldsymbol{\theta}$ values do not interfere with the probability of responding, inferences about $\boldsymbol{\xi}$, ignoring the process that causes the missing data, is appropriate. In this particular case, the missing data mechanism is ignorable, also assuming that the missing data are missing at random, and equation (4.4) simplifies to

$$p\big(\boldsymbol{\xi} \mid \mathbf{y}_{obs}, \mathbf{d}\big) \propto \int p\big(\mathbf{y}_{obs} \mid \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\xi}\big)p(\boldsymbol{\theta})p(\boldsymbol{\xi})d\boldsymbol{\theta}. \tag{4.5}$$

As a result, inferences based on the distribution (4.5) are equivalent to inferences based on the full distribution, see the right-hand side of Equation (4.4).

4.3.1 The splitter item technique

In a rich-data situation the data can be split in two samples according to the scores of one item, the splitter item. This method of splitting the data is quite common and can be used for general problems like model assessment and selection. In this particular case, it is of interest to test whether the manner of splitting affects the statistical inference. That is, the splitter item technique is used to test whether the process that causes the missing data can be ignored.

Two samples are obtained when the observed item response data are divided according to the scores on a particular item k. In Figure 4.1 the splitting of the data $(\mathbf{y}_{obs}, \mathbf{d})$ is given in a diagram. The missing data indicator of item k splits the data in two samples. The first sample, denoted as the observed group; the observed item responses of individual i except those to item k, $\mathbf{y}_{i,obs}^{(-k)}$ with $d_{ik} = 1$, $i = 1, \ldots, n$. The second sample, denoted as the missing group: the observed item responses of individual i except those to item k, $\mathbf{y}_{i,mis}^{(-k)}$ with $d_{ik} = 0$, $i = 1, \ldots, n$.

The occurrence of missing data patterns is modeled by a latent variable model. As a result, this splitting of the data in two groups depends on the values for $\boldsymbol{\zeta}$, $\boldsymbol{\theta}$ and some model parameters $\boldsymbol{\phi}$. After splitting the data, the marginal posterior distributions of the item parameters for the observed data are $p_{mis}\big(\boldsymbol{\xi} \mid \mathbf{y}_{obs}, \mathbf{d}^{(-k)}, d_k = 0\big)$ and $p_{obs}\big(\boldsymbol{\xi} \mid \mathbf{y}_{obs}, \mathbf{d}^{(-k)}, d_k = 1\big)$ for the missing and observed group, respectively. It follows that the statistical inferences derived from these posterior distributions are different when $\boldsymbol{\theta}$ affects the probability of responding to an item and accordingly influences the way the data are divided.

For example, assume that students were permitted to choose a subset of items of varying difficulty and that the better students choose the easier items, since they can better decide which items are easier (Bradlow & Thomas, 1998). In this case, $\boldsymbol{\theta}$ represent the individuals' math abilities, and the data are divided in a high level and low level group. The weaker students may select the harder items and make them appear even more difficult. So, the estimates of the corresponding difficulty parameter are biased. The better students select the more easier items and make them appear even more easier. Analogous, biased estimates of the difficulty parameter are obtained.

In summary, the splitter item technique is used to detect a nonignorable missing data mechanism by comparing the marginal posterior distributions of the item parameters given both samples. Results can be compared using summary statistics for the sampled values from the marginal posterior distributions, such as the posterior mean or median.

4.3.2 BMI based on the splitter item technique

BMI were introduced by Fox and Glas (2005) for detecting model violations of the 2PNO model, differential item functioning and violations of the assumptions of local independence. They extended the 2PNO model with extra parameters such that the assumption to be tested is violated. An indication of a model violation is found when the estimated extra parameters are significantly different from zero. The marginal distribution of the extra parameters is unknown but samples from a so-called Bayesian modification (BM) distribution, that is a good approximation of the true marginal posterior distribution, can be obtained from an extra sampling step in an MCMC algorithm for sampling the parameters of the 2PNO model. BMI are useful when the number of model violations are large, or when estimating the parameters of the more general models is difficult and/or time-consuming.

The relation between the categorical outcomes and the underlying latent variables, equation (4.1) and (4.2), can also be explained in terms of a random variable z_{ik} with mean $a_k\theta_i - b_k$ or $a_k\theta_i$ for binary or polytomous outcomes, respectively, and variance 1 (see, Albert, 1992; Johnson & Albert, 1999). The binary response y_{ik} is the indicator of z_{ik} being positive, and the polytomous response can be viewed as an indicator of z_{ik} falling into one of the line segment associated with response categories. It follows that,

$$\begin{aligned} z_{ik} &= a_k\theta_i - b_k + e_{ik} \text{ such that } \begin{cases} y_{ik} = 0 \leftrightarrow z_{ik} \leq 0 \\ y_{ik} = 1 \leftrightarrow z_{ik} > 0 \end{cases} \\ z_{ik} &= a_k\theta_i + e_{ik} \text{ such that } y_{ik} = c \leftrightarrow \kappa_{kc-1} < z_{ik} < \kappa_{kc}, \end{aligned} \tag{4.6}$$

where $-\infty = \kappa_{k0} < \kappa_{k1} < \ldots < \kappa_{kC} = \infty$, and z_{ik} normally distributed. The resulting model is the ordinal probit model for dichotomous or polytomous data, respectively. This ordinal probit model will be considered as the null-model.

The data are divided, according to the values of splitter item, in an observed ($j = 1$) and a missing group ($j = 0$). In case of nonignorable nonresponses, the item parameter estimates given the item responses of the observed group will differ from the item parameter estimates given the item responses of the missing group. This item bias can be modeled by extending the null model, equation (4.6), with fixed group effects that represent the item bias due to nonignorable nonresponses since the grouping of the data is based upon the values (observed/missing) of the splitter item. Extending the two-parameter model for dichotomous data leads to

$$z_{ijk} = a_k\theta_{ij} - b_k + \big(\lambda_{1jk}\omega_{ij} + \lambda_{2jk}\big) + e_{ijk}, \tag{4.7}$$

where $j = 0, 1$, and ω_{ij} is an explanatory variable which might be θ or an observed covariate either within the test (test score) or outside the test. The magnitude of the extra parameters, $\boldsymbol{\lambda}_{jk} = (\lambda_{1jk}, \lambda_{2jk})$, depends on the extent to which the difference z_{ijk} and $a_k\theta_{ij} - b_k$ is properly modeled. It follows that the response mechanism is not ignorable when the estimated group effects are significantly different from zero. Note that item bias in the discrimination parameter is modeled when one of the explanatory variables is a function of $\boldsymbol{\theta}$.

This fixed effects model in (4.7) can be written as a linear regression model using an indicator variable $\mathbf{x}_2$. The ith case of $\mathbf{x}_2$ equals one when the ith case of the splitter item is observed and zero otherwise or vice versa. In this case, the fixed effects model is identified by fixing one of the two group effects to zero, and the subscript j of parameter $\boldsymbol{\lambda}$ can be dropped. The multiple regression model then is as follows:

$$\mathbf{z}_k = a_k\boldsymbol{\theta} - b_k + \mathbf{x}_2\big(\lambda_{1k}\boldsymbol{\omega} + \lambda_{2k}\big) + \mathbf{e}_k, \tag{4.8}$$

where the $\mathbf{z}_k$ and $\boldsymbol{\theta}$ are the augmented responses and latent attitudes or abilities of all respondents, respectively. In the same way the ordinal probit model, equation (4.6), can be extended to handle item bias. It follows that,

$$\mathbf{z}_k = a_k\boldsymbol{\theta} + \mathbf{x}_2\big(\lambda_{1k}\boldsymbol{\omega} + \lambda_{2k}\big) + \mathbf{e}_k, \tag{4.9}$$

where $\mathbf{e}_k$ are independent standard normally distributed. As for the two-parameter model, item bias in the discrimination parameter is modeled when one of the explanatory variables is a function of $\boldsymbol{\theta}$. Item bias in the threshold parameters is modeled by λ_{2k} since it allows thresholds to vary across groups. In the simplest case,

$$\begin{aligned} P\big(z_{ijk} \leq \kappa_{kc} \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k, \lambda_2\big) &= \Phi\big(\kappa_{kc} - (a_k\theta_{ij} + \lambda_{2k})\big) \\ &= \Phi\big((\kappa_{kc} - \lambda_{2k}) - a_k\theta_{ij}\big). \end{aligned} \tag{4.10}$$

As a result, in group j the original thresholds for item k, $\boldsymbol{\kappa}_k$, are simultaneously shifted yielding the thresholds $\boldsymbol{\kappa}_k + \lambda_{2k}$. Thus, the effective thresholds vary across groups when the missing data mechanism cannot be ignored. The thresholds for item k, $\boldsymbol{\kappa}_k$, represent the average across groups.

4.4 Bayesian Estimation

Bayesian inference typically requires the computation of the posterior distribution for a collection of random variables (parameters or unknown observables). Therefore, numerous simulation-based methods have been developed and implemented within the Bayesian paradigm, e.g. importance sampling (Chen, Shao, & Ibrahim, 2000; Ripley, 1987), and Markov Chains Monte Carlo (MCMC) algorithms (see for e.g., Robert & Casella 1999; Gelfand & Smith, 1990; Gelman, Carlin, Stern & Rubin, 2004). In specific, MCMC procedures for sampling the parameters of logistic and probit IRT models were formulated by, among others, Albert (1992), Fox and Glas (2001, 2003), Hendrawan (2004), Johnson and Albert (1999), Maris and Maris (2002), and Patz and Junker (1999a, 1999b). A wide range of MCMC algorithms were developed for other latent variable models (e.g. Casella & Robert 1999; Congdon, 2002). More specific, Bradlow and Zaslavsky (1999) developed different MCMC schemes for sampling the parameters of a latent variable model for a missing data mechanism.

When employing the splitter item technique, and modeling both observed data sets, all parameters are sampled using MCMC. This way, values are sampled from $p_{mis}\big(\boldsymbol{\xi} \mid \mathbf{y}_{obs}, \mathbf{d}^{(-k)}, d_k = 0\big)$ and $p_{obs}\big(\boldsymbol{\xi} \mid \mathbf{y}_{obs}, \mathbf{d}^{(-k)}, d_k = 1\big)$ by sampling from their full conditionals. These group specific sampling steps are easily derived from the general procedure for sampling item parameter values. For example, in case of a normal ogive model for binary item responses with splitter item K. Let t denote the iteration number of the Markov chain.

Algorithm 1

Sample augmented data $\mathbf{z}$*, for* $i = 1, \ldots, n_j, k = 1, \ldots, K-1$ *and* $j = 0, 1$*:*

$$z_{ijk}^{(t)} \mid \mathbf{y}_{obs}, \mathbf{d}^{(-k)}, \theta_i^{(t-1)}, \boldsymbol{\xi}_{jk}^{(t-1)} \sim \begin{cases} \mathcal{N}\big(a_{jk}\theta_i - b_{jk}, 1\big), \\ \quad \textit{if } y_{ijk} \textit{ is missing} \\ \mathcal{N}\big(a_{jk}\theta_i - b_{jk}, 1\big) I(z_{ijk} > 0), \\ \quad \textit{if } y_{ijk} = 1 \\ \mathcal{N}\big(a_{jk}\theta_i - b_{jk}, 1\big) I(z_{ijk} < 0), \\ \quad \textit{if } y_{ijk} = 0 \end{cases}$$

Sample latent parameters $\boldsymbol{\theta}$, for $i = 1, \ldots, n_j$, and $j = 0, 1$:

$$\theta_{ij}^{(t)} \mid \mathbf{z}_j^{(t)}, \boldsymbol{\xi}_j^{(t-1)} \sim \mathcal{N}\big((\mathbf{a}_j^t \mathbf{a}_j)^{-1} \mathbf{a}_j^t (\mathbf{z}_{ij} + \mathbf{b}_j), (\mathbf{a}_j^t \mathbf{a}_j)^{-1}\big) p(\theta_{ij})$$

where $\mathbf{a}_j = (a_{j1}, \ldots, a_{jK-1})$.

Sample item parameters $\boldsymbol{\xi}_{jk}$, for $k = 1, \ldots, K-1$ and $j = 0, 1$:

$$\boldsymbol{\xi}_{jk}^{(t)} \mid \mathbf{z}_{jk}^{(t)}, \boldsymbol{\theta}_j^{(t)} \sim \mathcal{N}\Big(\hat{\boldsymbol{\xi}}_{jk}, \big(\mathbf{H}_j^t \mathbf{H}_j\big)^{-1}\Big) p(\boldsymbol{\xi}_{jk})$$

where $\mathbf{H}_j = (\boldsymbol{\theta}_j, \mathbf{1}_{n_j})$ and $\hat{\boldsymbol{\xi}}_{jk} = \big(\mathbf{H}_j^t \mathbf{H}_j\big)^{-1} \mathbf{H}_j^t \mathbf{z}_{jk}$.

In summary, values from the marginal distribution of the group specific item parameter estimates are easily obtained using MCMC when modeling the observed data. Then, summaries of these marginal posterior distributions can be used to decide whether the missing data process can be safely ignored.

4.4.1 Sampling BMI parameters

The BMI values are sampled as an extra step in an MCMC algorithm for estimating the null-model parameters. This way, sampled values of the BM distribution are obtained when estimating the parameters of the null-model. The extra step in the MCMC algorithm consists of sampling values of $\boldsymbol{\lambda}$ given sampled values of the IRT null-model parameters. These extra draws do not influence the chain, and the Markov chain remains restricted to the manifold of the posterior corresponding to the null-model. It will be shown that the resulting estimate of the marginal posterior of $\boldsymbol{\lambda}$ is a good approximation of the true marginal posterior distribution.

A general model that includes the models in equation (4.8) and (4.9) for the observed data, with the extra BMI parameters $\boldsymbol{\lambda}$ for modeling the item bias can be written as,

$$\mathbf{z}_k = \mathbf{x}_1 \boldsymbol{\gamma}_k + \mathbf{x}_2 \boldsymbol{\lambda}_k + \mathbf{e}_k, \tag{4.11}$$

where $\mathbf{x}_1$ is an $n \times r$ and $\mathbf{x}_2$ an $n \times 2$ matrix with rank r and 2, respectively, and $\mathbf{e}_k$ are normally distributed with variance σ^2. Note that $\mathbf{x}_1 = \big(\boldsymbol{\theta}, -\mathbf{1}\big)$ with $r = 2$ and $\boldsymbol{\gamma}_k = \big(a_k, b_k\big)$ for binomial observed data and $\mathbf{x}_1 = \boldsymbol{\theta}$ with $r = 1$ and $\boldsymbol{\gamma}_k = a_k$ for ordinal polytomous data. The full conditional distribution of the BMI parameters can be specified explicitly. Therefore, define the least squares estimate of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\lambda}_k$ as

$$\begin{pmatrix}\hat{\boldsymbol{\gamma}}_k\\ \hat{\boldsymbol{\lambda}}_k\end{pmatrix} = \begin{pmatrix}\mathbf{x}_1^t\mathbf{x}_1 & \mathbf{x}_1^t\mathbf{x}_2\\ \mathbf{x}_2^t\mathbf{x}_1 & \mathbf{x}_2^t\mathbf{x}_2\end{pmatrix}^{-1}\mathbf{x}^t\mathbf{z}_k = \begin{pmatrix}\mathbf{v}_{11} & \mathbf{v}_{12}\\ \mathbf{v}_{21} & \mathbf{v}_{22}\end{pmatrix}\mathbf{x}^t\mathbf{z}_k = \mathbf{v}\mathbf{x}^t\mathbf{z}_k,$$

with $\mathbf{v} = (\mathbf{x}^t\mathbf{x})^{-1}$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$. It follows from linear regression theory (see, e.g., Box & Tiao, 1973, p. 116-118) that the true marginal distribution of $\boldsymbol{\lambda}_k$ given $\mathbf{z}_k$ according to the full model is a bivariate t-distribution $t_2\left[\hat{\boldsymbol{\lambda}}_k, s^2\mathbf{v}_{22}, n-(r+2)\right]$, where $r+2$ equals the dimension of $\left(\boldsymbol{\gamma}_k, \boldsymbol{\lambda}_k\right)$ and s^2 is an estimate of the residual variance of the model in equation (4.11) using a noninformative reference prior for $\boldsymbol{\lambda}_k$ and $\boldsymbol{\gamma}_k$.

