MULTILEVEL IRT:
A BAYESIAN PERSPECTIVE ON ESTIMATING PARAMETERS AND TESTING STATISTICAL HYPOTHESES

JEAN-PAUL FOX
University of Twente, Enschede

Samenstelling promotiecommisie

Voorzitter/secretaris	Prof. dr. J.M. Pieters
Promotor	Prof. dr. C.A.W. Glas
	Prof. dr. W.J. van der Linden
Leden	Prof. dr. R.J. Bosker
	Prof. dr. J.J.C.M. Hox
	Dr. W.C.M. Kallenberg
	Prof. dr. T.A.B. Snijders
	Prof. dr. N.D. Verhelst

Fox, Jean-Paul

Multilevel IRT: A Bayesian perspective on estimating parameters and testing statistical hypotheses / Jean-Paul Fox

Proefschrift Universiteit Twente, Enschede. - Met lit. opg. - Met samenvatting in het Nederlands.

ISBN: 90-365-1640-4

Druk: PrintPartners Ipskamp B.V., Enschede

MULTILEVEL IRT:
A BAYESIAN PERSPECTIVE ON ESTIMATING PARAMETERS AND TESTING STATISTICAL HYPOTHESES

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van rector magnificus,
prof. dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 7 september 2001 te 15.00 uur.

door

Gerardus Johannes Andre Fox

geboren op 20 december 1970
te Haaksbergen

Acknowledgments

This thesis is the result of my four-year research at the Department of Educational Measurement and Data Analysis (OMD) at the University of Twente. I thank my colleagues for the pleasant and supportive atmosphere that was always present. I have been fortunate with the opportunity of writing this dissertation at OMD.

In particular I would like to thank Cees Glas for providing corrections and guidance for clarifying some of the discussions, which wasn't always easy. I also like to thank Wim van der Linden for reviewing some of the chapters.

As the last one of *the group of four*, I would like to thank, Anton Béguin, Edith van Krimpen-Stoop, and Bernard Veldkamp for the coffee breaks in the good old days, and the necessary distraction from time to time. The discussions with Anton and Bernard were times not to forget.

My thanks also to my family for their support, in particular, my parents who always believed in me, but remained wondering what kept me busy all the time. Well, this is it.

Finally, I want to thank Miranda. Writing this thesis required much time and energy. In a period where you wanted to move on it must have felt as if time stood still. My sincere thanks for your support and patience.

Enschede, june 2001
Jean-Paul Fox

Contents

Acknowledgments **v**

List of Figures **ix**

List of Tables **xi**

List of Symbols **xiii**

1. Introduction **1**
1. Multilevel Data 1
2. Measurement Error 2
3. Objectives and Outline 4

2. The Effects of Measurement Error in a Multilevel Model **7**
1. School Effectiveness Research 7
2. Models for Measurement Error 9
3. Multilevel IRT 10
4. Ignorable and Non-Ignorable Measurement Error 14
5. An Illustrative Example 19
6. Discussion 24

3. Bayesian Estimation of a Multilevel IRT Model **29**
1. Introduction 29
2. Multilevel IRT Models 32
3. MCMC Estimation Procedure for a Multilevel IRT Model 33
4. Simulated and Real-Data Examples 40
5. Discussion 51

4. Bayesian Modeling of Measurement Error in Predictor Variables **55**
1. Introduction 55
2. The Structural Multilevel Model 57
3. Measurement Error Models 59

4. MCMC Estimation Procedure for a Multilevel Model with Measurement Error 61
5. Measurement Error in Correlated Predictor Variables 68
6. A Simulation Study 71
7. Example of Measurement Error in Hierarchical Models 76
8. Discussion 81

5. Bayesian Model Checking and Residual Analysis 85
1. Introduction 85
2. Bayesian Residual Analysis 87
3. Detection of Outliers 91
4. Heteroscedasticity 93
5. Choice of Priors 97
6. An Analysis of a Dutch Primary School Mathematics Test 100
7. Discussion 108

6. A Stochastic EM Approach 109
1. Introduction 109
2. A Multilevel IRT Model 111
3. The SEM Algorithm 113
4. Maximum Likelihood Estimation 114
5. Implementation of the SEM Algorithm 115
6. SEM in Comparison with the Gibbs Sampling Approach 121
7. A Dutch Primary School Language Test 122
8. Discussion 129

Epilogue 137

Summary in Dutch 141

References 145

Author Index 155

Subject Index 159

List of Figures

2.1 A path diagram of a Multilevel IRT model, where item response theory models measure the latent variables within the structural multilevel model. 11

2.2 Expected posterior estimates and predictions of the dependent values given the true independent variables. 25

2.3 Expected posterior estimates and predictions of the dependent values given that the explanatory variables at Level 1 and Level 2 are measured with an error. 26

3.1 Posterior densities of a_k for items 2, 5, 7 and 8. Dotted line is an estimate of density after 500 values, and solid line is an estimate of density after $20,000$ values. 43

3.2 Expected posterior estimate and prediction of students' abilities in a Cito and non-Cito school as a function of SES, controlling for ISI and Gender. 49

3.3 Students' N-scores and predicted N-scores in a Cito and non-Cito school as a function of SES, controlling for ISI and Gender. 50

4.1 The $E\left[L^2\right]$ for the multilevel IRT model and the multilevel true score model. 76

4.2 Density plots of the observed and replicated data using the normal ogive model and the classical true score model. 77

4.3 The $E\left[L^2\right]$ and $E\left[L_1^2\right]$ for different values of the error variance to model the latent predictor variables on Level 1 and Level 2 with the classical true score model. 81

5.1 Bayesian latent residuals plotted against the probabilities of a correct response and the outlying probabilities. 102
5.2 Posterior densities of the Bayesian latent residuals corresponding to Item 17 for a number of students. 103
5.3 Normal probability plot of standardized residuals at Level 1. 105
5.4 Testing heteroscedasticity at Level 1. 106
6.1 Plausible region for (σ, τ), generated by the stochastic EM algorithm. 128

List of Tables

2.1	Parameter estimates of the multilevel model with measurement error in the dependent variable.	21
2.2	Parameter estimates of the multilevel model with measurement error in both the dependent and independent variables.	23
3.1	Item parameter estimates of the normal ogive IRT model using the Gibbs sampler.	42
3.2	Parameter estimates of the multilevel model, with the Gibbs sampler and HLM for Windows.	44
3.3	Parameter recovery of the multilevel model with standardized true latent scores and Z-scores as dependent variables.	45
3.4	Parameter estimates of the multilevel model with the Gibbs sampler and HLM using N-scores and rescaled N-scores as dependent variables.	47
4.1	Item parameter estimates of the normal ogive IRT model at Level 1.	72
4.2	Parameter estimates of the multilevel model with measurement error in the covariates.	73
4.3	Parameter estimates of the multilevel model with the normal ogive and the classical true score model as measurement error models.	79
4.4	Parameter estimates of the multilevel model with the normal ogive and the classical true score model as measurement error models on both levels.	80
5.1	Parameter estimates of a multilevel IRT model with explanatory variable End on Level 2.	101

5.2 Parameter estimates of a multilevel IRT model with explanatory variables ISI and SES on Level 1 and End on Level 2. 104
5.3 Parameter estimates of a multilevel IRT model with explanatory variables ISI and SES on Level 1 and End on Level 2 using proper informative priors. 107
6.1 Parameter estimates of the discrimination parameter with SEM and the Gibbs sampler. 124
6.2 Parameter estimates of the difficulty parameter with SEM and the Gibbs sampler. 125
6.3 Parameter estimates of the multilevel model with the Gibbs sampler, stochastic EM, and HLM using sum scores. 126

List of Symbols

General Notation

$P(.)$	Probability of an event
$p(.)$	Marginal density function
$p(. \mid .)$	Conditional density function
$\Phi(.)$	Standard normal cumulative distribution
$\phi(.)$	Density of the standard normal distribution
$F(.)$	F distribution
$\chi^2(.)$	Chi-square distribution
$Inv-\chi^2(.)$	Inverse-chi-square distribution
$Inv-Wishart(.)$	Inverse-wishart distribution
N	Number of respondents
J	Number of groups
K	Number of items
n_j	Number of respondents within group j
ij	Index of respondent i in group j
p_{ijk}	Probability of a correct response of person ij on item k
$\mathbf{r}$	Realized residuals at the item level

Symbols of the Structural Multilevel model

$\boldsymbol{\theta}$	Dependent or independent latent variable(s) at Level 1
$\boldsymbol{\omega}$	Dependent latent variable at Level 1
$\boldsymbol{\zeta}$	Independent latent variables at Level 2
Λ	Independent variables at Level 1 without an error
Γ	Independent variables at Level 2 without an error
$\mathbf{Y}, \mathbf{y}$	Dependent variables at Level 1
$\mathbf{X}, \mathbf{x}$	Independent variables at Level 1
$\mathbf{W}, \mathbf{w}$	Independent variables at Level 2
Q	Maximum number of covariates at Level 1
S	Maximum number of covariates at Level 2
Ω	Set of independent variables at Level 1

Ω^-	Set of independent variables at Level 1 without $\boldsymbol{\theta}_q$
Ψ	Set of independent variables at Level 2
Ψ^-	Set of independent variables at Level 2 without $\boldsymbol{\zeta}_s$
$\boldsymbol{\beta}^{(\Omega)}$	Set of regression coefficents at Level 1 relating to Ω^-
$\boldsymbol{\gamma}^{(\Psi)}$	Set of fixed effects at Level 2 relating to Ψ^-
β_q	Regression coefficient q at Level 1
β_{qj}	Random regression coefficient q of group j at Level 1
γ_{qs}	Fixed effect s relating to β_{qj}
$\mathbf{e}$	Residuals at Level 1
$\mathbf{u}$	Residuals at Level 2
σ^2	Residual variance at Level 1
$\mathbf{T}$	Residual variance matrix at Level 2
τ^2_{qj}	Residual variance of the regression of β_{qj} on W_{qj}
ρ	Intraclass correlation coefficient

