Behaviormetrika
Vol.33, No.1, 2006, 1–16

Fixed Effects IRT Model

Jean-Paul Fox*, Jonald Pimentel*, and Cees Glas*

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.

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

Key Words and Phrases: analysis of variance, differential item functioning, fixed effect, item response theory model, MCMC.

* University of Twente, Department of research methodology, measurement and data analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands.

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.

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

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

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

$$b_{kj} = b_k + \lambda_{kj}, \tag{2}$$

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

$$P\big(y_{ik} = 1 \mid \theta_i, a_k, b_k, \boldsymbol{\lambda}_k\big) = \begin{cases} \big[1 + \exp(-D(a_k\theta_i - (b_k + \mathbf{x}_i^t\boldsymbol{\lambda}_k)))\big]^{-1} & \text{for the 2PL} \\ \Phi\big(a_k\theta_i - (b_k + \mathbf{x}_i^t\boldsymbol{\lambda}_k)\big) & \text{for the 2PNO.} \end{cases} \tag{3}$$

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 (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.

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 $\boldsymbol{\lambda}$, $p(\boldsymbol{\lambda} \mid \mathbf{y})$. 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.

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, eg., 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,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:

$$P\big(y_{iksr} = 1 \mid \theta_i, a_k, b_k, \lambda_{k1s}, \lambda_{k2r}\big) = \Phi\big(a_k\theta_i - (b_k + \lambda_{k1s} + \lambda_{k2r})\big), \tag{4}$$

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 (3). Note that interaction effects between gender and nations are easily incorporated.

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.

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

$$z_{ik} \mid y_{ik}, \theta_i, a_k, b_k, \boldsymbol{\lambda}_k \sim \begin{cases} \mathcal{L}(0,1) & \text{for the 2PL} \\ \mathcal{N}(0,1) & \text{for the 2PNO,} \end{cases} \tag{5}$$

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

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

  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 $\mathbf{a}$, $\mathbf{b}$, $\boldsymbol{\lambda}$. 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 $\mathbf{z}_k^*$,

$$\begin{aligned}\mathbf{z}_k^* &= D(a_k\boldsymbol{\theta} - (b_k + \mathbf{x}\boldsymbol{\lambda}_k)) + \boldsymbol{\epsilon}_k \\ \mathbf{z}_k^* &= \mathbf{H}\Xi_k + \boldsymbol{\epsilon}_k\end{aligned} \tag{6}$$

where $\mathbf{H} = D(\boldsymbol{\theta}, -\mathbf{1}, -\mathbf{x})$, $\Xi_k = (a_k, b_k, \boldsymbol{\lambda}_k)^t$, 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

$$\Xi_k \mid \mathbf{z}_k^*, \boldsymbol{\theta} \sim \mathcal{N}\Big(\hat{\Xi}_k, c\big(\mathbf{H}^t\mathbf{H}\big)^{-1}\Big)p(\Xi_k), \tag{7}$$

where

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

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 (7) for the fixed effects 2PL model.

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

$$1 - \alpha \leq P\big(C_{\boldsymbol{\lambda}}(\mathbf{y}) \mid \mathbf{y}\big) = \int_{C_{\boldsymbol{\lambda}}(\mathbf{y})} p\big(\boldsymbol{\lambda} \mid \mathbf{y}\big)d\boldsymbol{\lambda}. \tag{8}$$

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

$$H_0 : |\boldsymbol{\lambda} - \boldsymbol{\lambda}_0| \leq \epsilon \text{ versus } H_1 : |\boldsymbol{\lambda} - \boldsymbol{\lambda}_0| > \epsilon, \tag{9}$$

where ϵ is "small". The point null hypothesis will be seen as an approximation for the small interval null as in Equation (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 $\Xi = (\mathbf{a}, \mathbf{b}, \boldsymbol{\theta})$ and assume that $p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid \mathbf{y}, M) = p(\Xi \mid \mathbf{y}, M_0)$. 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):

$$\begin{aligned} p(\mathbf{y} \mid M_0) &= \int p(\mathbf{y} \mid \Xi, M_0) p(\Xi \mid M_0) d\Xi \\ &= \int \frac{p(\Xi \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid M)} p(\boldsymbol{\lambda} = 0, \Xi \mid M) p(\mathbf{y} \mid \boldsymbol{\lambda} = \mathbf{0}, \Xi, M) d\Xi \\ &= p(\mathbf{y} \mid M) \int \frac{p(\Xi \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid M)} p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid \mathbf{y}, M) d\Xi. \end{aligned} \tag{10}$$