Let $\mathbf{z}$ and $\boldsymbol{\theta}$ be given. Then, the null-model is given by

$$\mathbf{z}_k = \mathbf{x}_1\boldsymbol{\gamma}_k + \mathbf{e}_k, \tag{4.12}$$

where $\mathbf{e}_k$ are normally distributed with variance σ_0^2. An MCMC algorithm is extended by sampling BMI values of $\boldsymbol{\lambda}$. Let t denote the iteration number of the Markov chain.

Algorithm 2

sample $\boldsymbol{\gamma}_k^{(t)} \mid \mathbf{z}_k, \sigma_0^{2^{(t-1)}}$

sample $\sigma_0^{2^{(t)}} \mid \mathbf{z}_k, \boldsymbol{\gamma}_k^{(t)}$

sample $\boldsymbol{\lambda}_k^{(t)} \mid \mathbf{z}_k, \boldsymbol{\gamma}_k^{(t)}, \sigma_0^{2^{(t)}}$

Notice that the sampled values of $\boldsymbol{\lambda}_k$ do not interfere with the sampling of the other parameters. Further, the grouping index j only applies for sampling BMI values. It is assumed that the elements of $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}_k$, and $\log(\sigma_0^2)$ are uniformly and independently distributed. Then, the full conditional distribution of $\boldsymbol{\gamma}_k$ equals

$$\boldsymbol{\gamma}_k \mid \mathbf{z}_k, \mathbf{x}_1, \sigma_0^2 \sim \mathcal{N}\left(\hat{\boldsymbol{\gamma}}_{k,\mathbf{x_1}}, \sigma_0^2\mathbf{v}_{11}\right) \tag{4.13}$$

where $\hat{\boldsymbol{\gamma}}_{k,\mathbf{x_1}}$ is the least squares estimate of $\boldsymbol{\gamma}_k$ given $\mathbf{x_1}$ and $\mathbf{z}$. Further, the full conditional of σ_0^2 equals

$$\sigma_0^{2^{(t)}} \mid \mathbf{z}_k, \boldsymbol{\gamma}_k \sim \mathcal{IG}\Big(\frac{n}{2}, \sum_i \Big(z_{ik} - \mathbf{x}_1\boldsymbol{\gamma}_k^{(t)}\Big)^2/2\Big). \tag{4.14}$$

Accordingly, as an extra MCMC step values of the BMI parameters are sampled from the full conditional distribution,

$$\boldsymbol{\lambda}_k \mid \mathbf{z}_k, \mathbf{x}, \boldsymbol{\gamma}_k, \sigma_0^2 \sim \mathcal{N}\Big(\hat{\boldsymbol{\lambda}}_k + \mathbf{v}_{12}^t \mathbf{v}_{11}^{-1} \big(\gamma_k - \hat{\gamma}_{k,\mathbf{x}_1}\big), \sigma_0^2 \big(\mathbf{v}_{22} - \mathbf{v}_{12}^t \mathbf{v}_{11}^{-1} \mathbf{v}_{12}\big)\Big). \tag{4.15}$$

The Bayesian Modification (BM) distribution is obtained by integrating the conditional distribution of $\boldsymbol{\lambda}_k$ with respect to the null model parameters using MCMC. That is, the MCMC algorithm is used for obtaining sampled values from the Bayesian Modification (BM) distribution. Let $\tilde{p}(.)$ denote the BM distribution, and $\big\{\big(\boldsymbol{\gamma}_k^{(m)}, \sigma_0^{2(m)}\big), m = 1, \ldots, M\big\}$ an MCMC sample from the joint posterior distribution $p\big(\boldsymbol{\gamma}_k, \sigma_0^2 \mid \mathbf{z}\big)$. It follows that

$$\begin{aligned}\tilde{p}\big(\boldsymbol{\lambda}_k \mid \mathbf{z}_k, \mathbf{x}\big) &= \int_0^\infty \int_\Omega \tilde{p}\big(\boldsymbol{\lambda}_k \mid \mathbf{z}_k, \mathbf{x}, \boldsymbol{\gamma}_k, \sigma_0^2\big) p\big(\boldsymbol{\gamma}_k, \sigma_0^2 \mid \mathbf{z}_k, \mathbf{x}_1\big) d\boldsymbol{\gamma}_k d\sigma_0^2 \\ &= \lim_{M \to \infty} 1/M \sum_{m=1}^{M} \tilde{p}\Big(\boldsymbol{\lambda}_k \mid \mathbf{z}_k, \mathbf{x}, \boldsymbol{\gamma}_k^{(m)}, \sigma_0^{2(m)}\Big),\end{aligned} \tag{4.16}$$

where $\Omega = \{\boldsymbol{\gamma}_k \in \mathbb{R}^r\}$. In the Appendix it is shown that the marginal BM distribution in equation (4.16) is the bivariate t-distribution $t_2\big[\hat{\boldsymbol{\lambda}}_k, s_0^2 \mathbf{v}_{22}, n - r\big]$, using a noninformative reference prior for $\boldsymbol{\lambda}_k$ and $\boldsymbol{\gamma}_k$. Here, s_0^2 is an estimate of the residual variance of the null-model in equation (4.12). As a result, the marginal BM distribution approximates the true marginal posterior distribution very good. It is not expected that the residual variance, s_0^2, differs much from the larger residual variance s^2 since the measurement null-model contains all relevant (latent) variables.

4.5 Simulated Examples

Bradlow and Thomas (1998) analyzed simulated response data from an examination that allowed students to choose a subset of items. In this choice-based examination, a subset of items was presented in pairs of items and the examinees choose to respond to one of them. No responses are given to those items that are not selected. The choice mechanism can only be ignored when the examinees randomly select items. Recently, the fairness of item-selection has been further investigated by Allen, Holland, and Thayer (2005).

The two experiments by Bradlow and Thomas (1998) are used to demonstrate the splitter item technique for detecting nonignorable missing data. In both experiments, 5000 abilities and 20 difficulty parameter values were generated from a standard normal distribution. These parameter values were used to generate item response data according the Rasch model. The last ten items were considered as paired items. In a third experiment polytomous IRT data were generated according the ordinal probit IRT model. A response mechanism simulated that respondents with a negative strong opinion are more likely to give a nonresponse. This corresponds with well-known cases where respondents refuse to give a socially undesirable answer.

In all three experiments below, Gibbs sampling algorithms were used for estimating the IRT models for binary and polytomous data. For each model, 20,000 iterations were used for estimating the model parameters with a burn-in period of 1,000 iterations. Convergence of the Markov chains was easily established using plots of sampled values and using several convergence diagnostics (Gelman et. al., 2004). For further details regarding the sampling procedure refer to Albert (1992) and Johnson and Albert (1999). All IRT models were identified by fixing the scale of the posterior distribution of the latent variable with mean zero and variance one.

4.5.1 Experiment 1

In the first experiment, the examinees with positive abilities chose the easier items within pairs with probability $p_1 = .95$ and the harder item with probability $p_2 = .05$. These examinees can often decide which of the two items is easier. The examinees with negative abilities chose one of the items randomly. So, the distribution of the response mechanism depends on the ability parameters $\boldsymbol{\theta}$ underlying the observed responses and the difficulty parameters. As a result, the missing data are missing at random but the parameters of the missing data process, $\boldsymbol{\phi}$, are not a priori distinct from $\boldsymbol{\theta}$. Inferences about $\boldsymbol{\theta}$ cannot be based on the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}_{obs})$ ignoring the missing data mechanism.

The BMI parameters $\boldsymbol{\lambda}$ were defined according to equation (4.8) with indicator variable $\mathbf{x}_2$ equal to one if the corresponding value for the splitter item was missing and zero otherwise. These BMI parameters represent item bias for the so-called missing group, the set of observed item response data with nonresponses for the splitter item. According to missing data mechanism it was expected that the item difficulties varied over groups since the distribution of abilities varied across groups.

In Table 4.1 are the posterior means and standard deviations given of the difficulty parameters for different subsets of item response data. The posterior means of $p\big(\mathbf{b} \mid \mathbf{y}_{obs}\big)$ correspond to the estimated item difficulties given all observed item response data. It can been seen that for each paired item the difficult item is overestimated and the easy item is underestimated. In most cases the better students choose the easier item and make them appear even more easier. This is in contrast to the harder items. These items were selected by the weaker students and they make them appear even more difficult. The observed item response data were grouped according to the values of item 20. This splitter item was paired with item 19, item 19 was the easier item. As a result, most of the better students chose to make item 19 and they are denoted as the missing group. Most of the weaker students chose to make the more difficult item 20, denoted as the observed group. It can be seen that the true item difficulties are highly overestimated by the posterior means of $p_{obs}\big(\mathbf{b} \mid \mathbf{y}_{obs}, \mathbf{d}_{20} = \mathbf{1}\big)$ since the item responses of the observed group correspond to the weaker students. Subsequently, the posterior means of $p_{mis}\big(\mathbf{b} \mid \mathbf{y}_{obs}, \mathbf{d}_{20} = \mathbf{0}\big)$ are lower than the true item difficulties. The difficulty parameter of item 19 could not be estimated since the students of the observed group did not respond to item 19. It can be concluded that the pattern of missing data is affected by the abilities of the students. That is, the examinee's propensity to respond correlates with their ability and this results in nonignorable missing data.

The BMI values are sampled under the null-model for the missing group. The posterior means are all negative indicating that the students in the missing group make the items appear more easier. The estimated BMI values cannot capture the difference between the estimates given all data and the estimates given only the item responses of the missing group. This follows from the fact that the other model parameters are estimated under the null model. However, the estimated BMI values are for most items significant given the 95% highest posterior density intervals (HPD), that is, the value zero was not contained in the HPD region. The grouping of the data according to splitter item 20 resulted in various significant fixed group effects indicating that the way of grouping the data (observed/missing) affects the results. The set of responses to the last ten items contains missing values that caused a reduction in size of the estimated BMI values.

The choice of the splitter item may affect the results. The last ten items can be considered as potential splitter items. All BMI values are sampled in one MCMC algorithm for estimating the null model for each possible splitter item. This procedure requires only extra sampling steps. In Figure 4.2, the posterior distributions of the BMI's are given for splitter item 12, 17, and 20. These items are all easy items in the pairs of items. It can be seen that the results are not depending on the splitter item since the estimated BMI values and their distributions are comparable.

4.5.2 Experiment 2

In the second simulation, each examinee first responded to both items of a pair, and, contingent on the responses, then submitted only one response and left the other response missing. This was done according to the following rule. For each of the paired items, each examinee chose randomly one of the items when both responses were either correct or incorrect. If one of the responses within a pair was correct and the other one incorrect, then the examinee chose the correct one with probability $p_1 = .75$ and the incorrect one with probability $1 - p_1 = .25$. The response mechanism cannot be ignored since the missing data are not missing at random. That is, the distribution of the missing data mechanism depends on the missing values.

In the second and third column of Table 4.2, the true and estimated item difficulties are given using all item response data. As expected, the item difficulties of the paired items are underestimated. This follows from the fact that students are tended to select items when they knew the correct answer. The last item was used as a splitter item and the data were grouped in an observed and a missing group. However, if examinees responded to item 20 they did not respond to item 19, and the other way around. As a result, the splitter item technique does not provide any additional information since the responses to item 19 of the missing group correspond with all observed responses to item 19, and the observed group did not respond to item 19. The patterns of missing data corresponding to a pair of items are only depending on the values of the missing item responses to this pair of items. Therefore, it was not expected to detect any differences between the estimated item difficulties corresponding to the grouped data.

In Table 4.2 it can been seen that the estimated posterior means based on the item responses of the observed group (fifth column) are slightly higher than the estimated posterior means based on the item responses of the missing group (seventh column). Because item 20 is much easier than item 19, the observed group consists of weaker students. They knew the correct answer to item 20 but not to item 19. However, this effect is in most cases very small and not significant given the posterior standard deviations. This is supported by the estimated BMI's, defined as in experiment 1. These estimated BMI values were around zero and not significant.

4.5.3 Experiment 3

In collecting data via surveys it is often assumed that the respondents are willing to cooperate and respond honestly to the survey questions. In case of sensitive topics, respondents may be tended to provide more socially desirable answers or provide nonresponses. In the same way, lack of a strong opinion or indifference can also lead to nonresponses (see, e.g., Baker and laird, 1988; Bradlow and Zaslavsky, 1999). De Leeuw, Hox, and Huisman (2003), and Sijtsma and van der Ark (2003) discuss several types of missing item scores. Item response data were generated according to the ordinal probit model with three response categories and discrimination parameters set to one. The attitude parameters were generated from a standard normal distribution. The ordered threshold parameters were generated from an uniform distribution restricted to the interval $[-.75, -.50] \cup [.50, .75]$. For the last ten items out of twenty items, nonresponses were generated. Respondents scoring in the lowest category had a probability of 40% of a nonresponse and others 20%. As a result, the missing data are not missing at random since the distribution of the missing data mechanism depends on the missing values. Respondents with low attitude values provided responses in the lowest category, and, subsequently, they had a higher propensity to give a nonresponse than respondents with high attitude values. The willingness to give a (negative, mild, positive) response was correlated with the respondents' attitude being measured, that is, the attitude parameter was correlated with the parameter of the missing data mechanism.

The last item was used as a splitter item, and the data were split in two groups, a missing and an observed group. The item parameters were estimated using MCMC given all observed data and the group specific item response data. In Figure 4.3 are the threshold estimates given corresponding to the full and partitioned data set. It can be seen that the true threshold values are underestimated when using all observed data. About 25% of the item response data is missing and about 44% of the missing item responses were scores in the lowest category causing an underestimation of the true threshold values. It appears as if respondents score relatively often in a second or third category. However, respondents, who were inclined to score in the first category, more often refused to give an answer. It follows that the estimated threshold parameters given the item responses of the missing group are higher than the estimates given all observed data. The missing group contain respondents with lower attitudes and they are more inclined to score in the first category. Most of the corresponding 95% HPD regions of the estimated threshold parameters given the item response data of the missing group do not contain the estimated threshold values given all observed item response data. The partitioning of the data based on the values of the splitter item (observed/missing) resulted in different item parameter estimates and, subsequently, the missing data are nonignorable.

Again, BMI values are sampled under the null-model for the missing group. In Figure 4.4 are the marginal posterior distributions given of all BMI parameters using the last item as the splitter item. The estimated BMI values captured the difference between the estimates given all data and the estimates given only the item responses of the missing group. That is, more than 50% of the estimated BMI values are significant given 95% highest posterior density intervals (HPD). Note that the estimated posterior means are negative since this corresponds with a shift upwards in threshold values, see Equation (4.10).

In all three experiments, the mechanism for generating missing item responses was known to produce nonignorable missing data. The proposed method detected nonignorable missing data in two experiments. In all three experiments, it was investigated that the splitter item technique did not detect significant differences in item parameter estimates across groups in case of ignorable missing item response data. In this case, the data were grouped completely at random, and significant differences were also not to be expected.

4.6 Discussion

The splitter item technique can be used to test whether the response mechanism leads to ignorable or nonignorable missing data. In this proposed procedure, it is tested if the item parameter estimates differ across the subsets of item response data. Differences in item parameter estimates across subsets indicate nonignorable missing data since the splitting of the data was done according to the values of the splitter item (observed/missing). Two methods were proposed for detecting differences in estimates across groups. In the first method, all parameters are estimated given the subsets of item response data. Then, summary statistics of the estimated marginal posterior distributions of the item parameters can be used for detecting differences. In the second method, parameters of an IRT model for binary or ordinal responses are estimated given all observed data using MCMC and, as an additional sampling step, BMI values are sampled. These BMI values provide information regarding any fluctuations in item parameter values across subsets of item response data. In the Appendix, it is shown that the Bayesian Modification distribution is a good approximation of the true marginal posterior distribution of the item parameters. As a result, the BMI values can be obtained as a by-product of the MCMC algorithm for estimating the parameters of an IRT model.