Symbols of the Measurement Error Model

Item Response Theory Model

a_k	Discrimination parameter relating to item k
b_k	Difficulty parameter relating to item k
c_k	Guessing parameter relating to item k
$\boldsymbol{\xi}_k$	Set of item parameters relating to item k
ω_{ij}	Ability parameter of person i in group j
θ_{ij}	Ability parameter of person i in group j
ζ_j	Characteristic of group j or of a person representing group j
η_{ijk}	Equals $a_k\theta_{ij} - b_k$
$\mathbf{Y}, \mathbf{y}$	Response patterns relating to $\boldsymbol{\omega}$ or $\boldsymbol{\theta}$
$\mathbf{X}, \mathbf{x}$	Response patterns relating to $\boldsymbol{\theta}$
$\mathbf{W}, \mathbf{w}$	Response patterns relating to $\boldsymbol{\zeta}$
ε	Latent residuals at the item level

Classical True Score Model

ω_{ij}	True score of person i in group j
θ_{ij}	True score of person i in group j
ζ_j	True characteristic of group j or of a person representing group j
$\mathbf{Y}, \mathbf{y}$	Observed scores relating to $\boldsymbol{\omega}$ or $\boldsymbol{\theta}$
$\mathbf{X}, \mathbf{x}$	Observed scores relating to $\boldsymbol{\theta}$
$\mathbf{W}, \mathbf{w}$	Observed scores relating to $\boldsymbol{\zeta}$
ε	Error scores
$\varepsilon^{(y)}$	Error scores relating to the observed scores $\mathbf{Y}$
$\varepsilon^{(x)}$	Error scores relating to the observed scores $\mathbf{X}$
$\varepsilon^{(w)}$	Error scores relating to the observed scores $\mathbf{W}$
φ	Group specific error variance

Chapter 1

Introduction

1. Multilevel Data

In a wide variety of research areas, analysts are confronted with hierarchical structured data. Examples of this nested structure of the data include longitudinal data where several observations are nested within individuals, cross-national data where observations are nested in geographical, political or administrative units, data from surveys where respondents are nested under an interviewer, and test data of students within schools (see, for example, Longford, 1993). The nested structure gives rise to multilevel data. The problem is properly analyzing the data taking the hierarchical structure into account.

There are two often criticized approaches for analyzing variables from different levels at one single level. The first is to disaggregate all higher order variables to the individual level. That is, data from higher levels are assigned to a much larger number of units at Level 1. In this approach, all disaggregated values are assumed to be independent of each other, which is a misspecification that threatens the validity of the inferences. In the second approach, the data at the individual level are aggregated to the higher level. As a result, all within group information is lost. This is especially serious because relations between the aggregated variables can be much stronger and different from the relations between non-aggregated variables (see, for instance, Snijders & Bosker, 1999, pp. 14). When the nested structure within multilevel data is ignored, standard errors are estimated with bias.

A class of models that takes the multilevel structure into account and makes it possible to incorporate variables from different aggregation levels is the class of so-called multilevel models. Multilevel models support

analyzing variables from different levels simultaneously, taking account of the various dependencies. These models entail a statistically more realistic specification of dependencies and do not waste information. The importance of a multilevel approach is fully described by Burstein (1980). Different methods and algorithms have been developed for fitting a multilevel model, and these have been implemented in standard software. The EM algorithm (Dempster, Laird, & Rubin, 1978), the iteratively reweighted least squares method of Goldstein (1986), and Fisher scoring algorithm (Longford, 1993) have become available in specialized software for fitting multilevel models (HLM, Raudenbush, Bryk, Cheong, & Congdon, 2000, MLwiN, Goldstein, Rasbash, Plewis, Draper, Brown, Yang, Woodhouse, & Healy, 1998, Mplus, Muthén & Muthén, 1998, and VARCL, Longford, 1990, respectively).

The field of multilevel research is broad and covers a wide range of problems in different areas. In social research, the basic problem is to relate specific attributes of individuals and characteristics of groups and structures in which the individuals function. In sociology, multilevel analysis is a particularly useful strategy for contextual analysis, which focuses on the effects of the social context on individual behavior (see, for example, Mason, Wong, & Entwisle, 1983). In the same way, relating micro and macro levels is an important problem in economics; for an overview, see Baltagi (1995). Moreover, within repeated measurements of a variable on a subject, interest is focused on the relationship of the variable to time (Bryk & Raudenbush, 1987; Goldstein, 1989; Longford, 1993). Further, Bryk and Raudenbush (1987) have introduced multilevel models in meta-analysis. The multilevel model has been used extensively in educational research, see, for example, Bock, (1989), Bryk and Raudenbush (1987), Goldstein (1995) and Hox (1995). Extensive overviews of multilevel models can be found in Hüttner and van den Eeden (1995), Kreft and de Leeuw (1998) and Longford (1993).

2. Measurement Error

In many research areas, such as physical or social sciences, studies may involve variables that cannot be observed directly or are observed subject to error. For example, a person's mathematical ability cannot be measured directly, only the performance on a number of mathematical test items. Also data collected from respondents contain response error. That is, there is response variation in answers to the same question when repeatedly administered to the same person. Measurement error can occur in both independent explanatory and dependent variables. The reliability of explanatory variables is an important methodological question. When the reliability is known, corrections can be made

(Fuller, 1987), or, if repeated measurements are available, the reliability can be incorporated in the model and estimated directly. The use of unreliable explanatory variables leads to biased estimation of regression coefficients and the resulting statistical inference can be very misleading unless careful adjustments are made (Carroll, Rupert, & Stefanski, 1995; Fuller, 1987). To correct for measurement error, data that allow for estimation of the parameters in the measurement error model are collected. Measurement error models have been applied in different research areas to model errors-in-variables problems, incorporating error in the response as well as in the covariates. In epidemiology, covariates, such as blood pressure or level of cholesterol, are frequently measured with error (see, for example, Buonaccorsi, 1991; Müller & Roeder, 1997; Wakefield & Morris, 1999). In educational research, students' pre-test scores, socio-economic status or intelligence are often used as explanatory variables in predicting students' examination results. Further, students' examination results or abilities are measured subject to error or cannot be observed directly. The measurement errors associated with the explanatory variables or variables that cannot be observed directly are often ignored or analyses are carried out using assumptions that may not always be realistic (see, for example, Aitkin & Longford, 1986; Goldstein, 1995).

Although the topic of modeling measurement error has received a considerable amount of attention in the frequentist literature, for the greater part, this attention is focused on linear measurement error models, more specifically, the classical additive measurement error model, e.g. Carroll et al. (1995), Fuller (1987), Goldstein (1995), and Longford (1993). The classical additive measurement error model is based on the assumption of homoscedasticity, which entails equal variance of measurement errors conditional on different levels of the dependent variable. Further, it is often assumed that the measurement error variance can be estimated from replicate measurements or validation data, or that it is a priori known for identification of the model. Often the measurement error models are very complex. For example, certain epidemiology studies involve nonlinear measurement error models to relate observed measurements to their true values (see, for example, Buonaccorsi & Tosteson, 1993; Carroll et al., 1995). In educational testing, item response models relate achievements of the students to their response patterns (see, for instance, Lord, 1980 or van der Linden & Hambleton, 1997).

Measurement error models are often calibrated using external data. To correct for measurement error in structural modeling, the estimates from the measurement error model are imputed in the estimation procedure for the parameters of the structural model. This method has

several drawbacks. In case of a single measurement with a linear regression calibration curve for the association of observed and true scores and a homoscedastic normally distributed error term, the results are exact (Buonaccorsi, 1991). But if a dependent or explanatory variable subject to measurement error in the structural model has a nonconstant conditional variance, the regression calibration approximation suggests a homoscedastic linear model given that the variances are heteroscedastic (Carroll et al., 1995, pp. 63). Also in case of a nonlinear measurement error model and a nonlinear structural model the estimates are biased in certain cases (Buonaccorsi & Tosteson, 1993; Carroll et al., 1995, pp. 62-69).

Until recently, measurement error received relatively little attention in the Bayesian literature (Zellner, 1971, pp. 114-161). Solutions for measurement error problems in a Bayesian analysis were mainly developed after the introduction of Markov chain Monte Carlo sampling (Gelfand & Smith, 1990; Geman & Geman, 1984); see, for example, Bernardinelli, Pascutto, Best, & Gilks (1997), Mallick and Gelfand (1996), Müller and Roeder (1997), Richardson (1996) or Wakefield and Morris (1999). The Bayesian framework provides a natural way of taking into account all sources of uncertainty in the estimation of the parameters. Also, the Bayesian approach is flexible; different sources of information are easily integrated and the computation of the posterior distributions, which usually involves high-dimensional integration, can be carried out straightforwardly by Markov chain Monte Carlo (MCMC) methods.

3. Objectives and Outline

In this thesis a new model is introduced for dealing with measurement error in both the dependent and independent variables of a structural multilevel model. It is shown that the measurement error can be modeled with an item response theory (IRT) model and it is shown that the parameters of the IRT model and the multilevel model can be estimated concurrently. The appropriateness of an IRT model for measurement error will be evaluated by a comparison with the classical true score model.

In Chapter 2, attention will be focused on effects of measurement error on estimating the parameters, where the response error is modeled with an item response model or a classical true score model. Expressions for the posterior estimates of the random regression coefficients will be derived to illustrate the influence of the measurement error on estimating these parameters. An artificial data set is used to show the effects of measurement error on both the dependent and independent variables on estimating all parameters of interest.

In Chapter 3, a description will be given of multilevel IRT modeling, where the dependent variable is measured with error. Further, advantages of modeling response error with an item response model are given in comparison to the use of observed scores. A fully Bayesian estimation procedure is used which resulted in a straightforward and easily to be implemented estimation procedure. Details of the MCMC algorithm are given in the same chapter. A note is given on other approaches for estimating the parameters. A simulated data set is used to illustrate the parameter recovery of the described Gibbs sampler. Further, a Dutch primary school mathematics test is analyzed to illustrate the practical impact of the proposed multilevel IRT model.