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

$$\begin{aligned} BF &= \int \frac{p(\Xi \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \Xi, \mathbf{y}, M) p(\Xi \mid \mathbf{y}, M) d\Xi \\ BF &= \mathcal{E}\left[\frac{p(\Xi \mid M_0)}{p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \Xi, \mathbf{y}, M) \right], \end{aligned} \tag{11}$$

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

$$\widehat{BF} = M^{-1} \sum_m \frac{p\big(\Xi^{(m)} \mid M_0\big)}{p\big(\boldsymbol{\lambda} = \mathbf{0}, \Xi^{(m)} \mid M\big)} p\big(\boldsymbol{\lambda} = \mathbf{0} \mid \Xi^{(m)}, \mathbf{y}, M\big). \tag{12}$$

A special case occurs when $p(\Xi \mid \boldsymbol{\lambda} = \mathbf{0}, M) = p(\Xi \mid M_0)$. Via Equation (11) it follows that

$$\begin{aligned} BF &= \int \frac{p(\Xi \mid \boldsymbol{\lambda} = \mathbf{0}, M)}{p(\boldsymbol{\lambda} = \mathbf{0}, \Xi \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \Xi, \mathbf{y}, M) p(\Xi \mid \mathbf{y}, M) d\Xi \\ &= \int \frac{1}{p(\boldsymbol{\lambda} = \mathbf{0} \mid M)} p(\boldsymbol{\lambda} = \mathbf{0} \mid \Xi, \mathbf{y}, M) p(\Xi \mid \mathbf{y}, M) d\Xi \\ &= \frac{p(\boldsymbol{\lambda} = \mathbf{0} \mid \mathbf{y}, M)}{p(\boldsymbol{\lambda} = \mathbf{0} \mid M)}, \end{aligned} \tag{13}$$

which is known as the Savage-Dickey density ratio (Dickey, 1971). Note that the Bayes factor in Equation (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.

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.

7.1 Simulation Study 1

Analogous to Bradlow and Thomas (1998), and Fox, Pimentel, and Glas (2005) 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 (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 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

Table 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	$-$	$-$	$-$	$-$

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.

7.2 Simulation Study 2

In this numerical example, data were analyzed to investigate the performance of the fixed effects 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) and nations (Dutch, non-Dutch) where a Dutch female was coded as $x_1 = 1$ and $x_2 = 1$. 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.

Table 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]

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 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 **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 and standard deviation of the fixed parameter estimates, $\boldsymbol{\lambda}$ are 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.

In Table 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 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

Table 3: Parameter estimates of the fixed effects IRT model given data generated under model M_3.

			$p(\mathbf{b} \mid \mathbf{y}, \boldsymbol{\lambda} = 0)$		$p(\mathbf{b} \mid \mathbf{y}, x_1 = 0, x_2 = 0)$	
	Item	**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