Further study is focused on the generalization of the BMI approach to more complex models. One of the main advantages of estimating IRT models using a fully Bayesian approach is that traditional frequentist approaches break down because of the infeasible numerical evaluation of the multiple integrals involved in solving the estimation equations. The splitter item technique in combination with BMI becomes particularly interesting when estimating complex IRT models, like testlet response models (Bradlow, Wainer and Wang, 1999), models with multidimensional latent abilities (Béguin and Glas, 2001), and multilevel IRT models (Fox, 2004; Fox and Glas, 2001, 2003) and it is in the realm of these models that more research needs to be done.

4.7 Appendix

The marginal BM distribution in equation (4.16) of the BMI parameter can be obtained by integration using a noninformative reference prior for $\boldsymbol{\lambda}_k$ and $\boldsymbol{\gamma}_k$.

$$\begin{aligned}\tilde{p}(\boldsymbol{\lambda}_k \mid \mathbf{z}_k, \mathbf{x}) &= \int_0^\infty \int_\Omega \tilde{p}(\boldsymbol{\lambda}_k \mid \mathbf{z}_k, \mathbf{x}, \boldsymbol{\gamma}_k, \sigma_0^2) p(\boldsymbol{\gamma}_k, \sigma_0^2 \mid \mathbf{z}_k, \mathbf{x}_1) d\boldsymbol{\gamma}_k d\sigma_0^2 \\ &\propto \int_0^\infty \int_\Omega \sigma_0^{-(n+r+1)} \exp\left(\frac{-1}{2\sigma_0^2}\left(\left[(\boldsymbol{\lambda}_k - \hat{\boldsymbol{\lambda}}_k) - \mathbf{v}_{12}^t \mathbf{v}_{11}^{-1}\right.\right.\right. \\ &\left.(\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_{k,\mathbf{x}_1})\right]^t \mathbf{w}_{22}\left[(\boldsymbol{\lambda}_k - \hat{\boldsymbol{\lambda}}_k) - \mathbf{v}_{12}^t \mathbf{v}_{11}^{-1}(\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_{k,\mathbf{x}_1})\right] \\ &\left.\left. + (n-2)s_0^2 + (\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_{k,\mathbf{x}_1})^t \mathbf{v}_{11}^{-1}(\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_{k,\mathbf{x}_1})\right)\right) d\boldsymbol{\gamma}_k d\sigma_0^2\end{aligned}$$

where $\Omega = \{\boldsymbol{\gamma}_k \in \mathbb{R}^2\}$ and $\mathbf{w}_{22} = \left(\mathbf{v}_{22} - \mathbf{v}_{12}^t \mathbf{v}_{11}^{-1} \mathbf{v}_{12}\right)^{-1}$. Terms can be gathered together by defining $\boldsymbol{\Lambda}_k = (\boldsymbol{\gamma}_k, \boldsymbol{\lambda}_k)$ and using a specific form of the inverse of a partitioned full rank symmetric matrix, that is,

$$\mathbf{v}^{-1} = \begin{bmatrix} \mathbf{v}_{11}^{-1} + \mathbf{v}_{11}^{-1}\mathbf{v}_{12}\mathbf{w}_{22}\mathbf{v}_{12}^t\mathbf{v}_{11}^{-1} & -\mathbf{v}_{11}^{-1}\mathbf{v}_{12}\mathbf{w}_{22} \\ -\mathbf{w}_{22}\mathbf{v}_{12}^t\mathbf{v}_{11}^{-1} & \mathbf{w}_{22} \end{bmatrix}. \tag{4.17}$$

It follows that

$$\begin{aligned}\tilde{p}(\boldsymbol{\Lambda}_k \mid \mathbf{z}_k) &\propto \int_0^\infty \int_\Omega \sigma_0^{-(n+r+1)} \exp\left(\frac{-1}{2\sigma_0^2}\left((n-2)s_0^2 + (\boldsymbol{\Lambda}_k - \right.\right. \\ &\left.\left.\hat{\boldsymbol{\Lambda}}_k)^t \mathbf{v}^{-1}(\boldsymbol{\Lambda}_k - \hat{\boldsymbol{\Lambda}}_k)\right)\right) d\boldsymbol{\gamma}_k d\sigma_0^2 \\ &\propto \int_\Omega \left[\nu\, s_0^2 + (\boldsymbol{\Lambda}_k - \hat{\boldsymbol{\Lambda}}_k)^t \mathbf{v}^{-1}(\boldsymbol{\Lambda}_k - \hat{\boldsymbol{\Lambda}}_k)\right]^{-(\nu+r+2)/2} d\boldsymbol{\gamma}_k,\end{aligned} \tag{4.18}$$

where $\nu = n - r$. This last step in equation (4.18) follows from the integral formula

$$\int_0^\infty x^{-(p+1)} \exp\left(\frac{-a}{x^2}\right) dx = \frac{1}{2} a^{-p/2} \Gamma(p/2), \tag{4.19}$$

with a, p greater than zero and $\Gamma(.)$ the Gamma function. Finally, the right hand side of equation (4.18) can be recognized as the multivariate t-distribution. As a result, the marginal distribution of $\boldsymbol{\lambda}_k$, a subset of $\mathbf{\Lambda}_k$, has a multivariate t-distribution (Box and Tiao, 1973)

$$\boldsymbol{\lambda}_k \mid \mathbf{z}_k \sim t_2 \left[\hat{\boldsymbol{\lambda}_k}, s_0^2 \mathbf{v}_{22}, \nu \right]. \tag{4.20}$$

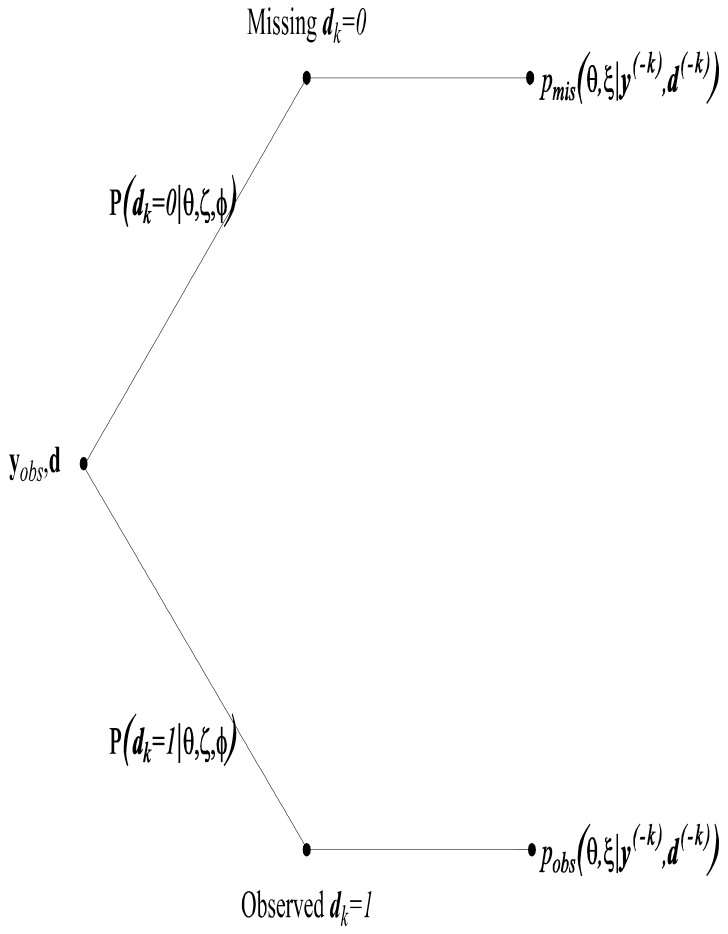
Missing $d_k=0$
$p_{mis}(\theta,\xi|y^{(-k)},d^{(-k)})$
$P(d_k=0|\theta,\zeta,\phi)$
y_{obs},d
$P(d_k=1|\theta,\zeta,\phi)$
$p_{obs}(\theta,\xi|y^{(-k)},d^{(-k)})$
Observed $d_k=1$

FIGURE 4.1. Splitting the observed item response data.

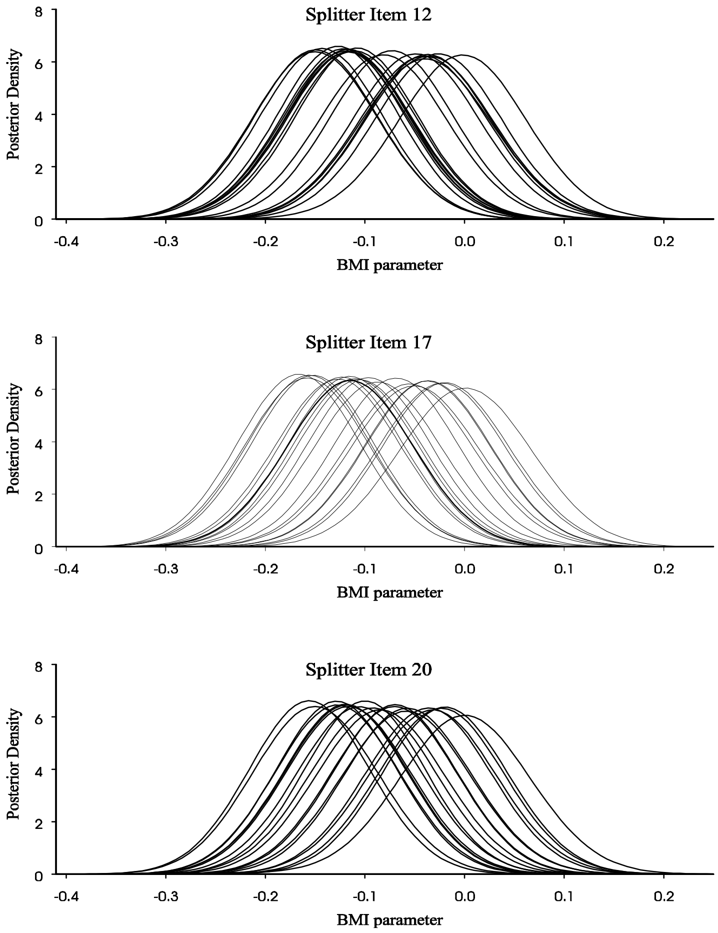

FIGURE 4.2. Posterior distributions of BMI parameters representing differences in item difficulties, corresponding to splitter items 12, 17, and 20.

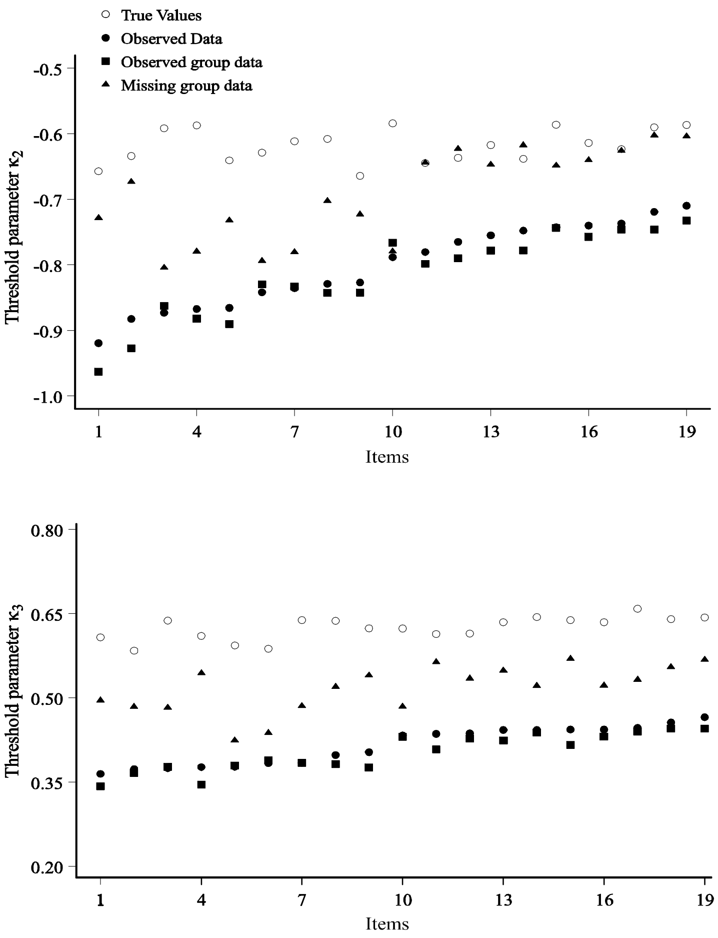

FIGURE 4.3. Threshold parameter estimates given all observed data, and the item responses of the missing and observed group where item 20 served as a splitter item.

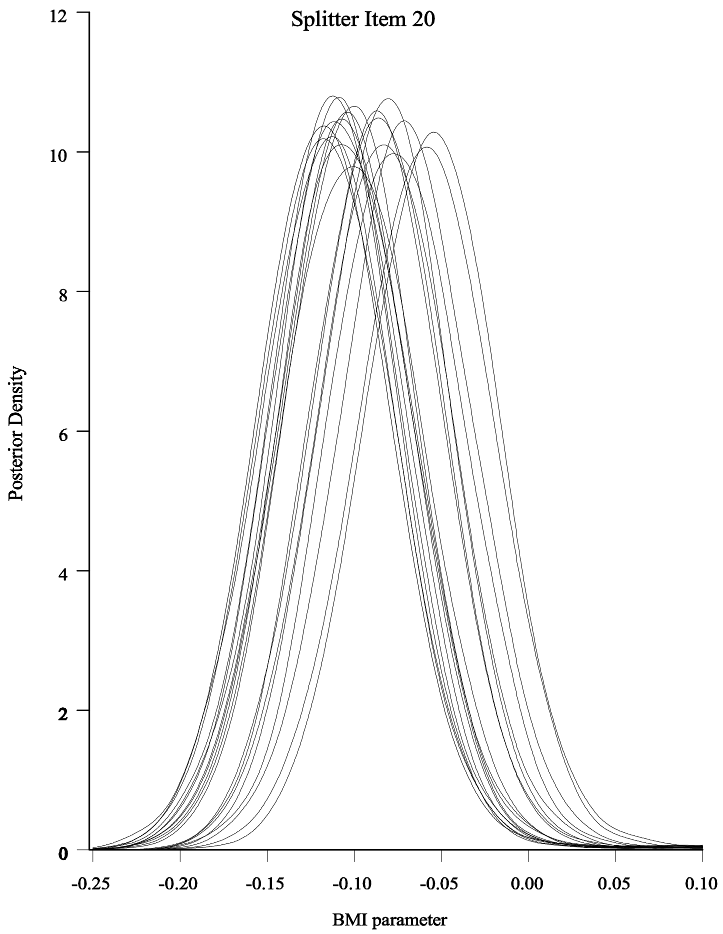

FIGURE 4.4. Posterior distributions of BMI parameters representing differences in item thresholds corresponding to splitter items 20.

TABLE 4.1. Experiment 1: The response mechanism depends upon the examinee ability and item difficulties.