Chapter 4 is a logical continuation of Chapter 3. In this chapter independent variables of the structural multilevel model are measured with an error and modeled by an item response model or a classical true score model. Advantages of modeling response error with an item response model, in comparison to the use of the classical true score model, are given. A detailed description of the MCMC estimation procedure is given, both for the case in which the independent variables are correlated and uncorrelated. The quality of the parameter recovery by the Gibbs sampler is shown using a simulated data set. Further, the fit of the structural multilevel model in combination with an item response model or a classical true score model are compared relative to each other. The influences of the group specific error variance is emphasized and illustrated using a real data set from a large scale study concerning a mathematics test (Bosker, Blatchford, & Meijnen, 1999; Hofman & Bosker, 1999).

In Chapter 5, an estimator of the Bayesian latent residuals and their variance is proposed. The Bayesian latent residuals are analyzed to check whether the assumptions in the multilevel IRT model are justified, for example, assumptions as homoscedasticity of variance or normality. Also, statistics to test the assumption of heteroscedasticity at Level 1 of the multilevel IRT model are developed. Outliers among the regression residuals are detected. The posterior distribution of the outliers can be used to compute the probability that an observation is an outlier. Further, the sensitivity of inferences to reasonable changes in the prior distributions is examined. Finally, the fit of several multilevel IRT models is discussed using the data utilized in Chapter 3.

In Chapter 6, another estimation procedure to estimate the parameters of a multilevel IRT model is discussed. It will be shown that a stochastic EM algorithm is an appealing alternative to the Gibbs sampler. The stochastic EM algorithm handles complex missing-data structures in which high-dimensional integration over nuisance parameters may be involved. This feature makes it attractive for estimating a mul-

tilevel IRT model with latent variables defined by a complex structural model. Further, parameter estimates by the stochastic EM algorithm appeared to be close to the maximum likelihood estimates. Both estimation methods, the Gibbs sampler and the stochastic EM algorithm, were compared using a Dutch primary school language test.

Finally, a summary of the main results is given, and some suggestions for further research are made. The chapters in this thesis are self-contained; hence, they can be read separately. Therefore, some overlap could not be avoided.

Chapter 2

The Effects of Measurement Error in a Multilevel Model

1. School Effectiveness Research

Monitoring student outcomes for evaluating teacher and school performance has a long history. A general overview with respect to the methodological aspects and findings in the field of school effectiveness research can be found in Scheerens and Bosker (1997). Methods and statistical modeling issues in school effectiveness studies are given in, for example, Aitkin and Longford (1986) and Goldstein (1997). The applications in this chapter focus on school effectiveness research with fundamental interest in the development of knowledge and skill of individual students in relation to school characteristics. Data are analyzed at the individual level and it is assumed that classrooms, schools, and experimental interventions have an effect on all students exposed to them. In school or teacher effectiveness research, both levels of the multilevel model are of importance because the objects of interest are schools and teachers as well as students. Interest may exist in the effect on student learning of the organizational structure of the school, characteristics of a teacher, and the characteristics of the student.

Multilevel models are used to make inferences about the relationships between explanatory variables and response or outcome variables within and between schools. This type of model simultaneously handles student level relationships and takes account of the way students are grouped in schools. Multilevel models incorporate a unique random effect for each organizational unit. Standard errors are estimated taking into account the variability of the random effects. This variation among the groups in their sets of coefficients can be modeled as multivariate outcomes which may, in turn, be predicted from Level 2 explanatory

variables. The most common multilevel model for analyzing continuous outcomes is a two-level model in which Level 1 regression parameters are assumed to be multivariate normally distributed across Level 2 units. Here, students (Level 1), indexed ij $(i = 1, \ldots, n_j, j = 1, \ldots, J)$, are nested within schools (Level 2), indexed j $(j = 1, \ldots, J)$. In its general form, Level 1 of the two level model consists of a regression model, for each of the J Level 2 groups $(j = 1, \ldots, J)$, in which the outcomes are modeled as a function of Q predictor variables. The outcomes or dependent variables in the regression on Level 1, such as, students' achievement or attendance, are denoted by ω_{ij} $(i = 1, \ldots, n_j, j = 1, \ldots, J)$. The Q explanatory variables at Level 1 contain information on students' characteristics, such as, gender and age, which are measured without error. Level 1 explanatory variables can also be latent variables, such as, socio-economic status, intelligence, community loyalty, or social consciousness. The unobserved Level 1 covariates are denoted by $\boldsymbol{\theta}$, the directly observed covariates by Λ. Level 1 of the model is formulated as

$$\omega_{ij} = \beta_{0j} + \ldots + \beta_{qj}\theta_{qij} + \beta_{(q+1)j}\Lambda_{(q+1)ij} + \ldots + \beta_{Qj}\Lambda_{Qij} + e_{ij}, \quad (2.1)$$

where the first q predictors correspond to unobservable variables and the remaining $Q-q$ predictors correspond to directly observed variables. Random error $\mathbf{e}_j$ is assumed to be normally distributed with mean $\mathbf{0}$ and variance $\sigma_j^2\mathbf{I}_{n_j}$. The regression parameters are treated as outcomes in a Level 2 model, although, the variation in the coefficients of one or more parameters could be constrained to zero. The Level 2 model, containing predictors with measurement error, $\boldsymbol{\zeta}$, and directly observed covariates, Γ, is formulated as

$$\beta_{qj} = \gamma_{q0} + \ldots + \gamma_{qs}\zeta_{sqj} + \gamma_{q(s+1)}\Gamma_{(s+1)qj} + \ldots + \gamma_{qS}\Gamma_{Sqj} + u_{qj}, \quad (2.2)$$

for $q = 0, \ldots, Q$, where the first s predictors correspond to unobservable variables and the remaining $S - s$ correspond to directly observed variables.

The set of variables $\boldsymbol{\theta}$ is never observed directly but supplemented information about $\boldsymbol{\theta}$, denoted as $\mathbf{X}$, is available. In this case, $\mathbf{X}$ is said to be a surrogate, that is, $\mathbf{X}$ is conditionally independent of $\boldsymbol{\omega}$ given the true covariates $\boldsymbol{\theta}$. In the same way, $\mathbf{Y}$ and $\mathbf{W}$ are defined as surrogates for $\boldsymbol{\omega}$ and $\boldsymbol{\zeta}$, respectively. For item responses, the distribution of the surrogate response depends only on the latent variable. All the information in the relationship between $\mathbf{X}$ and the predictors, $\boldsymbol{\theta}$, is explained by the latent variable. This is characteristic of nondifferential measurement error (Carroll et al., 1995, pp. 16-17). Accordingly, parameters

in response models can be estimated given the true dependent and explanatory variables, even when these variables $(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\zeta})$ are latent. The observed variables are also called manifest variables or proxies.

2. Models for Measurement Error

A psychological or educational test is a device for measuring the extent to which a person possesses a certain trait. These traits are, for example, intelligence, arithmetic and linguistic ability. Suppose that a test is administered repeatedly to a subject, that the person's properties do not change over the test period, and that successive measurements are unaffected by previous measurements. The average value of these observations will converge, with probability one, to a constant, called the true score of the subject. In practice, due to the limited number of items in the test and the response variation, the observed test scores deviate from the true score. Let Y_{ijk} denote the test score of a subject ij on item k, with an error of measurement ε_{ijk}. Then $Y_{ijk} - \varepsilon_{ijk}$ is the true measurement or the true score. Further, let y_{ijk} denote the realization of Y_{ijk}. The hypothetical distribution defined over the independent measurements on the same person is called the propensity distribution of the random variable Y_{ijk}. Accordingly, the true score of a person, denoted again as θ_{ij}, is defined as the expected value of the observed score Y_{ijk} with respect to the propensity distribution. The error of measurement ε_{ijk} is the discrepancy between the observed and the true score, formally,

$$Y_{ijk} = \theta_{ij} + \varepsilon_{ijk}. \tag{2.3}$$

A person has a fixed true score and on each occasion a particular observed and error score with probability governed by the propensity distribution. The classical test theory model is based on the concept of the true score and the assumption that error scores on different measurements are uncorrelated. An extensive treatment of the classical test theory model can be found in Lord and Novick (1968). The model is applied in a broad range of research areas where some characteristic is assessed by questionnaires or tests, for example, in the field of epidemiologic studies (see, e.g., Freedman, Carroll, & Wax, 1991; Rosner, Willett, & Spiegelman, 1989).

Another class of models to describe the relationship between an examinee's ability and responses is based on the characteristics of the items of the test. This class is labelled item response models. The dependence of the observed responses to binary scored items on the latent ability is fully specified by the item characteristic function, which is the regression of item score on the latent ability. The item response function is used to make inferences about the latent ability from the observed

item responses. The item characteristic functions cannot be observed directly because the ability parameter, $\boldsymbol{\theta}$, is not observed. But under certain assumptions it is possible to infer the information of interest from the examinee's responses to the test items, see, Lord and Novick (1968) or Lord (1980). One of the forms of the item response function for a dichotomous item is the normal ogive,

$$P\left(Y_{ijk}=1 \mid \theta_{ij}, a_{k,} b_k\right)=\Phi\left(a_k \theta_{ij}-b_k\right), \tag{2.4}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function, b_k is the ability level at the point of inflexion, where the probability of a correct response equals .5 and a_k is proportional to the slope of the curve at the inflexion point. The parameters a_k and b_k are called the discrimination and difficulty parameters, respectively. For extensions of this model to handle the effect of guessing or polytomously scored items, see, e.g., Hambleton and Swaminathan (1985) or van der Linden and Hambleton (1997).

The true score,

$$\sum_{k=1}^{K} P\left(Y_{ijk}=1 \mid \theta_{ij}\right), \tag{2.5}$$

is a monotonic transformation of the latent ability underlying the normal ogive model, formula (2.4). Every person with the same ability has the same expected number-right true score. Furthermore, the probability of a correct score is an increasing function of the ability; thus, the number-right true score is an increasing function of the ability. The true score, formula (2.5), and the latent ability are the same thing expressed on different scales of measurement (Lord & Novick, 1968, pp. 45-46). Since the true score and the latent ability are equivalent, the terms will be used interchangeably. Further, the context of the model under consideration will reveal whether $\boldsymbol{\theta}$ represents a true score or a latent ability.