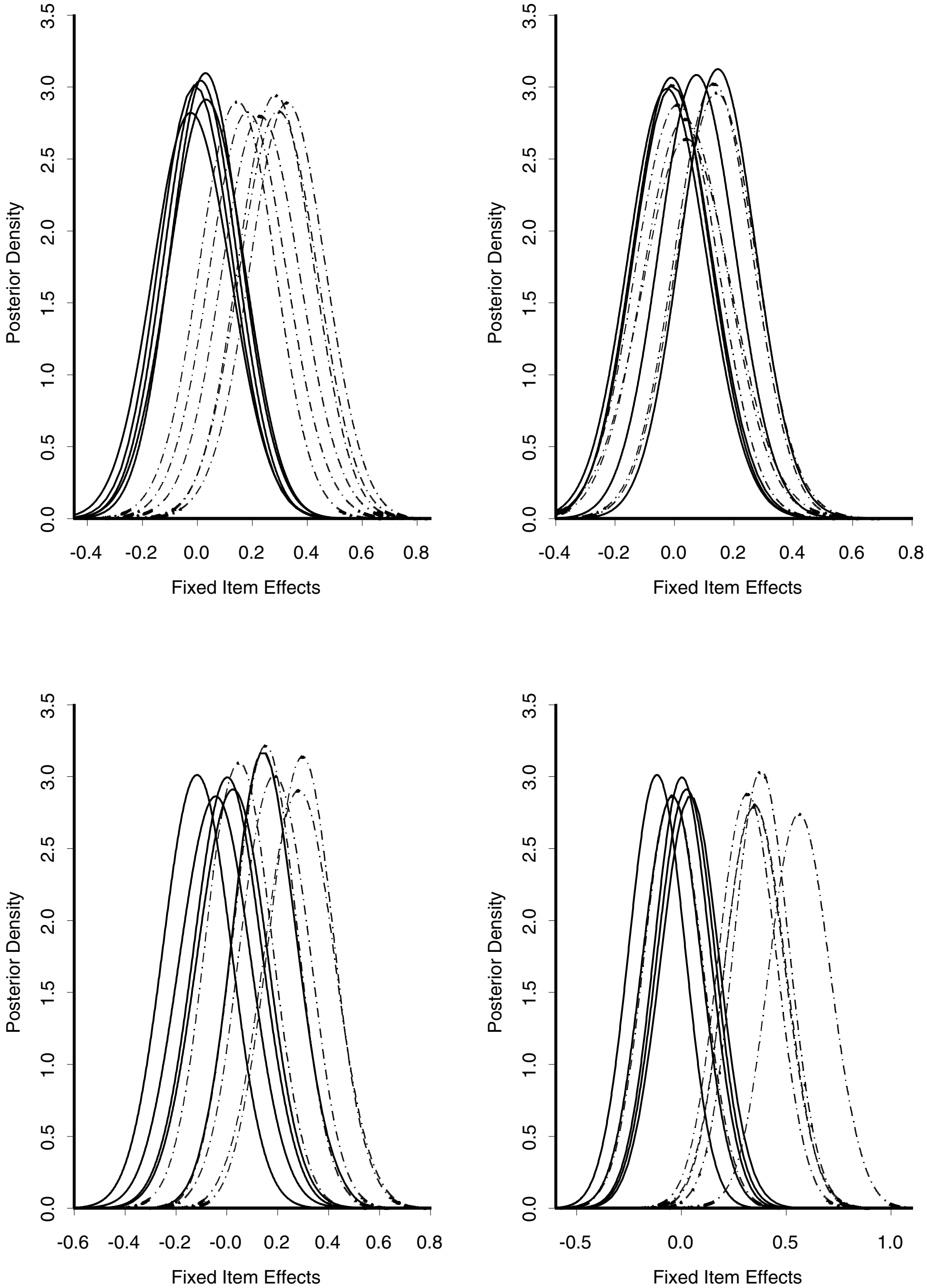

Figure 1: Posterior distributions of fixed effects parameters for the four groups.

.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.

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.

References

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

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

Bernaards, C.A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research, 34*, 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*, 321–364.

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.

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

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.

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

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

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. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*, 169–191.

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

Glas, C.A.W. (1998). Detection of differential item functioning using lagrange multiplier tests. *Statistica Sininca, 8*, 647–667.

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 The-*

ory (131–148). New York: Springer-Verlag.

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.

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.

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*, 1–17.

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.

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

Klugkist, I. (2004). *Inequality Constrained Normal Linear Models.* Unpublished doctoral dissertation, University of Utrecht, The Netherlands.

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.

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

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

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.

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.

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

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

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*, (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*, (67–113). Hillsdale, NJ: Erlbaum.

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

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.