				Splitter Item 20					
		p($\mathbf{b} \mid \mathbf{y}_{obs}$)		$p_{obs}(\mathbf{b} \mid \mathbf{y}_{obs})$		$p_{mis}(\mathbf{b} \mid \mathbf{y}_{obs})$		$\tilde{p}_{mis}(\boldsymbol{\lambda} \mid \mathbf{y}_{obs})$	
Item	**b**	Mean	sd	Mean	sd	Mean	sd	BMI$_{mis}$	HPD
1	.90	.91	.03	1.55	.07	.64	.03	−.12	[−.19, −.05]
2	.89	.91	.03	1.69	.08	.62	.03	−.12	[−.19, −.05]
3	.51	.48	.02	1.16	.06	.19	.03	−.16	[−.23, −.09]
4	1.17	1.13	.03	1.70	.08	.88	.03	−.09	[−.16, −.02]
5	1.50	1.48	.04	2.45	.18	1.18	.04	−.07	[−.14, −.01]
6	−.66	−.63	.02	.00	.04	−.90	.03	−.10	[−.17, −.04]
7	1.10	1.11	.03	1.81	.09	.81	.03	−.13	[−.19, −.05]
8	−1.41	−1.39	.04	−.74	.04	−1.71	.06	−.07	[−.15, −.01]
9	−.63	−.64	.03	.05	.04	−.95	.04	−.13	[−.19, −.05]
10	1.06	1.02	.03	1.75	.09	.72	.03	−.15	[−.22, −.07]
11	−1.10	−1.18	.04	−.52	.06	−1.44	.05	−.06	[−.14, .01]
12	.29	.44	.05	.97	.07	.16	.07	−.04	[−.11, .04]
13	.12	.12	.03	.78	.07	−.15	.03	−.10	[−.17, −.03]
14	.44	.63	.05	1.14	.07	.37	.07	−.03	[−.11, .05]
15	.24	.17	.03	.95	.08	−.12	.03	−.12	[−.19, −.05]
16	.29	.45	.04	.95	.07	.24	.07	−.02	[−.10, .05]
17	.00	.13	.05	.65	.06	−.21	.08	−.02	[−.10, .06]
18	−1.09	−1.09	.03	−.37	.06	−1.38	.05	−.08	[−.16, −.01]
19	−.15	−.18	.03	–	–	−.41	.03	−.06	[−.14, .02]
20	1.00	1.14	.05	–	–	–	–	–	–

TABLE 4.2. Experiment 2:The response mechanism depends upon the examinees unobserved item responses

				Splitter Item 20					
		p($\mathbf{b} \mid \mathbf{y}_{obs}$)		$p_{obs}(\mathbf{b} \mid \mathbf{y}_{obs})$		$p_{mis}(\mathbf{b} \mid \mathbf{y}_{obs})$		$\tilde{p}_{mis}(\boldsymbol{\lambda} \mid \mathbf{y}_{obs})$	
Item	**b**	Mean	sd	Mean	sd	Mean	sd	BMI$_{mis}$	HPD
1	−.25	−.27	.02	−.23	.03	−.32	.03	.02	[−.05, .08]
2	−.35	−.38	.02	−.36	.03	−.41	.03	.00	[−.06, .07]
3	.85	.84	.03	.85	.04	.83	.04	−.03	[−.10, .04]
4	−.20	−.25	.02	−.24	.03	−.25	.03	−.01	[−.07, .06]
5	.90	.88	.03	.92	.03	.82	.04	.00	[−.07, .07]
6	−.55	−.60	.02	−.59	.03	−.62	.03	−.01	[−.08, .06]
7	.35	.30	.02	.31	.03	.28	.03	−.01	[−.08, .06]
8	−.50	−.55	.02	−.52	.03	−.59	.03	−.02	[−.09, .05]
9	−.05	−.08	.02	−.03	.03	−.14	.03	.04	[−.03, .11]
10	.95	.92	.03	.96	.04	.89	.04	.03	[−.04, .10]
11	.30	.06	.03	.05	.04	−.03	.04	.00	[−.07, .07]
12	.75	.48	.03	.56	.04	.38	.05	.02	[−.05, .09]
13	−.45	−.73	.03	−.67	.04	−.83	.05	.03	[−.04, .10]
14	.70	.32	.03	.36	.05	.26	.05	−.00	[−.08, .07]
15	.25	−.13	.04	−.10	.04	−.18	.05	.01	[−.06, .09]
16	−.75	−1.03	.04	−.97	.05	−1.10	.06	−.00	[−.08, .06]
17	−.65	−1.06	.04	−.99	.05	−1.13	.07	−.00	[−.08, .07]
18	−1.00	−1.32	.05	−1.31	.06	−1.33	.07	−.00	[−.08, .07]
19	.10	−0.28	.03	–	–	−.28	.03	.00	[−.06, .07]
20	−.85	−1.10	.04	–	–	–	–	.00	[−.08, .08]

5
Fixed effect IRT Model

ABSTRACT: A fixed effect item response theory (IRT) model is developed for modeling group specific item parameters. Two applications are presented. The first application is that the proposed model can be used to detect whether a response mechanism is ignorable using the splitter item technique. The second application is the detection of differential item functioning. In the latter application, the fixed effect item parameters can model item parameter differences between groups. Simulation studies are presented to show the feasibility and performance of the method on both applications.

KEYWORDS: analysis of variance, differential item functioning, fixed effect, item response theory model, MCMC.

5.1 Introduction

Interest is often focused on the possibility that educational and psychological measures are biased against a particular group of respondents. So-called external bias occurs when test scores have different correlations with non-test variables for two or more groups of examinees. Another form of bias occurs when correlations among item responses differs across two or more groups. This measurement bias leads to noninvariant measurement scales (e.g., the measurement scale is not invariant across groups). This form of item bias is denoted as differential item functioning (DIF). DIF is often modeled using IRT. In the framework of IRT, an item displays DIF when any of the item parameters differs across groups. Statistics for detection of DIF based on IRT models are summarized in Muraki, Mislevy, and Bock (1987), and Thissen, Steinberg, and Wainer (1988, 1993), and references therein. The detection of DIF is complicated due to the fact that group differences in the distribution of the latent variable cause differences in response probabilities that as such are not signs of DIF. In other words, differences in the ability distribution between groups do not constitute DIF. Items are biased or noninvariant when respondents at the same level of the latent variable have different response distributions on the item.

Another common problem in educational and psychological measurement is the occurrence of nonignorable missing data. Rubin (1987) identified a number of situations in which statistical inferences based on the observed data and ignoring the distribution of the missing data indicators become biased. Roughly speaking, this bias does not occur if the distribution of the missing data indicator does not depend on the missing data. If the missing data cannot be ignored, a concurrent probability model must be defined for the observed and missing data, and inferences are made averaging over the missing data. Examples of such models were proposed by O'Muircheartaigh and Moustaki (1999, also see, Moustaki & O'Muircheartaigh, 2000; Moustaki & Knott, 2000; Bernaards & Sijtsma, 1999, 2000; Conaway, 1992; Park & Brown, 1994; Holman & Glas, 2005). Below it will be shown that a splitter item technique (Molenaar 1983; Van den Wollenberg, 1979) can be used for testing ignorability. In the splitter item technique, the sample of respondents are splitting up in two groups depending whether the response on the splitter item was observed or missing. Differences in item parameter estimates obtained in the two groups may then indicate nonignorable missing data.

In general, a unidimensional IRT model is appropriate for data in which a single common factor, say a latent variable, underlies the item responses. The person's response pattern on a particular set of items provides the basis for estimating the level on the latent variable level. IRT models involve an assumption about the distribution of the item response given the latent variable. Besides on the latent ability variable, the item response function also depends on item parameters which are distinct from the ability variable. In a fixed effect IRT model, group specific item parameters are added to the response function to model group specific fixed effects such as DIF, or differences in response behavior between subgroups formed using a splitter item that might indicate a violation of the ignorability assumption.

Though the approach that will be sketched below is quite general, the two-parameter logistic and normal ogive models will be used as an example. Estimation will be developed in a Bayesian framework. The development of powerful sampling-based estimation techniques have stimulated the application of Bayesian methods. Markov chain Monte Carlo (MCMC) methods, such as Gibbs sampling and Metropolis-Hastings (M-H), can be used to simultaneously estimate all model parameters. An MCMC implementation will be introduced for the sampling of all model parameters that combines various advantages of different MCMC schemes for sampling IRT parameters.

In the next section, a general notation is given for fixed effect IRT models. Then, it will be shown how the model can be used to detect nonignorable missing data when using the splitter item technique. Next, it will be shown how the model can be used to explore DIF. Both applications are illustrated using artificial data. The last section contains a discussion and suggestion for further research.

5.2 A Fixed Effects IRT Model

The two-parameter normal ogive (2PNO) and the two-parameter logistic (2PL) models can be used to describe the relationship between a set of binary response items and a latent variable. Let a response of a person i to an item labeled k be coded by a y_{ik}. The probability of a correct response of a person i on an item k is defined as

$$P\big(y_{ik} = 1 \mid \theta_i, a_k, b_k\big) = \begin{cases} \big[1 + \exp(-D(a_k\theta_i - b_k))\big]^{-1}, \\ \text{for the 2PL} \\ \Phi\big(a_k\theta_i - b_k\big), \\ \text{for the 2PNO,} \end{cases} \tag{5.1}$$

where a_k is the item discrimination parameter, and b_k is the item difficulty parameter in both models. The item parameters will also be denoted by ξ_k, with $\xi_k = (a_k, b_k)^t$. Function Φ is the cumulative standard normal distribution, and the factor D, usually taken to be 1.7, is a scaling factor introduced to scale the parameters of the logistic function as close as possible to the parameters of the normal ogive function.

Let λ_{kj} express the difference between a group j specific difficulty parameter, indexed k, and a fixed difficulty parameter b_k across groups indexed $j = 1, \ldots, J$. So, group specific difficulty parameters

b_{kj} can be expressed as

$$b_{kj} = b_k + \lambda_{kj}, \tag{5.2}$$

where the difference λ_{kj} is called a jth factor level effect or the jth treatment effect in ANOVA terms with the usual constraint that $\sum_j \lambda_{jk} = 0$ for $k = 1, \ldots, K$.

In a regression approach equivalent to an one-way ANOVA a design matrix $\mathbf{x}$ defines the grouping structure. Indicator variables are needed that take on values 0,1, or -1. It follows that:

$$a_k\boldsymbol{\theta} - (b_k + \mathbf{x}\boldsymbol{\lambda}_k) = \begin{pmatrix} \theta_{11} & -1 \\ \theta_{21} & -1 \\ \vdots & \vdots \\ \theta_{n_1 1} & -1 \\ \theta_{12} & -1 \\ \vdots & \vdots \\ \theta_{n_2,2} & -1 \\ \vdots & \vdots \\ \theta_{1J} & -1 \\ \vdots & \vdots \\ \theta_{1n_J} & -1 \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix} -$$

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & -1 \end{pmatrix} \begin{pmatrix} \lambda_{k1} \\ \lambda_{k2} \\ \lambda_{k3} \\ \vdots \\ \lambda_{kJ-1} \end{pmatrix},$$

such that $\lambda_{kJ} = -\lambda_{k1} - \lambda_{k2} - \cdots - \lambda_{kJ-1}$. The indicator variable $\mathbf{x}$ denotes the specific group-membership. As a result, in the fixed effects IRT model the probability of a correct response of a person i

on an item k, is defined as

$$P\big(y_{ik}=1 \mid \theta_i, a_k, b_k, \boldsymbol{\lambda}_k\big) = \begin{cases} \big[1+\exp(-D(a_k\theta_i-(b_k+\mathbf{x}_i^t\boldsymbol{\lambda}_k)))\big]^{-1}, \\ \text{for the 2PL} \\ \Phi\big(a_k\theta_i-(b_k+\mathbf{x}_i^t\boldsymbol{\lambda}_k)\big), \\ \text{for the 2PNO.} \end{cases} \tag{5.3}$$

In the present paper, attention is focused on differences in difficulty parameters. However, the fixed effects IRT model is easily extended to model differences in discrimination parameters across groups.

In the fixed effects IRT model, interest is focused in the individual group means of item parameters λ_{kj}, Equation (5.2), and they are of interest in themselves. The interest is not focused on the variance in item parameters across groups. In that case, the λ_{kj} are to be considered as random effects, and they are specified as independently distributed observations with a distribution. Subsequently, main interest is focused on this distribution. In the fixed effects approach it is a priori assumed that the λ_{kj} bear no strong relationship to one another. In the cases where might be more realistic to assume that the λ_{kj} are thought of as coming from a distribution, numerical problems occur when estimating variance components given a small number of groups. In this situation a fixed effects analysis can be very useful and avoids the complex statistical modeling of a mixture distribution and specification of hyperprior distributions. Fixed effects analyses from the Bayesian viewpoint have been tackled by, among others, Jeffreys (1961) and Lindley (1965). A random effects approach in IRT modeling has been considered by Janssen, Tuerlinckx, Meulders, and De Boeck (2000). In that approach, item parameters are considered as independent observations from a group specific population distribution, that is, the items in the test are seen as a random sample from this distribution. Subsequently, interest is focused on this item population distribution.

5.3 Testing for Non-Ignorable Missing Data

Holman and Glas (2005) propose an IRT model for taking non-ignorable missing data into account. In this model, the observed responses and the missing data indicators are modeled using distinct

IRT models, and the two latent variables associated with these two IRT models have a two-variate normal distribution. If the covariance between the two latent variables is non-zero, ignorability is violated. In that case, if the parameters of the IRT model for the observed responses are estimated ignoring the missingness, they prove to be biased (Holman & Glas, 2005). To assess this violation of ignorability, the data can be divided into two samples using a splitter item, say item k. The first group consists of respondents who have an observed response on this item, the second group consists of respondents who have a missing value on this item. Accordingly, the first sample will be denoted as the observed group; the observed item responses of individual i except those to item k, $\mathbf{y}_{i,obs}^{(-k)}$ with $d_{ik} = 1$, $i = 1, \ldots, n$. The second sample will be denoted as the missing group: the observed item responses of individual i except those to item k, $\mathbf{y}_{i,mis}^{(-k)}$ with $d_{ik} = 0$, $i = 1, \ldots, n$. In fact, the observed data is grouped in two sets. This can also be accomplished by specifying the indicator variable $\mathbf{x}$ in such a way that it represents the grouping structure defined by the splitter item. In that case, the fixed effects parameter $\boldsymbol{\lambda}$ represents item parameter differences between the observed and missing data set. Interest is focused on the marginal posterior distribution of $\boldsymbol{\lambda}$, $p\big(\boldsymbol{\lambda} \mid \mathbf{y}\big)$. When the missing data are nonignorable, the item parameters differ across groups, and the estimated $\boldsymbol{\lambda}$ values are different from zero. So, the splitter item technique is used to detect a nonignorable missing data mechanism by testing whether the fixed effects parameters are significantly different from zero.

5.4 Modeling Differential Item Functioning

The value of the ICC at a specific value for the latent variable corresponds to the conditional probability of a correct response given the level of the latent variable. When an ICC differ across groups then it is said that this item function differently and exhibit DIF. So, respondents across groups with the same level of the latent variable have different probabilities of scoring this item correct.

Several techniques for detection of DIF items based on IRT models have been proposed (see, .e.g., Glas, 2001; Glas & Verhelst, 1995; Hambleton & Rogers, 1989; Kelderman, 1989). In most cases, attention is focused on differences in response probabilities between groups conditional on the level of the latent variable. Thissen, Steinberg, and Wainer (1993), and Glas (1998, 2001) considered DIF as a special case of IRT model misfit. They both used statistical tests in an IRT framework to explore DIF. In a frequentist framework, Glas (1998, 2001) modeled DIF in a common IRT model using multiple background or categorical dummy variables, where these variables model DIF. In this approach, the parameters of the IRT model are estimated and Lagrange Multiplier (LM) tests for DIF, based on the model extension using background variables, are performed for each item. In the present Bayesian approach, all parameters of the fixed effects IRT model are simultaneously estimated. Subsequently, the Bayes factor can be used to identify DIF items.