3. Multilevel IRT

The combination of a multilevel model with one or more latent variables modeled by an item response model is called a multilevel IRT model. The structure of the model is depicted with a path diagram in Figure 2.1. The path diagram gives a representation of a system of simultaneous equations and presents the relationships within the model. It illustrates the combination of the structural model with the measurement error models. The symbols in the path diagram are defined as follows. Variables enclosed in a square box are observed without error and the unobserved or latent variables are enclosed in a circle. The error

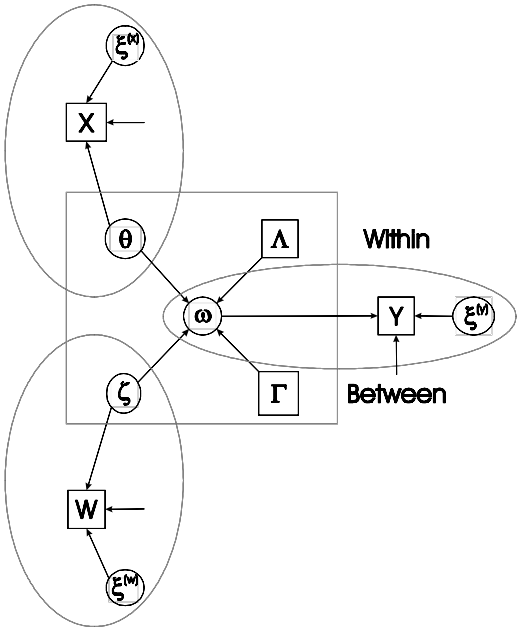

Figure 2.1. A path diagram of a Multilevel IRT model, where item response theory models measure the latent variables within the structural multilevel model.

terms are not enclosed and presented only as arrows on the square boxes. Straight single headed arrows between variables signify the assumption that a variable at the base of the arrow 'causes' variable at the head of the arrow. The square box with a dotted line, around the multilevel parameters, signifies the structural multilevel model. The upper part is denoted as the within-group regression, that is, regression at Level 1, and the lower part is denoted as the regression at Level 2 across groups. Accordingly, the regression at Level 1 contains two types of explanatory variables, observed or manifest and unobserved or latent variables and both are directly related to the unobserved dependent variable. Also Level 2 consists of observed and latent variables.

The model assumes that the latent variables within the structural multilevel model determine the responses to the items. That is, the latent variables $\boldsymbol{\omega}$, $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ determine the observed responses $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{W}$, respectively. The pair of a latent variable and an observed variable enclosed in an ellipse with a dotted line defines a measurement error model. In an item response theory model item parameters, denoted as $\boldsymbol{\xi}$, also determine the responses to the items.

The model in Figure 2.1 is not identified. Identification of the model is possible by fixing the origin and scale of the latent variables. Another way is to impose identifying restrictions on the item parameters of each test. In case of the classical true score model as measurement error model, the measurement error variances ought to be known, or estimated properly, to identify the model. One could, for example, from repeated measurements estimate the error variance.

Handling response error in both the dependent and independent variables in a multilevel model using item response models has several advantages in comparison to the use of the classical true score model; see, Chapter 3 and 4. In item response theory, measurement error can be defined locally, for instance, as the posterior variance of the ability parameter given a response pattern. This results in a more realistic, heteroscedastic treatment of the measurement error. Besides, the fact that in IRT reliability can be defined conditionally on the value of the latent variable offers the possibility of separating the influence of item difficulty and ability level, which supports the use of incomplete test administration designs, optimal test assembly, computer adaptive testing and test equating. Finally, the model is identified in a natural way, without needing any prior knowledge.

3.1 Markov chain Monte Carlo

Analyzing the joint posterior distribution of the parameters of interest in the model in (2.1) and (2.2) is infeasible. Computing expectations of marginal distributions using, for example, Gauss-Hermite quadrature is also quite difficult. We will return to this point in Chapter 3 and 6. Therefore, a sampling-based approach using an MCMC algorithm to obtain random draws from the joint posterior distribution of the parameters of interest given the data is considered. MCMC is a simulation based technique for sampling from high dimensional joint distributions. From a practical perspective, the Markov chains are relatively easy to construct and MCMC techniques are straightforward to implement. Besides, they are typically the only currently available techniques for exploring these high dimensional problems. In particular, the Gibbs sampler (Gelfand & Smith, 1990; Geman & Geman, 1984) is a procedure for sampling from the complete conditional distributions of all estimands. The algorithm is described as follows. Consider a joint distribution π defined on a set $\boldsymbol{\theta} \subset \mathbb{R}^k$ (in this section $\boldsymbol{\theta}$ is the generic parameter of π, not necessarily an ability parameter in an IRT model). The MCMC algorithm consists of specifying a Markov chain with stationary distribution π. The elements of $\boldsymbol{\theta}$ are partitioned into k components $(\theta_1, \ldots, \theta_k)$. Each component of $\boldsymbol{\theta}$ may be a scalar or a vector. One iteration of the Gibbs sampler is

defined as an updating of one component of $\boldsymbol{\theta}$. To obtain a sample from the target distribution π, the Gibbs sampler creates a transition from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$. Updating the first component, θ_1, consists of sampling from the full conditional distribution

$$\pi\left(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}\right)$$

which is the distribution of the first component of $\boldsymbol{\theta}$ conditional on all other components. Subsequently, $\theta_2^{(t+1)}$ is obtained as a draw from

$$\pi\left(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}\right),$$

and so on, until $\theta_k^{(t+1)}$ is drawn from

$$\pi\left(\theta_k \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)}\right),$$

which completes updating the components to $\boldsymbol{\theta}^{(t+1)}$.

The order of updating the different components is usually fixed, although this is not necessary. Random permutations of the updating order are acceptable. The choice of updating scheme can effect the convergence of the sampler (Roberts & Sahu, 1997), that is, a different updating strategy can make the algorithm convergence faster. In some applications a multivariate component sampler, instead of a single component sampler, is a more natural choice. This so-called blocking of the Gibbs sampler by blocking highly correlated components into a higher-dimensional component can improve the convergence of the Markov chain (Gelman, Carlin, Stern, & Rubin, 1995; Roberts & Sahu, 1997). On the other hand, updating in a block or group is often computationally more demanding than the corresponding componentwise updating scheme.

Running multiple chains reduces the variance of the parameter estimates attributable to the Gibbs sampler. This is useful in obtaining independent samples, but these are not required for estimating the parameters of interest. A very long run gives the best chance of finding new modes. However, inference from a Markov chain simulation is always problematic because there are areas of the target distribution that have not been covered by the finite chain. In practice, both methods are desirable, to check the behavior and convergence of the Markov chain. There are several methods for monitoring convergence, but despite much recent work, convergence diagnostics for the Gibbs sampler remains a topic for further research. The source of the problem is that the simulation converges to a target distribution rather than a target point. Different methods can be found in, for example, Brooks & Gelman (1998),

Cowles & Carlin (1996) and Gelman (1995). In the present chapter, convergence diagnostics and multiple chains from different starting points were used to verify that the Markov chain had converged. In addition, a visual inspection of the plot of random deviates against iteration was made to decide whether the Markov chain had converged.

A detailed description of the implementation of the Gibbs sampler to estimate the model in Figure 2.1 will not be given here. The full conditional distributions of the parameters of interest can be found in Chapter 3 and 4. Here, the Gibbs sampler is used to estimate parameters of the model to illustrate the effects of response error in both the dependent and independent variables of the structural multilevel model.

4. Ignorable and Non-Ignorable Measurement Error

This section focuses on problems associated with measurement error in the dependent and independent variables of a structural multilevel model. In certain cases, measurement error does not play a role. That is, the model for the latent variable also holds for the manifest variable with parameters unchanged, except that a measurement error variance component is added to the variance of the residuals (Carroll et al., 1995, pp. 229). An example is a structural linear regression model with measurement error in the dependent variable, where the measurement error is confounded with the residuals, resulting in greater variability of the parameter estimates. The measurement error is called ignorable in these cases. If the estimates of the regression coefficients are biased because measurement error in the manifest variable is ignored, then the measurement error is called non-ignorable. For example, in a linear regression model with measurement error in a covariate, the least squares regression coefficient is biased toward zero, that is, the regression coefficient is attenuated by the measurement error (Fuller, 1987, pp. 3).

Here it will be shown that response error in the dependent, independent, or both variables in a multilevel model is not ignorable. That is, the parameter estimates of the multilevel model are affected by the presence of the response error in the manifest variables. It will be shown that disattenuated parameter estimates of the structural multilevel model are obtained by modeling the response error in the manifest variables with a classical true score model. The generalization of the results from a multilevel true score model to a multilevel IRT model will be discussed at the end of this section.

Consider the linear regression model with the independent variable measured with error,

$$\omega_{ij} = \beta_0 + \beta_1 \theta_{ij} + e_{ij}, \tag{2.6}$$

where the equation errors are independent and normally distributed with mean zero and variance σ^2. It is assumed that the distribution of true scores, θ_{ij}, in the population is standard normal, that is, the θ_{ij} are unobservable independent realizations of a standard normal random variable. For a given person, the true score is a constant, but the observed score and error term are random variables, see formula (2.3).