As an example, consider items that may function differently across groups, say, gender and nations. To model differences in ICC's across gender ($s = 1, 2$) and nations ($r = 1, \ldots, R$) define a fixed effects (probit) IRT model as:

$$P\bigl(y_{iksr} = 1 \mid \theta_i, a_k, b_k, \lambda_{k1s}, \lambda_{k2r}\bigr) = \Phi\bigl(a_k\theta_i - (b_k + \lambda_{k1s} + \lambda_{k2r})\bigr), \tag{5.4}$$

where λ_{1s} is the main effect of being female (s = 2), and λ_{2r} is the main effect of being grouped in nation r, with $\lambda_{11} = 0$ and $\lambda_{21} = 0$ taken as a baseline related to a so-called focal group. Subsequently, let indicator variable $\mathbf{x}$ represent this grouping structure, and let the fixed effects IRT model with two grouping variables be given by (5.3). Note that interaction effects between gender and nations are easily incorporated.

5.5 Estimating Model Parameters

Direct posterior inference is not possible since the joint posterior distribution is very complex. However, samples from this distribution can be obtained using MCMC methods. Then, inferences concerning the model parameters can be made using the sampled values. Below, M-H and Gibbs sampling algorithms are used for sampling parameter values for the item parameters, fixed effects parameters, and the ability parameters from their posterior distributions. Using the method of data augmentation, realizations from a complicated posterior density can be obtained by augmenting the variables of interest by one or more additional variables such that sampling from the full conditional distributions is easy. Albert (1992) constructed an MCMC chain using the auxiliary variable method for estimating the two-parameter normal ogive model. Generating realizations from the full conditionals is complicated but with the introduction of this augmented variable the full conditionals are tractable and easy to simulate from. Maris and Maris (2002) developed an auxiliary variable method for logistic IRT models that handles different prior distributions in a flexible way. The augmented data are defined in such a way that each full conditional becomes an indicator function with bounds specified by the other parameter values. As a result, the sampling of the parameters is easy. However, the sampled values are highly correlated due to this incorporated dependency structure. As a result, the samples cannot be drawn freely from the target distribution but are restricted to a subspace specified by the other parameter values.

In the present paper, a combination of both methods for simultaneously estimating the parameters of a fixed effects two-parameter IRT model is outlined. In this approach, it is easy to handle different kinds of prior information, the convergence is fast, and the samples are not highly correlated. Fox and Hendrawan (2005) proposed this method for the MCMC estimation of two-parameter IRT models.

Let $\mathcal{L}(0,1)$ and $\mathcal{N}(0,1)$ denote the standard logistic and standard normal distribution function, respectively. Further, define augmented data $\mathbf{z}$,

$$z_{ik} \mid y_{ik}, \theta_i, a_k, b_k, \boldsymbol{\lambda}_k \sim \begin{cases} \mathcal{L}(0,1), \\ \text{for the 2PL} \\ \mathcal{N}(0,1), \\ \text{for the 2PNO,} \end{cases} \tag{5.5}$$

where y_{ik} is the indicator that assumes a value one if $z_{ik} > D((b_k + \mathbf{x}_i^t \boldsymbol{\lambda}_k) - a_k\theta_i)$ and zero otherwise ($D = 1.7$ for the 2PL and $D = 1$ for the 2PNO model). Note that the augmentation step defines a probit or logit analysis. The full conditional distribution of the model parameters are each tractable and easy to simulate from given the augmented data.

- Full conditional distribution of $\boldsymbol{\theta}$. The prior for $\boldsymbol{\theta}$ is a normal distribution with mean parameter μ and variance parameter σ. It follows that

$$
\begin{aligned}
p(\theta_i \mid \mathbf{y}, \mathbf{z}, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}, \mu, \sigma) \propto & \prod_k I(z_{ik} \geq D((b_k + \mathbf{x}_i^t \boldsymbol{\lambda}_k) - a_k\theta_i))^{y_{ik}} \\
& I(z_{ik} < D((b_k + \mathbf{x}_i^t \boldsymbol{\lambda}_k) - a_k\theta_i))^{1-y_{ik}} p(\theta_i \mid \mu, \sigma) \\
& = I\Bigg(\max_{k \mid y_{ik}=1} \frac{(b_k + \mathbf{x}_i^t \boldsymbol{\lambda}_k) - z_{ik}/D}{a_k} < \theta_i < \\
& \min_{k \mid y_{ik}=0} \frac{(b_k + \mathbf{x}_i^t \boldsymbol{\lambda}_k) - z_{ik}/D}{a_k} \Bigg) p(\theta_i \mid \mu, \sigma),
\end{aligned}
$$

where $I(\cdot)$ is an indicator function assuming a value one if the condition in the argument is fulfilled and is equal to zero otherwise.

- Full conditional distribution of **a**,**b**,$\boldsymbol{\lambda}$. The fixed effects parameters, $\boldsymbol{\lambda}$, are taken to be a priori exchangeable. That is, λ_{kj}, $j = 1, \ldots, J$ are assumed independent and normally distributed with mean zero and variance σ_λ, with a large value for σ_λ to specify a diffuse proper prior and to specify independence among the fixed effects parameters. Independent proper noninformative priors for the discrimination and difficulty parameters are specified, that is,

$$
p(a_k, b_k) = p(a_k)p(b_k) \propto I(a_k \in \mathcal{A}) I(b_k \in \mathcal{B}),
$$

where $\mathcal{A}$ and $\mathcal{B}$ are a sufficiently large bounded intervals in $\mathbb{R}^+$ and $\mathbb{R}$, respectively. As a result,

$$
p(\mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}) = p(a_k)p(b_k) \prod_j p(\lambda_{kj}) \propto \prod_{k,j} p(\lambda_{kj}) I(a_k \in \mathcal{A}) I(b_k \in \mathcal{B}).
$$

Define augmented data $\mathbf{z}_k^*$,

$$
\begin{aligned}
\mathbf{z}_k^* &= D(a_k \boldsymbol{\theta} - (b_k + \mathbf{x}\boldsymbol{\lambda}_k)) + \boldsymbol{\epsilon}_k \\
\mathbf{z}_k^* &= \mathbf{H}\boldsymbol{\Xi}_k + \boldsymbol{\epsilon}_k
\end{aligned}
\tag{5.6}
$$

where $\mathbf{H} = D(\boldsymbol{\theta}, -\mathbf{1}, -\mathbf{x})$, $\boldsymbol{\Xi}_k = (a_k, b_k, \boldsymbol{\lambda}_k)^t$, and $\boldsymbol{\epsilon}_k$ equals the augmented data $\mathbf{z}_k$ and they are standard normal or standard logistic distributed. The full conditional distribution can be specified as follows

$$
\boldsymbol{\Xi}_k \mid \mathbf{z}_k^*, \boldsymbol{\theta} \sim \mathcal{N}\Big(\hat{\boldsymbol{\Xi}}_k, c\big(\mathbf{H}^t\mathbf{H}\big)^{-1}\Big) p(\boldsymbol{\Xi}_k), \tag{5.7}
$$

where

$$\hat{\boldsymbol{\Xi}}_k = \left(\mathbf{H}^t\mathbf{H}\right)^{-1}\mathbf{H}^t\mathbf{z}_k^*,$$

and $c = 1$ or $c = \pi^2/3$ in case of 2PNO or 2PL augmented data, respectively. Note that the standard logistic cumulative distribution resembles the normal cumulative distribution with mean zero and variance $\pi^2/3$. A M-H probability can be used to correct any deficiencies in the approximation, since the tail of the logistic distribution is somewhat longer. However, almost every value is accepted since both distributions are quite comparable. In fact, a very good proposal distribution is specified in equation (5.7) for the fixed effects 2PL model.

5.6 Bayesian Inference

Summary statistics, such as the posterior mean or median, are used to report the results. A Bayesian confidence interval can provide information about the 'most likely' parameter values. In general, a $100(1-\alpha)\%$ credible set, $C_{\boldsymbol{\lambda}}(\mathbf{y})$, for $\boldsymbol{\lambda}$ is any set of values with

$$1-\alpha \leq P\left(C_{\boldsymbol{\lambda}}(\mathbf{y}) \mid \mathbf{y}\right) = \int_{C_{\boldsymbol{\lambda}}(\mathbf{y})} p\left(\boldsymbol{\lambda} \mid \mathbf{y}\right) d\boldsymbol{\lambda}. \tag{5.8}$$

It will be assumed that the (marginal) posterior density function is unimodal. The null-hypothesis $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ is of particular interest, however, it is not realistic to have a precise null-hypothesis. This is better represented as

$$H_0 : |\boldsymbol{\lambda} - \boldsymbol{\lambda}_0| \leq \epsilon \text{ versus } H_1 : |\boldsymbol{\lambda} - \boldsymbol{\lambda}_0| > \epsilon, \tag{5.9}$$

where ϵ is "small". The point null hypothesis will be seen as an approximation for the small interval null as in Equation (5.9). In general, a Bayesian confidence region can be determined and conclusions are directly drawn from this region. That is, $C_{\boldsymbol{\lambda}}(\mathbf{y})$ provides information about the location of $\boldsymbol{\lambda}$, its distance to $\boldsymbol{\lambda}_0$, and if this distance makes a practical difference. Berger and Delampady (1987) argued that Bayesian credible intervals are often inappropriate when testing $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ with a specific value $\boldsymbol{\lambda}_0$. They stated that the likelihood of a special point $\boldsymbol{\lambda}_0$, say, outside a confidence region $C_{\boldsymbol{\lambda}}(\mathbf{y})$ is often not too much smaller than the average likelihood in $C_{\boldsymbol{\lambda}}(\mathbf{y})$.

As a result, there is no strong evidence for rejecting $\boldsymbol{\lambda}_0$. Besides reporting a credible region, the Bayes factor can be used to test the null-hypothesis. Note that the computation of the Bayes factor against H_0 is easily constructed from the MCMC output for estimating the fixed effects IRT model parameters. Let M_0 denote the model with $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \mathbf{0}$, subsequently, $\boldsymbol{\lambda}$ is unconstrained in the fixed effects IRT model, denoted as M. Let $\boldsymbol{\Xi} = (\mathbf{a}, \mathbf{b}, \boldsymbol{\theta})$ and assume that $p(\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi} \mid \mathbf{y}, M) = p(\boldsymbol{\Xi} \mid \mathbf{y}, M_0)$. Then the marginal likelihood under model M_0 can be related to the marginal likelihood under the fixed effects IRT model M (see, e.g., Chen, Shao, & Ibrahim, 2000; Verdinelli & Wasserman, 1995):

$$\begin{aligned} p(\mathbf{y} \mid M_0) &= \int p(\mathbf{y} \mid \boldsymbol{\Xi}, M_0) p(\boldsymbol{\Xi} \mid M_0) d\boldsymbol{\Xi} \\ &= \int \frac{p(\boldsymbol{\Xi} \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi} \mid M)} p(\boldsymbol{\lambda} = 0, \boldsymbol{\Xi} \mid M) p(\mathbf{y} \mid \boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi}, M) d\boldsymbol{\Xi} \\ &= p(\mathbf{y} \mid M) \int \frac{p(\boldsymbol{\Xi} \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi} \mid M)} p(\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi} \mid \mathbf{y}, M) d\boldsymbol{\Xi}. \end{aligned} \tag{5.10}$$

As a result, the Bayes factor for testing the null-hypothesis $\boldsymbol{\lambda} = \mathbf{0}$ can be stated as:

$$\begin{aligned} BF &= \int \frac{p(\boldsymbol{\Xi} \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi} \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \boldsymbol{\Xi}, \mathbf{y}, M) p(\boldsymbol{\Xi} \mid \mathbf{y}, M) d\boldsymbol{\Xi} \\ BF &= \mathcal{E}\left[\frac{p(\boldsymbol{\Xi} \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi} \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \boldsymbol{\Xi}, \mathbf{y}, M)\right], \end{aligned} \tag{5.11}$$

where the expectation is taken with respect to the marginal posterior distribution $p(\boldsymbol{\Xi} \mid \mathbf{y}, M)$. A single MCMC output denoted as $\boldsymbol{\Xi}^{(m)}$, $(m = 1, \ldots, M)$ from the posterior distribution $p(\boldsymbol{\Xi} \mid \mathbf{y}, M)$ can be used to compute the Bayes factor. That is,

$$\widehat{BF} = M^{-1} \sum_m \frac{p\left(\boldsymbol{\Xi}^{(m)} \mid M_0\right)}{p\left(\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\Xi}^{(m)} \mid M\right)} p\left(\boldsymbol{\lambda} = \mathbf{0} \mid \boldsymbol{\Xi}^{(m)}, \mathbf{y}, M\right). \tag{5.12}$$

A special case occurs when $p(\Xi \mid \boldsymbol{\lambda} = \mathbf{0}, M) = p(\Xi \mid M_0)$. Via Equation (5.11) it follows that

$$
\begin{aligned}
BF &= \int \frac{p(\Xi \mid \boldsymbol{\lambda} = \mathbf{0}, M)}{p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \Xi, \mathbf{y}, M) p(\Xi \mid \mathbf{y}, M) d\Xi \\
&= \int \frac{1}{p(\boldsymbol{\lambda} = \mathbf{0} \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \Xi, \mathbf{y}, M) p(\Xi \mid \mathbf{y}, M) d\Xi \qquad (5.13) \\
&= \frac{p(\boldsymbol{\lambda} = \mathbf{0} \mid \mathbf{y}, M)}{p(\boldsymbol{\lambda} = \mathbf{0} \mid M)},
\end{aligned}
$$

which is known as the Savage-Dickey density ratio (Dickey, 1971). Note that the Bayes factor in Equation (5.13) is reduced to estimating the marginal posterior density $p(\boldsymbol{\lambda} \mid \mathbf{y}, M)$ at the point $\boldsymbol{\lambda} = \mathbf{0}$.

In a different way, Klugkist (2004) derived an expression for the Bayes factor, under comparable assumptions, that enables its computation via MCMC output under model M. In this approach the Bayes factor is expressed as a ratio of two proportions, a ratio of priors, and a ratio of posterior distributions, where the prior and posterior distributions are defined for the constrained and the unconstrained model. The ratios are estimated using the MCMC output.

5.7 Simulation Study

A simulation study was used to assess the performance of the MCMC algorithm and to illustrate the usefulness of the fixed effects IRT model. In simulation study 1, data were generated using a nonignorable missing data mechanism. In simulation study 2, data were generated given DIF items.

5.7.1 Simulation Study 1

Analogous to Bradlow and Thomas (1998) and the example in the previous chapter, response data were simulated as if students were allowed to choose a subset of items. In this setup, for a subset of items, responses were simulated for pairs of items. This was done in such a way that for each person one response was generated for each paired item. The response mechanism was such that an item response was generated for the easier items within pairs with probability $p_1 =$ .95 and the harder item with probability $p_2 = .05$ if the persons' ability level was positive. If the persons' ability level was negative, an item response was generated for one of the items at random. So, the distribution of the response mechanism depends on the ability parameters $\boldsymbol{\theta}$ underlying the observed responses and the difficulty parameters.

Two groups were identified as follows: one group of respondents, denoted as the observed group, responding to splitter item 20, and the other group of respondents not responding to the splitter item, denoted as the missing group. So, the last item, $k = 20$, was considered as a splitter item and the corresponding responses (observed/-missing) were considered as a group indicator. It was expected that the item difficulties varied over groups since the distribution of abilities varied across groups. In the fixed effects IRT model, the observed group was considered as the baseline group. The fixed effects in Equation (5.3) represent item parameter differences between this baseline group and the missing group.

In this simulation study, 5000 abilities and 20 difficulty parameter values were generated from a standard normal distribution. Discrimination parameters were generated from a log-normal distribution. These parameter values were used to generate item response data according the 2PNO model. The last ten items were considered as paired items. The Gibbs sampling algorithm was used for estimating the parameters of the fixed effects IRT model. A total of 10,000 iterations were used for estimating the model parameters with a burn-in period of 1,000 iterations. The fixed effects IRT model was identified by fixing the scale of the latent variable with mean zero and variance one.

In Table 5.1 are the posterior means and standard deviations given of the difficulty parameters for different subsets of item response data. The posterior means of $p\big(\mathbf{b} \mid \mathbf{y}_{obs}\big)$ correspond to the estimated item difficulties given all observed item response data. It follows that for most paired item the difficult item is overestimated and the easy item is underestimated. The observed group consisted of the better respondents, making item 20 since it was the easier item. The true item difficulties are highly underestimated in the baseline group, that is, the respondents in the observed group make the items appear more easy. The fixed effects are all positive and significantly different from zero given the 95% HPD regions. As a result, the difficulty parameter estimates for the missing group are a factor $\hat{\boldsymbol{\lambda}}$ higher in comparison to the difficulty parameter estimates in the observed group. The Bayes factor for testing the null-hypothesis $\boldsymbol{\lambda} = \mathbf{0}$ equals approximately zero. So, it can be concluded that the grouping of responses according to values of the splitter item affects the statistical inference. The difficulty parameter estimates vary across groups. The grouping of the data according to splitter item 20 resulted in significant fixed group effects indicating that the way of grouping the data (observed/missing) affects the results.

TABLE 5.1. Parameter estimates of the fixed effects IRT model using splitter item 20.

				Splitter Item 20			
		p($\mathbf{b} \mid \mathbf{y}_{obs}$)		$p_{obs}(\mathbf{b} \mid \mathbf{y}_{obs})$		$p(\boldsymbol{\lambda} \mid \mathbf{y}_{obs})$	
Item	**b**	Mean	sd	Mean	sd	Mean	HPD
1	−1.37	−1.37	.03	−1.71	.06	1.02	[.89, 1.16]
2	.05	.07	.02	−.20	.02	.98	[.87, 1.10]
3	−1.19	−1.14	.02	−1.44	.04	.93	[.80, 1.06]
4	−.42	−.42	.02	−.68	.02	.90	[.80, 1.00]
5	1.24	1.22	.02	.93	.03	1.14	[.92, 1.41]
6	−.77	−.78	.02	−1.07	.03	.93	[.82, 1.05]
7	−.65	−.72	.02	−1.00	.03	.92	[.81, 1.04]
8	−.24	−.25	.02	−.54	.02	.96	[.86, 1.07]
9	.74	.70	.02	.43	.02	.99	[.85, 1.16]
10	−.33	−.32	.02	−.57	.03	.83	[.74, .94]
11	.08	.21	.04	.00	.07	.72	[.55, .92]
12	−.21	−.24	.02	−.50	.02	.96	[.83, 1.10]
13	−.04	−.10	.02	−.38	.02	.98	[.85, 1.13]
14	.93	1.16	.05	.97	.08	.58	[.35, .81]
15	−.54	−.58	.02	−.86	.03	1.01	[.88, 1.14]
16	−.24	−.10	.05	−.39	.08	.78	[.61, .98]
17	.53	.75	.05	.52	.07	.70	[.50, .91]
18	−1.05	−1.09	.03	−1.33	.04	.86	[.71, 1.02]
19	−.40	−.32	.05	–	–	–	–
20	−1.00	−1.05	.03	–	–	–	–

5.7.2 Simulation Study 2

In this numerical example, data were analyzed to investigate the performance of the fixed effects IRT model for detecting DIF items. In four different setups, response patterns, $\mathbf{y}$, were generated according to a fixed effects 2PL model for 2000 persons and 10 items. DIF was imposed on the item difficulties. The respondents were grouped by gender (Male, Female) denoted by x_1 and nations (Dutch, non-Dutch) denoted by x_2 where a female Dutch was coded as $x_1 = 1$ and $x_2 = 1$ respectively. It was assumed that the groups of respondents are homogenous with respect to the latent variable. Three data sets were generated: (1) no DIF items denoted as model M_1, (2) main effect of gender where $\boldsymbol{\lambda}_1 = .25$ for the last five items, denoted as model M_2, and (3) main effects of gender and nations where $\boldsymbol{\lambda}_1 = .20$, $\boldsymbol{\lambda}_2 = .20$ for the last five items, denoted as model M_3.

The MCMC algorithm was used to simultaneously estimate all model parameters given the generated item response data using the 2PL. The convergence of the MCMC chains was checked and it was concluded the all MCMC chains converged within 1000 iterations. Then, $10,000$ iterations were made to estimate the posterior means and standard deviations. Each model was identified by fixing the scale of the latent variable to make the outcomes comparable.

Table 5.2 presents the fixed effects IRT parameter estimates given data generated under model M_1 and M_2. The simulated difficulty parameters are given under the label $\mathbf{b}$. The difficulty parameter estimates and their standard deviations of the null-model with $\boldsymbol{\lambda} = \mathbf{0}$ are given under the label $p(\mathbf{b} \mid \mathbf{y}, \boldsymbol{\lambda} = \mathbf{0})$. It can be seen that for data generated under model M_1, the difficulty parameter estimates of the null model resemble the true parameter values since there are no DIF items simulated. The simulated data were used to estimate the parameters of a fixed effects IRT model where the fixed effects represent a main effect of gender. This model assumes that the item parameters differ across groups of males and females. The difficulty parameters estimates corresponding to the female group using this fixed effects IRT model also resemble the true values. Note that the estimated standard deviations are slightly higher in comparison to the corresponding estimates of the null model. This follows from the fact that the the estimates of the fixed effects IRT model are group specific, and so they are based on less observations. The mean of the fixed parameter estimates, $\boldsymbol{\lambda}$ is given under the label $p(\boldsymbol{\lambda} \mid \mathbf{y})$. The estimated fixed effects are close to zero, and the 95% HPD regions show that none of the effects differ significantly from zero. This corresponds with the fact that the data were generated under model M_1 with no DIF items. The Bayes factor for testing the hypothesis $\boldsymbol{\lambda} = \mathbf{0}$ equals exp(8) and provides strong evidence that the null hypothesis should not be rejected.

The simulated difficulty parameters according to model M_2 are given under the label **b** and correspond to the baseline group (Female, $x_1 = 1$). For the last five items, a gender effect was imposed ($\lambda_k = .25, k = 6, \ldots, 10$), and it can be seen that the estimates of the difficulty parameters under the null model differ from the true values for these last five items. The parameter estimates of the baseline group according to the fixed effects IRT model resemble the simulated difficulty parameters since the model captures item parameter differences between groups. The true main effects are slightly overestimated by the estimated fixed effects parameters but they are all significant for last last items. The positive sign of the estimated fixed effects indicates that the item difficulties in the male group are more difficult. The estimated item difficulties in the male group are the sum of the estimated fixed effects and the estimated difficulties in the female group. Here, the Bayes factor equals exp(−34) and provides strong evidence that the null hypothesis should be rejected.

TABLE 5.2. Parameter estimates of the fixed effects IRT model given data generated under model M_1 and M_2.

			$\mathrm{p}(\mathbf{b} \mid \mathbf{y}, \boldsymbol{\lambda} = 0)$		$p(\mathbf{b} \mid \mathbf{y}, x_1 = 1)$		$p(\boldsymbol{\lambda} \mid \mathbf{y})$	
	Item	**b**	Mean	sd	Mean	sd	Mean	HPD
M_1	1	-1.09	-1.10	.05	-1.07	.07	.07	$[-.14, .29]$
	2	1.21	1.16	.05	1.20	.08	.07	$[-.14, .29]$
	3	1.46	1.57	.06	1.59	.10	.09	$[-.21, .37]$
	4	$-.41$	$-.42$	.03	$-.37$	.05	.09	$[-.06, .25]$
	5	.60	.66	.04	.74	.06	.15	$[-.05, .31]$
	6	$-.05$	$-.05$	.03	$-.06$	.04	$-.02$	$[-.17, .11]$
	7	$-.19$	$-.24$	.03	$-.27$	.05	$-.06$	$[-.22, .09]$
	8	$-.08$	$-.15$	.03	$-.11$	.04	.07	$[-.08, .22]$
	9	$-.34$	$-.39$	.03	$-.35$	.05	.09	$[-.06, .25]$
	10	.20	.19	.03	.14	.04	$-.11$	$[-.28, .03]$
M_2	1	$-.28$	$-.23$	.03	$-.29$	.05	$-.11$	$[-.27, .05]$
	2	-2.10	-2.28	.08	-2.55	.27	$-.61$	$[-1.72, .12]$
	3	$-.38$	$-.41$	.03	$-.38$	.05	.08	$[-.09, .26]$
	4	.59	.63	.04	.61	.05	$-.05$	$[-.23, .12]$
	5	.49	.58	.03	.57	.05	$-.02$	$[-.21, .16]$
	6	1.28	1.59	.06	1.37	.09	$-.39$	$[-.71, -.09]$
	7	-1.32	-1.16	.05	-1.32	.09	$-.31$	$[-.56, -.08]$
	8	$-.33$	$-.47$	.03	$-.33$	.05	$-.30$	$[-.48, -.14]$
	9	-1.01	$-.92$	.04	-1.06	.07	$-.30$	$[-.53, -.10]$
	10	$-.65$	$-.49$	.03	$-.63$	.05	$-.25$	$[-.41, -.09]$

TABLE 5.3. Parameter estimates of the fixed effects IRT model given data generated under model M_3.

			$\mathrm{p}(\mathbf{b} \mid \mathbf{y}, \boldsymbol{\lambda} = 0)$		$p(\mathbf{b} \mid \mathbf{y}, x_1 = 0, x_2 = 0)$	
	Item	$\mathbf{b}$	Mean	sd	Mean	sd
M_3	1	.04	.02	.03	$-.04$	.08
	2	.07	.09	.03	.09	.07
	3	.06	.06	.03	.11	.08
	4	.06	.07	.03	.11	.07
	5	.07	.08	.03	.12	.07
	6	$-.17$	.04	.03	$-.19$	.08
	7	.23	.39	.03	.21	.07
	8	$-.10$	.15	.03	$-.10$	.07
	9	$-.06$	.11	.03	$-.02$	.08
	10	.00	.22	.03	$-.02$	.07

In Table 5.3 presents the parameter estimates given data generated under model M_3. The true simulated difficulty parameters for the baseline group (non-Dutch Males) are given under the label **b**. The difficulty parameter estimates of the null-model, with fixed effects equal to zero, differ from the true values with respect to the last five items. The difficulty parameters of these DIF items are correctly estimated by the fixed effects IRT model. That is, the estimated difficulty parameters of the baseline group resemble the true parameters.

The fixed effects parameters are estimated for the four different groups. In Figure 5.1 are the estimated posterior distributions given of the group specific fixed effects parameters. The dotted lines correspond to the last five items of the test. In the group of Dutch-Females, the last five items are DIF items due to the main effect of gender with $\boldsymbol{\lambda}_1 = .2$. It can be seen that the fixed effects parameters of the DIF items are distributed around .2 but only two are significantly different from zero. The posterior distributions of the fixed effects parameters of the non-DIF items are centered around zero. The estimated posterior variances may seem large but they are based on the size of the groups and not the entire sample size. A main effect of nations, $\boldsymbol{\lambda}_2 = .2$, can be detected in the group of Dutch-Males. That is, three of the five posterior distributions of the fixed effects parameters corresponding to DIF items have a mean significantly different from zero. The true difficulty parameters in the group of Dutch-Females are much higher due to main effects of gender and nations. It can be seen that the corresponding estimates of the fixed effects are approximately .4 for the DIF items, and around zero for the non-DIF items. The Bayes factor equals $\exp(-56)$ and supports the fixed effects IRT model without restricting the fixed effects to be zero. In conclusion, the fixed effects IRT model captures differences in difficulty parameters across groups and detects DIF items. As a result, the measurements of the latent variable are more reliable since differences in item parameters across groups are taken into account.

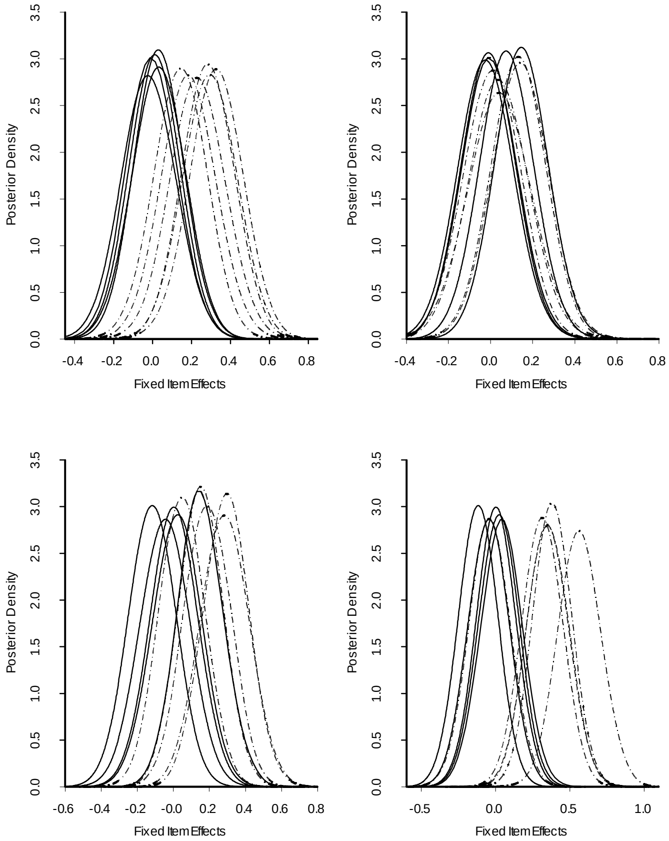

FIGURE 5.1. Posterior distributions of fixed effects parameters for the four groups. (Clockwise from top-left: Male-non-Dutch, Dutch-Male, Dutch-female, non-Dutch-Female)

5.8 Discussion

Fixed effects IRT models consisting of difficulty parameters that are allowed to vary across groups, are discussed. In contrast to random effects item parameters, interest is focused on the fixed effects and not on the variance in item parameters across groups. Two applications are considered: (1) detecting nonignorable missing data, and (2) detecting and/or modeling DIF items. It was shown that the fixed effects IRT model can be used for detecting nonignorable missing data in combination with the splitter item technique. That is, the observations of the splitter item (observed/missing) defines the grouping of observed item response data, and the fixed effects parameters model item parameter differences between these groups. Significant fixed effects parameters indicate item parameter differences between groups. In the second simulation study, it was shown that the fixed effects parameters can comprehend DIF items since differences in item parameters between groups are properly modeled. So, the fixed effects IRT model can be used to measure a latent variable in the presence of DIF items. It can also be used to detect DIF items in combination with a Bayes factor for testing the hypothesis that the fixed effects are zero.

It was shown that the proposed MCMC method for simultaneously estimating all parameters yields acceptable estimates. The estimation method can handle the 2PL and 2PNO model in three comparable sampling steps. This analogy makes the implementation easier. In general, the 2PNO model may be preferred since it has some computational advantages.

It has been shown that the Bayes factor for testing the null-hypothesis that all fixed effects are zero follows from evaluating the marginal posterior distribution of the fixed effects parameters in the point zero. This approach can be extended to facilitate the computation of Bayes factors for other hypothesis concerning problems of choosing between alternative models. For example, in the same way it can be tested whether all item discrimination parameters are equal. Bayesian inference concerning the fixed effects IRT model can also be based on HPD regions. Therefore, HPD region can be defined for the fixed effects IRT model to test hypotheses by deciding if a given point lies inside or outside the confidence region. Then, for example, testing the equality of difficulty parameters across groups, all fixed effects are zero, can be done by computing the probability on a HPD region that just includes the point zero.