In the classical true score model, inferences about θ_{ij} are made from the responses x_{ijk} for $k = 1, \ldots, K$, which are related to θ_{ij} via

$$X_{ij} = \theta_{ij} + \varepsilon_{ij}^{(x)}, \tag{2.7}$$

where x_{ij} is a realization of X_{ij}, the observed total score of person ij, and $\varepsilon_{ij}^{(x)}$ an error term that is independent of θ_{ij} and e_{ij}. The superscript x denotes the connection with the observed variable X_{ij}. Further, it is assumed that $\varepsilon_{ij}^{(x)}$ are independent normally distributed with mean zero and variance φ_x, where, again, the subscript x denotes the connection with the observed variable X_{ij}. One of the consequences of the measurement error in the independent variable can be seen in the posterior expectation of the regression coefficient β_1 given the variables ω_{ij}, x_{ij} and the parameters β_0, σ^2 and φ_x. This posterior expectation is derived from the conditional distribution of θ_{ij} given x_{ij} and φ_x,

$$f(\theta_{ij} \mid x_{ij}, \varphi_x) \propto f(x_{ij} \mid \theta_{ij}, \varphi_x)\, f(\theta_{ij}; 0, 1), \tag{2.8}$$

where the right-hand-side consists of a product of normal densities. Due to standard properties of normal distributions (e.g., see, Box & Tiao, 1973; Lindley & Smith, 1972) the full conditional posterior density of θ_{ij} given x_{ij} and φ_x is also normally distributed and is given by

$$\theta_{ij} \mid X_{ij}, \varphi_x \sim N\left(\frac{\varphi_x^{-1}}{1+\varphi_x^{-1}} x_{ij}, \frac{1}{1+\varphi_x^{-1}}\right). \tag{2.9}$$

Below, $\varphi_x^{-1}/\left(1+\varphi_x^{-1}\right)$ will be denoted by λ_x. The regression on Level 1 imposes a density $f\left(\omega_{ij} \mid \boldsymbol{\beta}, \theta_{ij}, \sigma^2\right)$ that can be considered as a likelihood, and formula (2.9) can be regarded as the prior for the unobserved θ_{ij}. Accordingly, it follows that the conditional posterior distribution of ω_{ij} is given by

$$f\left(\omega_{ij} \mid \boldsymbol{\beta}, \theta_{ij}, \sigma^2, x_{ij}, \varphi_x\right) \propto f\left(\omega_{ij} \mid \boldsymbol{\beta}, \theta_{ij}, \sigma^2\right) f(\theta_{ij} \mid x_{ij}, \varphi_x).$$

Due to properties of normal distributions (Lindley & Smith, 1972), the conditional distribution of ω_{ij} is also normally distributed, that is,

$$\omega_{ij} \mid \boldsymbol{\beta}, \sigma^2, X_{ij}, \varphi_x \sim N\left(\beta_0 + \lambda_x \beta_1 x_{ij}, \sigma^2 + \beta_1^2 (1-\lambda_x)\right). \tag{2.10}$$

In the same way it follows that, given a uniform prior for β_1, the conditional posterior of β_1 given $\boldsymbol{\omega}, \beta_0, \sigma^2, \mathbf{x}$ and φ_x is normal with expectation

$$E\left[\beta_1 \mid \boldsymbol{\omega}, \mathbf{x}, \beta_0, \sigma^2, \varphi_x\right] = \lambda_x^{-1}\widehat{\beta}_1, \tag{2.11}$$

where $\widehat{\beta}_1$ is the least squares estimator in the regression of $\boldsymbol{\omega} - \beta_0$ on $\mathbf{x}$. Because of the measurement error in the explanatory variable, the least squares regression coefficient is biased toward zero, that is, the regression coefficient is attenuated by the measurement error. The ratio λ_x defines the degree of attenuation, which is a measure of the degree of true score variation relative to observed score variation. In the social science literature, this ratio is called the reliability of X_{ij}. From (2.11) it can be seen that if the ratio λ_x is known, it is possible to construct an unbiased estimator of β_1. Several techniques for estimating this model, given λ_x, can be found in Fuller (1987). The effect of errors in variables on ordinary least squares estimators is well known, and is described in, for example, Cochran (1968) and Fuller (1987).

Next, suppose the intercept and slope of model (2.6) are random coefficients, that is, the coefficients vary over Level 2 groups. The coefficients are treated as outcomes in a Level 2 model given by

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j}, \end{aligned} \tag{2.12}$$

where the Level 2 error terms $\mathbf{u}_j$ have a multivariate normal distribution with mean zero and covariance matrix $\mathbf{T}$. In the sequel, it will be assumed that the errors on Level 2 are uncorrelated. That is, the covariance matrix $\mathbf{T}$ consists of diagonal elements $\text{var}\,(u_{0j}) = \tau_0^2$ and $\text{var}\,(u_{1j}) = \tau_1^2$. Suppose that the dependent variable ω_{ij} is not observed exactly, but its error-prone version Y_{ij} is available. So

$$Y_{ij} = \omega_{ij} + \varepsilon_{ij}^{(y)}, \tag{2.13}$$

where the measurement errors $\varepsilon_{ij}^{(y)}$ are independent of ω_{ij} and e_{ij}, and independent normally distributed with mean zero and variance φ_y. The superscript and subscript y emphasize the connection with the observed total score Y_{ij}. Again, the conditional posterior distribution of $\mathbf{Y}_j$, the observed scores of students in group j, given $\boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2$ and φ_y follows from the standard properties of normal distributions, that is,

$$f\left(\mathbf{y}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2, \varphi_y\right) \propto f\left(\mathbf{y}_j \mid \boldsymbol{\omega}_j, \varphi_y\right) f\left(\boldsymbol{\omega}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2\right),$$

where the second factor on the right-hand side defines the distribution of the true scores $\boldsymbol{\omega}_j$ in the population. As a result,

$$\mathbf{Y}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2, \varphi_y \sim N\left(\beta_{0j} + \beta_{1j}\boldsymbol{\theta}_j, \left(\varphi_y + \sigma^2\right)\mathbf{I}_{n_j}\right), \tag{2.14}$$

where $\mathbf{I}_{n_j}$ is the identity matrix of dimension n_j. Obviously, the measurement error in the dependent variable results in an extra variance component φ_y. Combining this conditional distribution of $\mathbf{Y}_j$ with the prior knowledge about $\boldsymbol{\beta}_j$, in formula (2.12), results in the conditional posterior distribution of $\boldsymbol{\beta}_j$ given $\mathbf{y}_j, \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \zeta_j$ and φ_y. Define $\Sigma_j = \left(\sigma^2 + \varphi_y\right)\left(\mathbf{H}_j^t \mathbf{H}_j\right)^{-1}$, where $\mathbf{H}_j = \left[\mathbf{1}_{n_j}, \boldsymbol{\theta}_j\right]$. Then

$$\boldsymbol{\beta}_j \mid \mathbf{Y}_j, \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \zeta_j, \varphi_y \sim N\left(\frac{\Sigma_j^{-1}\widehat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1}\mathbf{A}\boldsymbol{\gamma}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}, \frac{\mathbf{1}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}\right), \tag{2.15}$$

where $\mathbf{A}$ defines the structure of the explanatory variables on Level 2. The posterior expectation of $\boldsymbol{\beta}_j$ is the well-known composite or shrinkage estimator, where the amount of weight placed on the estimates depends on their precision. Notice that the usual least squares estimator, $\widehat{\boldsymbol{\beta}}_j$, based on the linear regression on Level 1 given $\boldsymbol{\theta}_j$ and $\mathbf{Y}_j$, is weighted by Σ_j^{-1}, which contains the measurement error in the dependent variable. Thus, the estimator of $\boldsymbol{\beta}_j$ is not equivalent to the standard least squares estimator of $\boldsymbol{\beta}$, and as consequence, the measurement error in the dependent variable of a structural multilevel model is not ignorable. The estimates of the random regression coefficients are attenuated when the measurement error in the dependent variable is ignored because the least squares estimator $\widehat{\boldsymbol{\beta}}_j$ is attenuated by the measurement error.

Next, it will be shown that the posterior expectation of $\boldsymbol{\beta}_j$ given the manifest variables is affected by measurement error in the explanatory variable on Level 1. From formula (2.10) and (2.14) the conditional distribution of $\mathbf{Y}_j$ can be derived as

$$\mathbf{Y}_j \mid \mathbf{X}_j, \boldsymbol{\beta}_j, \sigma^2, \varphi_y, \varphi_x \sim N\left(\beta_{0j} + \lambda_x \beta_{1j}\mathbf{x}_j, \left(\varphi_y + \sigma^2 + \beta_{1j}^2\left(1 - \lambda_x\right)\right)\mathbf{I}_{n_j}\right) \tag{2.16}$$

The conditional posterior distribution of $\boldsymbol{\beta}_j$ can be derived by considering this conditional distribution of $\mathbf{Y}_j$ as the likelihood and formula (2.12) as the prior for its parameter vector $\boldsymbol{\beta}_j$. To obtain an analytical expression for this conditional posterior distribution, it must be assumed that the variance in (2.16) is known. Denote this variance, for group j, as $\mathbf{C}_j$. In practice, an empirical Bayes estimator could be used. Define $\Sigma_j = \mathbf{C}_j\left(\mathbf{H}_j^t \mathbf{H}_j\right)^{-1}$, where $\mathbf{H}_j = [1, \lambda_x \mathbf{x}_j]$. Then it follows that

$$\boldsymbol{\beta}_j \mid \mathbf{Y}_j, \mathbf{X}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \zeta_j, \varphi_y, \varphi_x \sim N\left(\frac{\Sigma_j^{-1}\widehat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1}\mathbf{A}\boldsymbol{\gamma}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}, \frac{\mathbf{1}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}\right), \tag{2.17}$$

where the other variables are defined as in formula (2.15) . The posterior expectation is a shrinkage estimator where $\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{H}_j^t \mathbf{H}_j\right)^{-1} \mathbf{H}_j^t \mathbf{y}_j$ and the variance of $\widehat{\boldsymbol{\beta}}_j$ increases due to the measurement error in the dependent and independent variables. Besides the measurement error in the dependent variable, the reliability ratio λ_x further influences the least squares regression coefficients $\widehat{\boldsymbol{\beta}}_j$.

Finally, assume that the explanatory variable on Level 2, $\boldsymbol{\zeta}$, is unobserved and instead a variable $\mathbf{W}$ is observed with measurement error variance φ_w, that is,

$$W_j = \zeta_j + \varepsilon_j^{(w)},$$

where the measurement errors $\varepsilon_j^{(w)}$ are independent of ζ_j and u_{0j}, and independently normally distributed with mean zero and variance φ_w. Further, it is assumed that the true scores, ζ_j, in the population are standard normally distributed. Analogous to the derivation of (2.10) , it follows that

$$\beta_{0j} \mid W_j, \boldsymbol{\gamma}, \tau_0^2, \varphi_w \sim N\left(\gamma_{00} + \lambda_w \gamma_{01} w_j, \tau_0^2 + \gamma_{01}^2 \left(1 - \lambda_w\right)\right), \qquad (2.18)$$

where $\lambda_w = \varphi_w^{-1} / \left(1 + \varphi_w^{-1}\right)$. Again, the posterior expectation of $\boldsymbol{\beta}_j$ can be derived by combining the prior information for β_{0j} and the standard prior information for β_{1j}, from (2.12) , with the likelihood in formula (2.16). Hence the conditional posterior distribution of $\boldsymbol{\beta}_j$ is equivalent to formula (2.17) , except that the first diagonal-element of $\mathbf{T}$ is increased by $\gamma_{01}^2 \left(1 - \lambda_w\right)$, and the first row of $\mathbf{A} = (1, \lambda_w W_j, 0)$. Accordingly, the shrinkage estimator is a combination of two weighted estimators, where both parts are influenced by measurement error in the dependent and independent variables. As a consequence, the measurement error is not ignorable and ignoring it leads to attenuated estimates of the random regression coefficients.