Finally, the extension of the fixed effects IRT model to capture differences in discrimination parameters across groups is easily done by extending the design matrix $\mathbf{x}$. In that case, the design matrix is extended with the latent variable and the fixed effects parameters represent difficulty and discrimination parameter differences across groups. Further research will also focus on population group differences in the distribution of the latent variable. The framework of the multilevel IRT model (Fox, 2004; Fox & Glas, 2001, 2003) can be used to model population differences on the latent variable but it assumes that item response curves are the same for all groups. Problems occur due to the fact that the fixed effects parameters and population parameters vary across the same groups, which results in an identification problem. A possible solution might be found in finding identifying constraints such that the scale of the latent variable is identified and common across groups, and item and fixed effects parameters can be estimated with respect to this scale.

Synopsis

The handling of nonignorable missing data in psychometrics is not fully developed, but in recent years the attention for these problems in the application of latent variable modeling (see for instance Moustaki, 1996; O'Muircheartaigh & Moustaki,1999; Moustaki & Knott, 2000, Holman & Glas, 2005) is much increased. In educational measurement, most literature and software packages ignore missing data. However, this is inappropriate when the ignorability principle defined by Rubin (1987) does not hold. In these cases, the estimation of parameters ignoring the missing data often leads to biased results (Holman & Glas, 2005).

This thesis discusses methods to detect nonignorable missing data and methods to adjust for the bias caused by nonignorable missing data, both by introducing a model for the missing data indicator using item response theory (IRT) models.

In Chapter 2, a model based procedure that handles nonignorable missing data in the framework of IRT is presented. The relevant IRT model for the observed data is estimated in combination with an IRT model for the missing data process. The two IRT models are connected by invoking the assumption that their latent person parameters have a joint multivariate normal distribution. The model parameters are estimated using marginal maximum likelihood. As an example, the generalized partial credit model is used to model the observed data while the Rasch model is used to model the missing data process. The simulation studies conducted with both dichotomous and polytomous data show that the bias in the item parameter estimates obtained ignoring the missing data process are reduced by using an explicit latent variable model for the missing data process in the estimation. The bias is further reduced when observed covariates are included in the IRT model for missing data in the estimation of the data.

The approach in Chapter 2 is further elaborated in Chapter 3 for a situation where a test is administered in a limited-time condition. The time limit condition leaves items at the end of the test unanswered by the examinee, i.e., the missing data in particular appear consecutively in the items at the end of a test. The cause of this missingness is usually related to the person's ability: the lower the ability the larger the number of items that are left unanswered at the end of the test. Thus, in this case the mechanism causing the missing data should not be ignored. Following the method in Chapter 2, the data are modeled using a combination of two IRT models: The observed response data are modeled by the generalized partial credit model (in particular, 2PL model) and the missing data are modeled by the sequential model also known as the steps model. Again, the two IRT models are connected by invoking the assumption that their latent person parameters have a joint multivariate normal distribution and the parameters are estimated using marginal maximum likelihood. Results of the simulation studies show that when the model for the missing data process is included in the estimation together with the model for the observed data, the bias in the item parameter estimates remains comparable to the base line obtained with ignorable missing data. Further, excluding the model for the missing data process leads to considerable bias, that increased with the extent of the violation of ignorability. A real data set was analyzed to assess the impact of the model in practice. Specifically, including the missing data process lead to an increase of the estimate of the global reliability of the test.

In Chapter 4, two methods based on the splitter item technique are proposed to detect a nonignorable missing data process. So the method aims at making decisions whether the missing data are ignorable or not. The sample of respondents is divided into two groups. The first group consists of respondents that have an observation on the splitter item and the second group consists of respondents that do not have a response on the splitter item. Then, it is tested whether the item parameter estimates differ across the two groups. Two methods are considered. Both apply to IRT models for binary or ordinal responses estimated using a Bayesian method and MCMC. In the first method, all parameters of an IRT model for binary or ordinal responses are estimated given the subsets of item response data. Then, summary statistics of the estimated marginal posterior distributions of the item parameters are used for detection of differences. In the second method, values of IRT model parameters and, as an additional sampling step, values of so-called Bayesian modification indices (BMI) are sampled using MCMC. These BMI values provide information regarding any fluctuations in item parameter values across groups. They are estimated using MCMC and do not interfere with the estimation of the other model parameters. So the BMI values are obtained as a by-product of the MCMC algorithm for estimating the parameters of an IRT model. It is shown that the BMI distribution is a good approximation of the true marginal posterior distribution of the group-specific item parameters.

The last chapter, Chapter 5 of this thesis discusses fixed effects IRT models that include item parameters that are allowed to vary across groups. The models are used for modeling group specific item parameters. The proposed models were applied to detection of nonignorable missing data and for detecting and modeling differential item functioning (DIF). For the detection of nonignorable missing data, a splitter item defines the grouping of the item response data. The partitioning of the sample is based on whether or not there is a response on the splitter item. The fixed effects item parameters (as opposed to the often used random effects item parameters) model the group effects. Significant fixed effect parameters indicate that the item parameters differ between groups. The estimates are computed in a Bayesian framework using MCMC. The MCMC estimation method handles the 2PL and 2PNO model in three comparable sampling steps. The Bayesian inference concerning the fixed effects IRT model is based on HPD regions and Bayes factors. The HPD region is defined for the fixed effects IRT model in testing hypotheses in deciding if a given point lies inside or outside the confidence region. This way the null-hypothesis stating that all fixed effects are zero (the group-specific item parameters are equal) can be tested.

It is shown that a Bayes factor can be used for testing comparable null-hypotheses. Simulation studies are used to evaluate the performance of the procedure.

Samenvatting

Methoden voor analyses van data met niet-negeerbare ontbrekende gegevens (non-ignorable missing data) zijn in de psychometrie nog niet volledig ontwikkeld, maar de afgelopen jaren is de aandacht voor het probleem van ontbrekende gegevens in relatie tot de schatting van parameters in modellen met latente variabelen sterk toegenomen (Moustaki, 1996; O'Muircheartaigh & Moustaki,1999; Moustaki & Knott, 2000, Holman & Glas, 2005). In de meeste literatuur over onderwijskundig meten en de meeste software die in dat kader gebruikt wordt, worden ontbrekende gegevens genegeerd. Dit is echter niet correct als niet aan het negeerbaarheidprincipe (ignorability principle, Rubin, 1987) voldaan is. In die gevallen zijn de parameterschattingen onder een model waarbij de ontbrekende gegevens genegeerd worden bijzonder onzuiver (Holman & Glas, 2005).

In dit proefschrift wordt een aantal methoden voor het ontdekken van niet-negeerbare ontbrekende gegevens en methoden voor het corrigeren van de schattingen gepresenteerd. In beide gevallen gebeurt dit door het postuleren van een item response theorie (IRT) model voor de indicator voor de ontbrekende gegevens.

In hoofdstuk 2 wordt een op een IRT model gebaseerde methode voor de het analyseren van data met niet-negeerbare ontbrekende gegevens gepresenteerd. Het IRT model voor de geobserveerde data wordt simultaan geschat met een IRT model voor de indicator voor de ontbrekende gegevens. De twee IRT modellen worden verbonden via de veronderstelling dat hun latente persoonsparameters een gezamenlijke multivariaat normale verdeling hebben. De parameters van het complete model worden geschat met een marginale grootste-aannemelijkheid methode (marginal maximum likelihood estimation method). Als voorbeeld wordt het gegeneraliseerde partial credit model gebruikt voor de geobserveerde data en het Rasch model voor de indicator voor de ontbrekende gegevens. Met simulatiestudies wordt aangetoond dat zowel voor dichtoom als voor polytoom gescoorde antwoorden, de onzuiverheid in de schattingen gereduceerd wordt door de introductie van een IRT model voor de indicator voor de ontbrekende gegevens. De onzuiverheid wordt verder gereduceerd als er covariaten in dit laatste model worden opgenomen.

Deze aanpak wordt verder uitgewerkt in hoofdstuk 3, voor een test die is afgenomen onder tijdsdruk. Door tijdsdruk worden items aan het eind van de test niet beantwoord. Het patroon van de ontbrekende gegevens hangt meestal samen met het vaardigheidsniveau van de studenten: hoe lager het vaardigheidsniveau, hoe meer items aan het eind van de test niet gemaakt worden. Daarom mag het mechanisme dat de ontbrekende gegevens veroorzaakt heeft niet worden genegeerd. Net als in hoofdstuk 2 worden de data gemodelleerd met een combinatie van twee IRT modellen: de observaties worden gemodelleerd met het gegeneraliseerde partial credit model, terwijl de indicator voor de ontbrekende gegevens wordt gemodelleerd met het z.g. sequentiële model, c.q. het stapjesmodel. Ook hier worden de twee IRT modellen worden verbonden via de veronderstelling dat hun latente persoonsparameters een gezamenlijke multivariaat normale verdeling hebben en de schattingen worden berekend met een marginale grootste-aannemelijkheid methode. Met simulatiestudies wordt aangetoond dat de schattingen met behulp van dit model met data met niet-negeerbare ontbrekende gegevens dezelfde precisie hebben als schattingen met geobserveerde data met negeerbare ontbrekende gegevens via een IRT model zonder extra model voor de indicator variabele. Verder blijkt ook hier dat het negeren van niet-negeerbare ontbrekende gegevens leidt tot ernstige onzuiverheid in de parameterschattingen. Om de impact van de methode in de praktijk te evalueren wordt reële data van een toets, gemaakt onder tijdsdruk, geanalyseerd. Met deze data wordt getoond dat het modelleren van de indicator voor ontbrekende gegevens kan leiden tot een belangrijke verandering van de schatting van de betrouwbaarheid.

In hoofdstuk 4 worden twee methoden voor het ontdekken van niet-negeerbare ontbrekende gegevens voorgesteld die gebaseerd zijn op de splitter-item techniek. Dus het doel van de methode is om vast te stellen of de ontbrekende gegevens negeerbaar zijn, of niet. Hiertoe wordt de steekproef van studenten verdeeld in twee groepen. De eerste groep bestaat uit studenten die een antwoord gaven op het splitter-item, de tweede groep bestaat uit studenten waar het splitter-item niet beantwoord is. Daarna wordt getoetst of the schattingen van de itemparameters verschillen voor de twee groepen. Dit wordt gedaan met twee methoden. Beide methoden hebben betrekking op IRT modellen voor binaire of ordinale responsie in een Bayesiaans raamwerk. De schattingen worden berekend met een iteratief simulatieproces dat bekend staat onder het acroniem MCMC. In de eerste methode worden alle itemparameters van een model voor binaire of ordinale responsie geschat op de twee deelsteekproeven. Daarna worden functies van geschatte marginale a-posteriori verdelingen van de itemparameters gebruikt om verschillen vast te stellen. In de tweede methode worden in de MCMC procedure via simulatie zogenaamde Bayesiaanse modificatieindices (BMI) gegenereerd. Deze indices geven informatie over de verschillen tussen de itemparameters tussen de groepen. De simulatie van de indices heeft geen invloed op de simulatie van de modelparameters; de gesimuleerde indices zijn een bijproduct van de MCMC simulatie. Er wordt aangetoond dat de verdeling van deze indices een goede benadering is van de verdeling de groepsafhankelijke itemparameters.

In het laatste hoofdstuk, hoofdstuk 5, wordt een zogenaamd fixed-effects IRT model besproken waarin de itemparameters kunnen variëren tussen groepen. Deze modellen worden gebruikt voor het modelleren van groepsspecifieke itemparameters. De modellen kunnen worden toegepast voor het opsporen van niet-negeerbare ontbrekende gegevens en voor het opsporen van vraagonzuiverheid (differential item functioning, DIF). Wanneer het doel is het opsporen van niet-negeerbare ontbrekende gegevens wordt opnieuw een splitter-item gebruikt voor het groeperen van de responsiedata. Ook hier is de indeling van de steekproef gebaseerd op het al of niet geven van een response op het splitter-item. Fixed-effects itemparameters (in tegenstelling tot de vaak gebruikte random effects itemparameters) modelleren de groepseffecten. Significante waarden van de fixed-effects itemparameters geven aan dat de itemparameters variëren tussen groepen. De parameter schattingen worden berekend met een MCMC methode. Het toetsen van de hypothese dat alle fixed-effects gelijk zijn aan nul (de itemparameters variëren niet tussen de groepen) kan worden gedaan op basis van een betrouwbaarheidsinterval en op basis van een Bayesiaanse toets die bekend staat als Bayes factor. De methoden worden geëvalueerd met behulp van onderzoek op basis van simulaties.

Bibliography

Ackerman, T.A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement 20,* 309-310.

Ackerman, T.A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement 20,* 311-329.

Adams, R.J., Wilson, M.R., & Wu, M. (1997). Multilevel item response theory models: an approach to errors in variables of regression. *Journal of Educational and Behavioral Statistics, 22,* 47-76.

AERA, APA, & NCME. (1985). *Standards for educational and psychological tests.* Washington DC: American Psychological Association, American Educational research Association, National Council on Measurement in Education.

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

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

Allen, N.L., Holland, P.W.,& Thayer, D.T. (2005). Measuring the benefits of examinee-selected questions. *Journal of Educational Measurement, 42*, 27-51.

Allison, P.D. (2001). *Missing data.* Thousand Oaks, CA: Sage Publications.

Andersen, E.B. (1973). *Conditional inference and models for measuring.* Unpublished dissertation, Mentalhygienisk Forskningsinstitut, Copenhagen.

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

Angoff, W.H. (1993). Perspective on differential item functioning methodology. In P.W.Holland & H. Wainer (Eds.), *Differential item functioning* (pp.3-23). Hillsdale, N.J.: Erlbaum.

Baker, S.G. & Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association, 83*, 62-69.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York, NJ: Dekker.

Baker, F.B. (1998). An investigation of item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement 22,* 153-169.

Bartholomew, D.J.,& Knott, M. (1999). *Latent variable models and factor analysis* (2nd edition). London: Oxford University Press

Bechger, T.M., Maris, G., Verstralen,H.H.F.M.,& Béguin, A.A.(2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27,*319-334

Béguin, A.A., & Glas, C.A.W. (1998). *MCMC estimation of multidimensional IRT models. [*Research Report 98-14], University of Twente, Enschede.

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

Berger, J.O., & Delampady, M. (1987). Testing precise hypothesis. *Statistical Science, 2*, 317-352.

Berger, M.P.F. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika, 57,* 521-538.

Bernaards, C.A. (2000).*Nonresponse and factor analysis, with applications to rating scale data.* Unpublished doctoral thesis, Univeriteit Utrecht,The Netherlands.

Bernaards, C.A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research, 34(3),* 277-313.

Bernaards, C.A. , & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35(3),* 321-364.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores.* (pp.395-479). 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., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement 6,* 431-444.

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: Springer.

Box, G.E.P., & Tiao, G.C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wiley.

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.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets.*Psychometrika, 64*, 153-168.

Bradlow, E.T., & Zaslavsky, A.M. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *Journal of the American Statistical Association, 94*, 43-52.

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

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note. *The American Statistician, 36,* 153-157.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Chang, H.-H., & Ying, Z. (1999). Nonlinear sequential designs for logistic item response theory models, with applications to computerized adaptive tests. *Annals of Statistics.* (In press).

Chen, M.-H., & Shao, Q.-M.,& Ibrahim J.G. (2000). *Monte Carlo methods in Bayesian computation.* New-York: Springer-Verlag.

Chib, S.(1995). Marginal likelihood from the Gibbs output.*Journal of American Statistical Association, 90(432)*, 1313-1321.

Cochran, W.G. (1977). *Sampling techniques.* New York, NJ: Wiley.

Cole, N.S. (1993). History and development of DIF. In P.W.Holland & H. Wainer (Eds.), *Differential item functioning* (pp.25-29). Hillsdale, NJ.: Erlbaum.

Cook, T.D., & Campbell, D.T. (1979). *Quasi- experimentation, design & analysis issues for field settings.* Chicago, IL: Rand McNally College Publishing Company.

Coombs, C.H. (1960). *A theory of data.* Ann Arbor, MA: Mathesis Press.

Coombs, C.H., Dawes, R.M.,& Tversky, A. (1970). *Mathematical psychology: An elementary introduction.* Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Conaway, M.R.(1992). The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association: Theory and Methods, 87*, 817-824.