Besides the effect of measurement error on the estimates of random regression coefficients, a perhaps less well-recognized effect is the increased variance of the observed dependent variable given the observed explanatory variables. Without measurement error in the explanatory variables the residual variance of Y_{ij} is

$$\operatorname{var}\left(Y_{ij} \mid \theta_{ij}, \zeta_j\right) = \tau_0^2 + \tau_1^2 \theta_{ij}^2 + \sigma^2 + \varphi_y.$$

By taking into account the measurement error in the independent variables, the residual variance of the manifest variable, Y_{ij}, increases to

$$\operatorname{var}\left(Y_{ij} \mid x_{ij}, w_j\right) = C_{ij} + \mathbf{H}_{ij} \mathbf{T}^{-1} \mathbf{H}_{ij}^t,$$

where $C_{ij} = \left(\varphi_y + \sigma^2 + \beta_{1j}^2 (1 - \lambda_x)\right)$, $\mathbf{H}_{ij} = [1, \lambda_x x_{ij}]$ and $\mathbf{T}$ is the diagonal matrix with elements $\left(\tau_0^2 + \gamma_{01}^2 (1 - \lambda_w), \tau_1^2\right)$. Notice that the response variance in the variance component of the dependent variable is just an extra variance component, but the measurement error variance in the explanatory variables causes a complex variance structure. The structure gets even more complex if the variables or error terms are correlated (Schaalje & Butts, 1993).

This overview of non-ignorable measurement error is based on the classical true score model. The conditional distributions of the random regression coefficients are derived by using the standard properties of the normal distribution. If the response error is modeled by an item response model, the conditional distributions of these parameters can be found in the same way by introducing an augmented variable $\mathbf{Z}$. Interpret the observation Z_{ijk} as an indicator that a continuous variable with normal density is negative or positive. Denote this continuous variable as $Z_{ijk}^{(x)}$, where the superscript x denotes the connection with the observed response variable X_{ijk}. It is assumed that $X_{ijk} = 1$ if $Z_{ijk}^{(x)} > 0$ and $X_{ijk} = 0$ otherwise. It follows that the conditional distribution $Z_{ijk}^{(x)}$ given θ_{ij} and $\xi_k^{(x)}$ is normal. This distribution can be used to obtain the conditional distributions of the random regression parameters in the same way as above. Expanding the two parameter normal ogive model to a three parameter normal ogive model to correct for guessing can be done by introducing an extra augmented variable (Johnson & Albert, 1999, pp. 204-205). Further, observed ordinal data can be modeled by assuming that a latent variable underlies the ordinal response (Johnson & Albert, 1999, pp. 127-133).

5. An Illustrative Example

In this section, the effects of measurement error in dependent and explanatory variables at different levels in a structural multilevel model are demonstrated using a simulation study. Further, a numerical example is analyzed to compare the effects of modeling measurement error in dependent and independent variables with an item response model and a classical true score model. The model is given by

$$\begin{aligned} \omega_{ij} &= \beta_{0j} + \beta_{1j}\theta_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j}, \end{aligned} \tag{2.19}$$

where $e_{ij} \sim N\left(0, \sigma^2\right)$ and $\mathbf{u}_j \sim N\left(0, \mathbf{T}\right)$. Furthermore, it is assumed that the surrogates $\mathbf{Y}, \mathbf{X}$ and $\mathbf{W}$ are related to the latent predictors $\boldsymbol{\omega}, \boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, respectively, through a two-parameter normal ogive model.

For the simulation studies, both of the latent predictors, $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, were drawn from the standard normal distribution. The latent dependent variable $\boldsymbol{\omega}$ was generated according to the above model. Response patterns were generated according to a normal ogive model for tests of 40 items. For tests related to the dependent and independent variables at Level 1, $6,000$ response patterns were simulated. The total number of groups was $J = 200$, each group or class consisting of 20 to 40 individuals. For the test related to the latent covariate $\boldsymbol{\zeta}$ at Level 2, 200 response patterns were generated. The generated values of the fixed and random effects, $\boldsymbol{\gamma}, \sigma^2$ and $\mathbf{T}$, are shown under the label Generated in Table 2.1.

5.1 Explanatory Variables Without Measurement Error

In the first simulation study, no response error in the explanatory variables on Level 1 and Level 2 was present, that is, the latent predictors $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ were observed directly without an error. The dependent variable was unobserved but information about $\boldsymbol{\omega}$, denoted as $\mathbf{Y}$, is available. The data were simulated by the multilevel IRT model. The structural multilevel model with measurement error in the dependent variable was estimated with the Gibbs sampler, using the normal ogive model and the classical true score model as measurement error models. Noninformative priors were used for the fixed effects and variance components in the multilevel model; see, Chapter 3 and 4. Also, the methods for computing starting values can be found there. After a burn-in period of $1,000$ iterations, $20,000$ iterations were made to estimate the parameters of the structural model with the two-parameter normal ogive model. For the classical true score model, 500 iterations were necessary as a burn-in period and $5,000$ iterations were used to estimate the parameters. Convergence of the Gibbs sampler was investigated by running multiple chains from different starting points to verify that they yielded similar answers and by plotting the MCMC iterations to verify convergence. For a comprehensive discussion of convergence of the Gibbs sampler, see Cowles and Carlin (1996).

In Table 2.1, the expected a posteriori estimates of the parameters of the multilevel IRT model obtained from the Gibbs sampler are given under the label IRT Model, denoted as Model M_1. Parameter estimates of the structural multilevel model using the classical true score model are given under the label Classical True Score Model, denoted as Model M_{c1}. The multilevel IRT model M_1 was identified by fixing a discrim-

Table 2.1. Parameter estimates of the multilevel model with measurement error in the dependent variable.

	Generated	IRT Model M_1			Classical True Score Model M_{c1}		
Fixed Effects	Coeff.	Coeff.	s.d.	HPD	Coeff.	s.d.	HPD
γ_{00}	.000	−.032	.042	[−.101, .039]	−.056	.032	[−.107, −.007]
γ_{01}	.100	.082	.026	[.040, .124]	.078	.026	[.038, .121]
γ_{10}	.100	.055	.034	[−.002, .109]	.054	.028	[.012, .103]
Random Effects	Var. Comp.	Var. Comp.	s.d.	HPD	Var. Comp.	s.d.	HPD
τ_0^2	.200	.234	.028	[.186, .287]	.200	.022	[.165, .236]
τ_1^2	.200	.201	.023	[.159, .247]	.138	.016	[.115, .167]
τ_{01}^2	.100	.169	.025	[.131, .211]	.118	.015	[.094, .143]
σ^2	.500	.513	.028	[.460, .573]	.435	.010	[.418, .450]

ination and a difficulty parameter to ensure that the latent dependent variable was scaled the same way as in the data generation phase. The structural model with the classical true score model as measurement error model was identified by specifying the parameters of the measurement error distribution. Therefore, the group specific error variance was a priori estimated. The unbiased estimates of the error variances of individual examinees were averaged to obtain the group specific error variance (Lord & Novick, 1968). The group specific error variance relating to the unweighted sums of item responses or test scores Y_{ij}, φ_y was .118, for every individual ij. The observed sum scores were scaled in the same way as the true latent dependent variable ω_{ij}. The reported standard deviations are the posterior standard deviations. The 90% highest posterior probability (HPD) intervals for parameters of interest were computed from a sample from their marginal posterior distribution using the Gibbs sampler (see, Chen & Shao, 1999).

The true parameter values were well within the HPD intervals obtained from the multilevel IRT model, M_1, except for the covariance of the Level 2 residuals, τ_{01}^2, which was too high. Further, the fixed effect, γ_{10}, was not significantly different from zero. The parameter estimates of the random and fixed effects are given under the label Classical True Score Model, Model M_{c1}. Here, more parameter estimates differed from

the true parameter values. Specifically, the variance at Level 1 and the variance of the residuals of the random slope parameter were too low. As a result, the estimates of the slope parameters corresponding to the different groups were more alike in comparison to the corresponding estimates resulting from the multilevel IRT model and the true simulated values. In the fit of Model M_{c1}, the slope parameter estimates vary less across groups. The estimates of the variance components affected the estimate of the intraclass correlation coefficient. This is the proportion of the total residual variation that is due to variation in the random intercepts, after controlling for the Level 1 predictor variable. The simulating values implied an intraclass correlation coefficient of $\rho = .286$, the multilevel IRT estimate was $\rho = .313$, and the Model M_{c1} estimate $\rho = .314$. These estimates were based on iterates of the variance components and were not based on the posterior means of the variance components.

5.2 Explanatory Variables With Measurement Error

In the second simulation study, both the dependent and independent observed variables had measurement error. Table 2.2 presents the results of estimating the parameters of the multilevel model using observed scores, denoted as Model M_o, using a normal ogive model as measurement error model, denoted as Model M_2, and using the classical true score model as measurement error model, denoted as Model M_{c2}, both for the dependent and independent variables. In the estimation procedure, all uncertainties were taken into account, where the group specific error variances for the sum scores relating to the Level 1 and Level 2 predictors, φ_x and φ_w, were .103 and .109, respectively. The multilevel IRT model, where measurement error in the covariates was modeled by a normal ogive model, Model M_2, was identified by fixing a discrimination and a difficulty parameter of each test. Model M_{c2} was identified by specifying the response variance of the observed scores. The true parameters were the same as in Table 2.1. The true parameter values were well within the HPD regions of the multilevel IRT estimates, Model M_2. That is, the parameter estimates were almost the same as the parameter estimates resulting from Model M_1, where the true parameter values were used for the predictor variables instead of modeling the variables with an IRT model. The same applied to the parameter estimates of Model M_{c2} which were comparable to the estimates of Model M_{c1}. Subsequently, the deficiencies of the fit of model M_{c1} also applied to the fit of Model M_{c2}. The posterior variances of the estimates of Model M_2 and M_{c2} were slightly higher in comparison to Model M_1 and M_{c1} because the measurement errors in the predictor variables were taken into account, but the differences were rather small. The estimates given

Table 2.2. Parameter estimates of the multilevel model with measurement error in both the dependent and independent variables.