Conaway, M.R.(1994). Causal nonresponse models for repeated categorical measurements. *Biometrics,50*, 1102-1116.

Congdon, P. (2002). *Bayesian statistical modeling.* New York: Wiley. Coombs, C.H., & Kao, R.C. (1955). *Nonmetric factor analysis.*[Engg. Res. Bull.,38]. Ann Arbor: University of Michigan Press.

Copas, A.J., & Farewell, V.T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-response' variable. *Journal of the Royal Statistical Society, A, 161*, 385-396

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

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

De Leeuw, E,D., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of official statistics, 19*153-176 (24).

De Leeuw, J., & Verhelst, N. D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183- 196.

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

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education, 15(1)*,15-31.

Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42*, 204-223.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. Dempster, N. Liard, & D. Rubin). *J. R. Statist. Soc.,B, 39,* 29.

Embretson, S.E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201-212.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fay, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association, 81*, 354-365.

Fischer, G.H. & Scheiblechner, H.H. (1970). Algorithmen und programme für das probabilistische testmodell von Rasch. *Psychologische Beiträge, 12,* 23-51.

Fischer, G.H. (1974). *Einführung in die theorie psychologischer tests introduction to the theory of psychological tests .* Bern: Huber.

Fischer, G.H. (1993). Notes on the Mantel -Haenszel procedure and another chi-square test for the assessment of DIF *Methodika, 7,* 88-100.

Fischer, G.H. (1995). Derivations of the Rasch model. In G.H.Fischer & I.W.Molenaar (Eds.), *Rasch models: foundations, recent developments and applications,* (pp.39-52). New York, NJ: Springer.

Fischer, G.H., & Molenaar I.W. (1995). *Rasch models. Their foundation, recent developments and applications.* New York, NJ: Springer.

Fox, J.P.(2001). *Multilevel IRT: A Bayesian perspective on estimating parameters and testing statistical hypothesis.* Unpublished Doctoral Dissertation, University of Twente, The Netherlands.

Fox, J.-P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269-286.

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

Fox, J.-P., & Glas, C.A.W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*, 169-191.

Fox, J.-P. (2004). Modelling response error in school effectiveness research. *Statistica Neerlandica, 58*, 138-160.

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

Fox, J.-P., & Hendrawan, I. (2005). Bayesian inference for linear models with latent dependent variables. *Submitted for publication.*

Fox, J.-P., Pimentel, J.L., & Glas, C.A.W. (2005). Detecting nonignorable missing data using the splitter item technique. *Submitted for publication.*

Fox, J.-P., Pimentel, J.L., & Glas, C.A.W. (2005). Fixed effect IRT model. *Submitted for publication.*

Fraser, C. (1988). NOHARM: *A Computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* NSW: University. of New England.

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. (2004). *Bayesian data analysis (2nd)*, New York: Chapman & 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. (1988). The Rasch model and multi-stage testing. *Journal of Educational Statistics, 13,* 45- 52.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* Arnhem: Cito.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* Unpublished doctoral thesis, Twente University, the Netherlands.

Glas, C.A.W. & Verhelst, N.D. (1989) Extensions of the partial credit model. *Psychometrika 54,* 635-659.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson, (Ed.), *Objective measurement: Theory into practice* (Volume 1) (pp. 236-258). Norwood, NJ: Ablex Publishing Corporation.

Glas C.A.W., & Ellis, J.L. (1993) *RSP, Rasch scaling program, computer program and user's manual.* Groningen: ProGAMMA.

Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In G.H.Fischer & I.W.Molenaar (Eds.), *Rasch models: foundations, recent developments and applications.* (pp.69-96). New York, NJ: Springer.

Glas, C.A.W., & Verhelst, N.D. (1995). Tests of fit for polytomous Rasch models. In G.H Fisher & I.W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (325-352). New York: Springer-Verlag.

Glas, C.A.W., & Béguin, A.A. (1996). *Appropriateness of IRT observed score equating.* Research Report 96-04, University of Twente, Enschede.

Glas, C.A.W. (1997). Testing the generalized partial credit model. In M. Wilson, G. Engelhard, Jr., & K. Draney (Eds.), *Objective measurement: Theory into practice, Vol. 4,* (pp.237-260). New Jersey, NJ: Ablex Publishing Corporation.

Glas, C.A.W. (1998). Detection of differential item functioning using lagrange multiplier tests. *Statistica Sininca, 8*, 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. (2000). Item calibration and parameter drift. In W.J. van der Linden & C.A.W.Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp.183-200). Boston MA: Kluwer-Nijhoff Publishing.

Glas, C.A.W., Wainer, H., & Bradlow (2000). MML and EAP estimates for the testlet response model. In W.J. van der Linden & C.A.W.Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp.271-287). Boston MA: Kluwer-Nijhoff Publishing.

Glas, C.A.W., & van der Linden, W.J. (2001a). *Computerized adaptive testing using item shells.* Twente University, OMD Research Report 01-10.

Glas, C.A.W., & van der Linden, W.J. (2001b). *Modeling variability in item parameters in educational measurement.* Twente University, OMD Research Report 01-11.

Glas, C.A.W. (2001). Differential item functioning depending on general covariates. In A. Boomsma, M.A.J. van Duijn, & T.A.B. Snijders (Eds.), *Essays on item response theory*(131-148). New York: Springer-Verlag.

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.

Glas, C.A.W. (2005) Structural item response models. *In Encyclopedia of social measurement, Volume 2*, 697-704. Elsevier Inc.

Green, P.E., & Park, T.A. (2003). A Bayesian hierarchical model for categorical data with nonignorable nonresponse. *Biometrics, 59*, 886-896.

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

Hambleton, R. K., & Swaminatan, H. (1985). *Item response theory: Principles and applications* (2nd edition). Boston MA: Kluwer Academic Publishers.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrika,* 46, 931-961.

Heitjan, D.F.(1994). Ignorability in general incomplete-data models, *Biometrika*,**81(4)**,701-708.

Hendrawan, I. (2004). *Statistical test of item response models: power and robustness.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G.H.Fischer & I.W.Molenaar (Eds.), *Rasch models: foundations, recent developments and applications.* (pp.53-68). New York, NJ: Springer.

Holman,R.,& Glas, C.A.W. (2005). Modelling non-ignorable missing data mechanism with item response theory models.*British Journal of Mathematical and Statistical Psychology,58*,1-18

Holman,R. (2005). *Statistical item response theory in clinical outcome measurement.* Unpublished doctoral dissertation, University of Amsterdam, The Netherlands

Huisman, M. (1998). Missing data in behavioral science research: Investigation of a collection of data sets. *Kwantitatieve Methoden, 57,* 69-93.

Huisman. M. (1999).*Item nonresponse: Occurence, causes, and imputation of missing answers to test items.*Unpublished Doctoral thesis, Rijksuniversiteit Groningen, The Netherlands.

Jansen, M.G.H. (1997). Rasch model for speed tests and some extensions with applications to incomplete designs. *Journal of Educational and Behavioral Statistics, 22,* 125-140.

Jansen, M.G.H., & Glas, C.A.W. (2000). Statistical tests for differential test functioning in Rasch's model for speed tests. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on item response theory* (pp.149-162). New York, NJ: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285-306.

Jeffreys, H. (1961). *Theory of probability (3rd Ed.).* Oxford: Clarendon Press.

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

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54*, 681-697.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887-903.

Klugkist,I. (2004). *Inequality constrained normal linear models.* Unpublished doctoral dissertation, University of Utrecht, The Netherlands.

Leeuw, E.D. de, Hox, J.J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics, 19*, 153-176.

Lehmann, E.L. (1983). *The theory of point estimation.* New York: Springer.

Lindley, D.V. (1965). *Introduction to probability of statistics from a Bayesian viewpoint (2 vols - Part I: Probability and Part II: Inference).* Cambridge: Cambridge University Press.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association, 77*, 237-250.

Little, R.J.A., & Rubin, D.B. (1987) *Statistical analysis with missing data.* New York: Wiley

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

Lord, F.M. (1983). Unbiased estimators of ability parameters, their variance and of their parallel-forms reliability. *Psychometrika, 48,* 233-245.

Lord, F.M. (1983). Small *N* justifies Rasch model. In D.J.Weiss (Ed.), *New horizons in testing.* (pp. 51-61). New York, NJ: Academic Press.

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

Lunn, D.J.,Thomas, A.,Best,N.,& Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility *Statistics and Computing, 10* , 325-337.

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.

Maris, G., & Maris, E. (2002). A MCMC-method for models with continuous latent responses. *Psychometrika, 67*, 335-350.

Masters,G.N. (1982) A Rasch model for partial credit scoring. *Psychometrika, 47, No.2.*

Masters, G.N., & Wright, B.D. (1997) The partial credit model. In van der Linden, W.J. & Hambleton, R.K. *Handbook of modern item response theory.* New York: Springer.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psyhometric monogaphs,* No. 15.

McDonald, R.P. (1997). Normal-Ogive multidimensional model. In van der Linden, W. J., & Hambleton, R. K. *Handbook of modern item response theory.* New York: Springer.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: a review and new developments, *Applied Measurement in Education*, *8*, 261-272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Mellenbergh, G.J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research 29,* 223-236.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R.J., & Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661-680.

Mislevy, R.J., & Bock, R.D. (1990). *PC-BILOG. Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Mislevy, R.J., & Wu, P.K. (1996). *Missing responses and IRT ability estimation: omits, choice, time limits, and adaptive testing.* ETS Research Report RR-96-30-ONR. Princeton, NJ: Educational Testing Service.

Mislevy, R.J., & Chang, H.H. (1998). *Does adaptive testing violate local independence?* ETS Research Report RR-98-33. Princeton, NJ: Educational Testing Service.

Mislevy, R.J., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65,* 149-156.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* Den Haag: Mouton.

Molenaar, I.W. (1983). *Item steps.* Heymans Bulletins Psychologische instituten R.U.Groningen, nr. HB-83-630-EX.

Molenaar, I.W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika,48*, 49-72.

Molenaar, I.W. (1995). Estimation of item parameters. In G. H. Fischer, & I.W.Molenaar (Eds.), *Rasch models: foundations, recent developments and applications.* New York, NJ: Springer.

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. (2000). A latent Variable model for Ordinal Variables. *Applied Psychological Measurement,24(3),* 211-223.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika, 65,* 391-411.

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, A, 163,* 445-459.

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

Muraki, E., Mislevy, R. J., & Bock, R. D. (1987). *BIMAIN: A program for item pool maintenance in the presence of item parameter drift and item bias [software manual].* Mooresville, IN:Scientific Software.

Muraki, E., & Bock, R.D. (1991). *PARSCALE: Parameter scaling of rating data* [computer program]. Chicago, IL: Scientific Software, Inc.

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

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered, categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, L. K., & Muthén, B. O.(1998). *MPLUS: The comprehensive modeling program for applied researcher, user's guide.* Los Angeles, CA: Muthén & Muthén.

Neyman, J. & Scott, E.L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16*, 1-32.

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.

Park, T., & Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association, 89,* 44-52.

Park, T. & Brown, M.B. (1997). Loglinear models for a binary response with nonignorable nonresponse. *Compuational Statistics and Data Analysis, 24*, 417-432.

Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

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.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.). *Handbook of modern item response theory* (pp.271-286). New York: Springer.

Rigdon S.E., & Tsutakawa, R.K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567-574.

Ripley, B.D. (1987). *Stochastic simulation.* New York: Wiley.

Robert, C.P., & Casella, G. (1999). *Monte Carlo statistical methods.* New York, NY: Springer.

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

Rubin, D.B. (1978). Multiple imputations in sample surveys- a phenomenological Bayesian approach to nonresponse. In: *Imputations and editing of faulty or missing survey data*(pp.1-23), U.S. Department of Commerce.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York:John Wiley & Sons.

Schafer J.L. (1997). *Analysis of incomplete multivariate data.* New York, NJ: Chapman and Hall.

Schafer,J.L., & Graham,J.W.(2002). Missing data: Our view of the state of the art.*Psychological methods, 7*, 147-177.

Scheerens,J., Glas, C.A.W.,& Thomas, S.M. (2003). *Educational evaluation, assessment, and monitoring : a systemic approach.* Lisse: Swets & Zeitlinger.

Sijtsma, K., & Van der Ark, L.A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505-528.

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

Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation for the one-parameter logistic model, *Psychometrika, 47,* 349-364.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66,* 331-342.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46*, 359-372.

Tanner, M.A. (1993). *Tools for statistical inference.* New York: Springer.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-110.

Thissen D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175-186.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), *Test validity*, (pp.147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice*, (pp.67-113). Hillsdale, NJ: Erlbaum.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39-55.

van den Wollenberg, A.L. (1979). *The Rasch model and time limit tests.* Unpublished doctoral dissertation, University of Nijmegen, The Netherlands.

van den Wollenberg, A.L. (1979). *The Rasch model and time limit tests.* Nijmegen: Studentenpers.

van den Wollenberg, A.L. (1982). Two new tests for the Rasch model. *Psychometrika, 47,* 123-140.

van der Linden, W. J.(2005). Classical test theory. *In encyclopedia of social measurement, volume 2*, 301-307. Elsevier Inc.

van der Linden, W. J.(2005). Item response theory. *In encyclopedia of social measurement, volume 2*, 379-387. Elsevier Inc.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer Verlag.

van Dijk, H., & Tellegen, P. (2004). *NIO, Nederlandse Intelligentietest voor Onderwijsniveau.* Amsterdam: Boom Test Uitgevers.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association, 90,* 614-618.

Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.J.H.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices discrimination indices in the one parameter logistic model. elementary symmetric functions and their first and second derivatives.* Measurement and Research Department reports, 91-10. Arnhem: Cito.

Verhelst, N.D., & Verstralen, H.H.F.M. (1991). *The partial credit model with non-sequential solution strategies. Measurement and Research Department reports, 91-5,* Arnhem, Cito.

Verhelst, N.D., & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58,* 395-415.

Verhelst, N.D. (1993). *On the standard errors of parameter estimates in the Rasch model.* Measurement and Research Department Reports, 93-1. Cito: Arnhem.

Verhelst, N.D., & Glas, C.A.W. (1995a). Dynamic generalizations of the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: their foundations, recent developments and applications* (pp. 181-202). New York: Springer.

Verhelst, N.D., & Glas, C.A.W. (1995b). The generalized one parameter model: OPLM. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: their foundations, recent developments and applications* (pp. 213-237). New York: Springer.

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 Measurement, the Netherlands.

Verhelst, N.D., Glas, C.A.W., & de Vries, H.H. (1997). A steps model to analyze partial credit. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory.* (pp.123-138). New York : Springer.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3pl model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 245-269). Boston, MA: Kluwer Academic Publishers.

Wetherill, G.B. (1977). *Sampling inspection and statistical quality control* (2nd edition). London: Chapman and Hall.

Wilson, D.T., Wood, R., & Gibbons, R.D. (1991) *TESTFACT: Test scoring, item statistics, and item factor analysis* (computer software). Chicago: Scientific Software International, Inc.

Wilson, M., & Masters, G.N. (1993). The partial credit model and null categories. *Psychometrika, 58,* 85-99.

Wingersky, M.S., Barton M.A., & Lord, F.M. (1982). *LOGIST.* (Computer software). Princeton, NJ: Educational Testing Service.

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

Wu, M.L., Adams, R.J. & Wilson, M.R. (1997). *ConQuest*: *Generalised item response modelling software, Draft release 2.* Australian Council for Educational Research.

Yen, W.M. (1981). Using simultaneous results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Zimowski, M.F.,Muraki, E.,Mislevy,R.J.,& Bock,R.D.(1996).*Bilog MG:Multiple-group IRT analysis and test maintenance for binary items.*Chicago: Scientific Software International Inc.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (2002). *Bilog-MG.* Lincolnwood, IL, Scientific Software International Inc.