	Observed M_o		IRT Model M_2			Classical True Score Model M_{c2}		
Fixed Effects	Coeff.	s.d.	Coeff.	s.d.	HPD	Coeff.	s.d.	HPD
γ_{00}	$-.057$	.032	$-.048$	.045	$[-.120, .027]$	$-.058$	.034	$[-.112, .000]$
γ_{01}	.058	.026	.081	.030	$[.032, .130]$	.058	.026	$[.018, .103]$
γ_{10}	.050	.026	.055	.034	$[.000, .110]$	.049	.026	$[.005, .091]$
Random Effects	Var. Comp.	s.d.	Var. Comp.	s.d.	HPD	Var. Comp.	s.d.	HPD
τ_0^2	.201	.023	.233	.030	$[.184, .278]$	.200	.023	$[.165, .238]$
τ_1^2	.126	.015	.200	.027	$[.157, .241]$	.138	.014	$[.098, .144]$
τ_{01}^2	.110	.015	.167	.024	$[.128, .204]$	.118	.015	$[.083, .131]$
σ^2	.560	.010	.515	.035	$[.454, .562]$	.435	.010	$[.416, .450]$

under the label Observed resulted from estimating the multilevel model using observed scores for both the dependent and independent variables, ignoring measurement error in all variables. It was verified that taking account of measurement error in the observed variables resulted in different parameter estimates, especially for the variance components.

Table 2.1 and 2.2 show that the estimates of the variance components were attenuated because the measurement error was ignored. As seen in the preceding section, the estimates of the random intercept and random slope parameters were strongly influenced by the variance components. The effects of measurement error in the dependent and independent variables were also reflected in the estimates of the random regression parameters. Figure 2.2 shows the expected a posteriori estimates of the dependent values in an arbitrary group using Model M_1 and M_{c1}. There was no horizontal shift in the estimates because both models used the true independent variables. The estimates of both models were quite close to the true values, but the more extreme values were better estimated by Model M_1, where the normal ogive model was the measurement model. The regression predicted by Model M_1 resulted in a higher intercept, the slope parameter nearly equaled the true slope parameter. The regression lines were based on posterior means of the random regression coefficients. The predicted regression slope, using Model M_{c1}, was

of opposite sign and resulted in different conclusions. In the same group as in Figure 2.2, the expected a posteriori estimates of the dependent values based on dependent and independent variables measured with an error, using the classical true score model and the normal ogive model, are given in Figure 2.3. The horizontal shifts in the expected a posteriori estimates, in relation to the estimates in Figure 2.2, were caused by the measurement error in the independent variables. The estimates were shrunk towards the mean of both variables. The estimates following from Model M_2 were closer to the more extreme true values. As a result, the predicted regression according to Model M_2 had a wider range, and was closer to the true regression. As in Figure 2.2, the slope estimate of the predicted regression of Model M_{c2} was positive, even though the true parameter slope was negative. In this group, the predicted regression based on observed scores, Model M_o, followed the regression of Model M_{c2}, and seemed to follow the true regression better. Notice that the predictions are slightly better in Figure 2.3, where the explanatory variables were modeled with the classical true score model or the normal ogive model. It seemed that the more complex model, which takes measurement error in all variables into account, was more flexible, resulting in a better fit of the model. Both figures indicate that the normal ogive model for the measurement error model yielded better estimates of the outcomes, especially, in case of the more extreme values. Further, the estimates of the random regression coefficients depended on the values of the variance components and were sensitive to measurement error in the variables. As shown in Figure 2.2 and 2.3, measurement error in the dependent and independent variables may lead to incorrect conclusions.

6. Discussion

Errors in the dependent or independent variables of a multilevel model are modeled by an item response model or a classical true score model. The Gibbs sampler can be used to estimate all parameters. Other estimation procedures, such as error calibration methods (Carroll et. al., 1995), do not take all parameter variability into account.

A fully Bayesian approach accommodates both covariate and response measurement error, and provides more reliable estimates of the variability of the model parameters. On the other hand, the Bayesian approach is computer intensive and still unrecognized in many working environments. Besides, the lack of programs for handling measurement errors in major statistical computer packages further impedes the use of structural multilevel models.

In this study, the consequences of ignoring measurement error are examined to evaluate estimation methods that are able to handle mea-

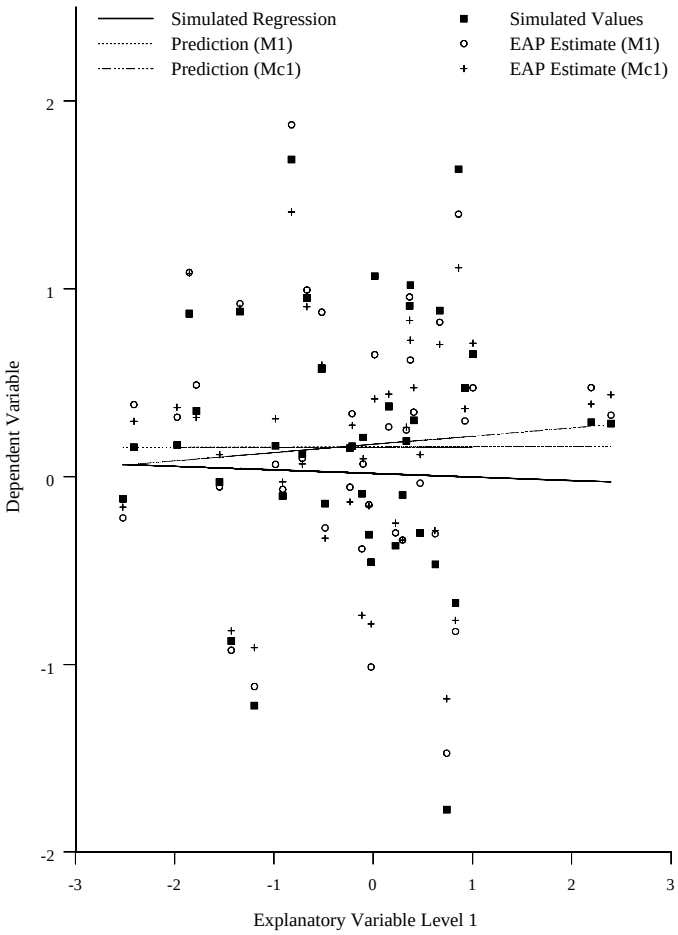

Figure 2.2. Expected posterior estimates and predictions of the dependent values given the true independent variables.

surement error in both the explanatory and independent variables of a structural multilevel model. It was shown that the estimates of the variance components and random regression coefficients are sensitive to measurement error in both the dependent and explanatory variables. Simulation studies were used to exemplify the impact of the measurement error. Other forms of measurement error can be handled similarly,

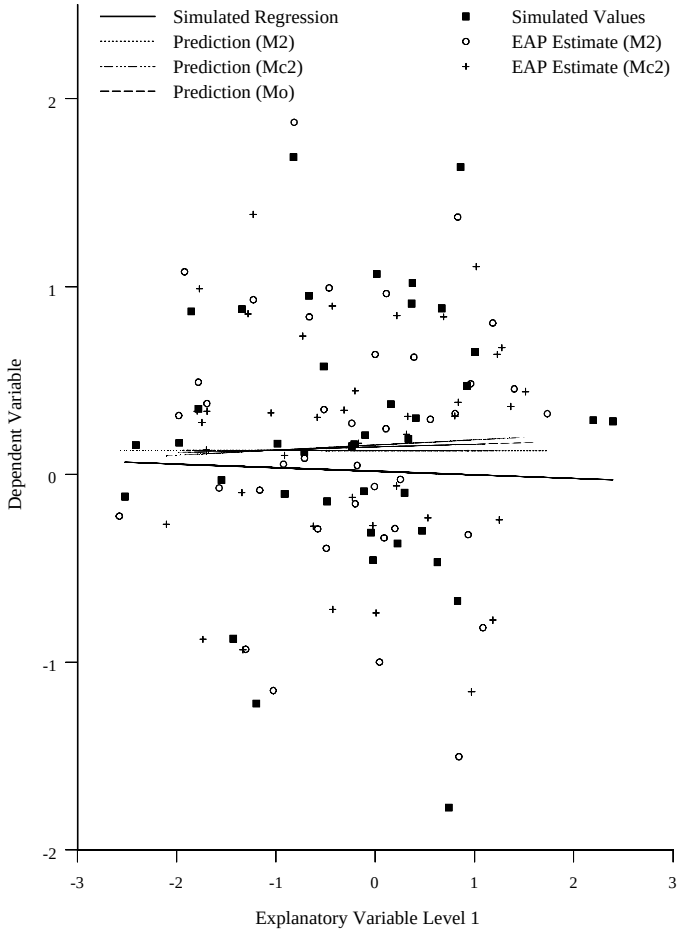

Figure 2.3. Expected posterior estimates and predictions of the dependent values given that the explanatory variables at Level 1 and Level 2 are measured with an error.

but information concerning the probability structure is necessary. Notice that the classical true score model as measurement error model requires a priori estimates of the group specific error variances. These estimates strongly affect the parameter estimates (see, Chapter 4). That is, a small change in the a priori estimates could lead to different conclusions. A detailed description of the Bayesian estimation procedure can be found

in the following chapters. The procedure is flexible in the sense that other measurement error models, and other priors can be used. This supports a more realistic way of modeling measurement error. Also, the estimation procedure can handle multilevel models with three or more levels. It takes the full error structure into account and allows for errors in both the dependent and independent variables.

Chapter 3

Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling

Abstract In this chapter, a two-level regression model is imposed on the ability parameters in an item response theory (IRT) model. The advantage of using latent rather than observed scores as dependent variables of a multilevel model is that it offers the possibility of separating the influence of item difficulty and ability level and modeling response variation and measurement error. Another advantage is that, contrary to observed scores, latent scores are test-independent, which offers the possibility of using results from different tests in one analysis where the parameters of the IRT model and the multilevel model can be concurrently estimated. The two-parameter normal ogive model is used for the IRT measurement model. It will be shown that the parameters of the two-parameter normal ogive model and the multilevel model can be estimated in a Bayesian framework using Gibbs sampling. Examples using simulated and real data are given.

Keywords: Bayes estimates, Gibbs sampler, item response theory (IRT), Markov chain Monte Carlo, multilevel model, normal ogive model.

1. Introduction

In educational and social research, there is a growing interest in the problems associated with describing the relations between variables of different aggregation level. In school effectiveness research, one may, for instance, be interested in the effects of the school budget on the educational achievement of the students. However, the former variable is defined on the school level while the latter variable is defined on the level of students. This gives rise to problems of properly model-

ing dependencies between these variables. These problems can be coped with using multilevel models (Bryk & Raudenbush, 1992; de Leeuw & Kreft, 1986; Goldstein, 1995; Longford, 1993; Raudenbush, 1988). In the above example, students are nested in schools, and in a multilevel model the students would make up a first level and the schools a secondary level. Although most applications of the multilevel paradigm are found in regression and analysis of variance models (see, for instance, Bryk & Raudenbush), multilevel modeling does, in principle, apply to all statistical modeling of data where elementary units are nested within aggregates. Longford, for instance, gives examples of multilevel factor analytical models and generalized linear models.

Also in the field of IRT models some applications of the multilevel paradigm can be found. Adams, Wilson and Wu (1997) discuss the treatment of latent proficiency variables as outcomes in a regression analysis. They show that a regression model on latent proficiency variables can be viewed as a two-level model where the first level consists of the item response measurement model which serves as a within-student model and the second level consists of a model on the student population distribution, which serves as a between-students model. Further, Adams et al. show that this approach results in an appropriate treatment of measurement error in the dependent variable of the regression model. Another application of multilevel modeling in the framework of IRT models was given by Mislevy and Bock (1989) where group-level and student-level effects are combined in an hierarchical IRT model. Both applications can be viewed as special cases of the general approach presented here. This general approach entails a multilevel regression model on the latent proficiency variables allowing for predictors on the student-level and group-level. The motivation for this approach is twofold. Firstly, linear multilevel models are based on the assumption of homoscedasticity, that is, it is assumed that the error component is independent of the outcome variable (i.e., the score of the test taker). In IRT, measurement error can be defined locally, for instance, as the posterior variance of the ability parameter given a response pattern. This local definition of measurement error results in heteroscedasticity: In the Rasch model, for instance, the posterior variance of the ability parameter given an extreme score is greater than the posterior variance of the ability parameter given an intermediate score (see, for instance, Hoijtink & Boomsma, 1995, pp. 59, Table 4.1). So summing up, the first motive for an IRT approach to multilevel models presented here is the more realistic treatment of measurement error. The second motive is that, contrary to observed scores, latent scores are test-independent, which offers the possibility of analyzing data from incomplete designs, such as, for instance, matrix-sampled

educational assessments, where different (groups of) persons respond to different (sets of) items.

An important difference between the approach by Adams et al. (1997) and Mislevy and Bock (1989) and the present one is the estimation procedure: In the earlier approaches marginal maximum likelihood (MML) and Bayes modal procedures (see, for instance, Bock & Aitkin, 1981; Mislevy, 1986) were used, while the present approach entails a fully Bayesian procedure. Below, it will be shown that adopting a fully Bayesian framework results in a straightforward and easily implemented estimation procedure. The procedure has several advantages. First, a fully Bayesian procedure supports definition of a full probability model for quantifying uncertainty in statistical inferences (see, for instance, Gelman, Carlin, Stern, & Rubin, 1995, pp. 3). Both knowledge about previous research and the data collection process can be incorporated in the model. Second, estimates of model parameters that might otherwise be poorly determined by the data can be enhanced by imposing restrictions on these parameters via their prior distributions. For example, priors can be placed on the variance components in case of a small number of Level 2 units (see, for example, Seltzer, Wong, & Bryk, 1996). The third, and probably most important advantage, has to do with the following. The framework used here is closely related to the framework introduced by Albert (1992). Recently, this framework has been further elaborated for estimation of IRT models with multiple raters (Patz & Junker, 1999b), testlet structures (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow, & Du, 2000), latent classes (Hoijtink & Molenaar, 1997) and multidimensional latent abilities (Béguin & Glas, 2001). The unifying theme of these applications is the use of a Markov chain Monte Carlo (MCMC) method for Bayesian inferences. The motivation for the recent interest in Bayesian inference and MCMC might be that the complex dependency structures in the mentioned models require the evaluation of multiple integrals to solve the estimation equations in an MML or Bayes modal framework (Patz & Junker, 1999a). In the sequel, it will become clear that these problems are easily avoided in an MCMC framework. This point will be returned to in the discussion section.

This chapter consists of five sections. After this introduction section, a general multilevel IRT model will be presented. In the next section, an MCMC estimation procedure will be described. Then, in the following section, examples of the procedure will be given. And finally, the last section contains a discussion and suggestions for further research.

2. Multilevel IRT Models

2.1 One-Way Random Effects IRT ANOVA

Before describing the complete model considered here, a special case will be presented first to illustrate the dependency structure of a multilevel IRT model. Consider a population of units, say schools, from which a sample of units indexed $j = 1, \ldots, J$ is drawn. Individuals, say students indexed $i = 1, \ldots, n_j$, are nested within units. In this framework, Bryk and Raudenbush (1992) consider a two-Level one-way random effects ANOVA model. For the first level, the model is given by

$$Y_{ij} = \beta_j + e_{ij}, \text{ with } e_{ij} \sim N(0, \sigma^2), \tag{3.1}$$

the second level is given by

$$\beta_j = \gamma + u_j, \text{ with } u_j \sim N(0, \tau^2). \tag{3.2}$$

So the model entails that the Level 1 unit means are sampled from a normal distribution with mean γ and variance τ^2. Persons within a unit are independent and the disturbances of the regression coefficients in different schools are uncorrelated. This model can be generalized to an IRT framework by imposing the linear structure on unobserved latent variables θ_{ij} rather than on observed variables Y_{ij}. The assumption is introduced that unidimensional ability parameters θ_{ij} are independent and normally distributed given β_j. So let $\theta_{ij} \mid \beta_j \sim N(\beta_j, \sigma^2)$. Further, $\beta_j \sim N(\gamma, \tau^2)$. Combining these two assumptions, it follows that the joint distribution of the ability parameters and the random regression coefficient in group j is multivariate normal, that is,

$$\begin{bmatrix} \theta_{1j} \\ \theta_{2j} \\ \vdots \\ \theta_{n_j j} \\ \beta_j \end{bmatrix} \sim N \left[\begin{bmatrix} \gamma \\ \gamma \\ \vdots \\ \gamma \\ \gamma \end{bmatrix}, \begin{bmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 & \tau^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \cdots & \tau^2 & \tau^2 \end{bmatrix} \right]. \tag{3.3}$$

So, though local independence holds within groups, over groups the ability parameters of the respondents are dependent. As noted above, these kinds of complex correlated structures suggest using a fully Bayesian rather than an MML or Bayes modal approach. However, this does not mean that the latter two approaches are completely infeasible for the present model, this point will be returned to in the discussion.

2.2 A Multilevel IRT Model

Bryk and Raudenbush (1992) present the above one-way random effects ANOVA model as a special case of a general model. In an IRT context, this model translates to a model given by

$$\theta_{ij} = \beta_{0j} + \ldots + \beta_{qj}X_{qij} + \ldots + \beta_{Qj}X_{Qij} + e_{ij}, \tag{3.4}$$

with $e_{ij} \sim N(0, \sigma^2)$, and

$$\beta_{qj} = \gamma_{q0} + \ldots + \gamma_{qs}W_{sqj} + \ldots + \gamma_{qS}W_{Sqj} + u_{qj}, \tag{3.5}$$

for $q = 0, \ldots, Q$. The Level 2 error terms, u_{qj}, $q = 0, \ldots, Q$, have a multivariate normal distribution with a mean equal to zero and a covariance matrix $\mathbf{T}$. In (3.4), X_{qij} and β_{qj} are Level 1 predictor variables and regression coefficients, respectively. The latter are assumed to be random variables modeled by (3.5), where W_{sqj} and γ_{qs} are Level 2 predictor variables and regression coefficients, respectively.

In the above formulation, the coefficients of all the predictors in the Level 1 model are treated as random, that is, as varying across Level 2 units. In certain applications, it can be desirable to constrain the effects of one or more of the Level 1 predictors to be identical across Level 2 units. This is accomplished by reformulating the hierarchical model as a mixed model (Raudenbush, 1988). The issues and procedures discussed below also apply to these mixed model settings.

Up to this point, the ability parameter $\boldsymbol{\theta}$ is unspecified and unknown. In the next section, an IRT model and an estimation procedure will be introduced.

3. An MCMC Estimation Procedure for a Multilevel IRT Model

Recently, Albert (1992) derived a procedure for simulating sampling from the posterior distribution of the item and person parameters of the two-parameter normal ogive model using the Gibbs sampler (Gelfand, Hills, Racine-Poon, & Smith, 1990; Gelman et al., 1995; Geman & Geman, 1984). In this paper, this approach will be generalized to the multilevel IRT model considered above. In the normal ogive model, the probability of a correct response of a person indexed ij on an item indexed k ($k = 1, ..., K$), $Y_{ijk} = 1$, is given by

$$P\left(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right) = \Phi\left(a_k\theta_{ij} - b_k\right), \tag{3.6}$$

where Φ denotes the cumulative standard normal distribution function, and a_k and b_k are the discrimination and difficulty parameter of item

k, respectively. Below, the parameters of item k will also be denoted by ξ_k, with $\xi_k = (a_k, b_k)^t$ (note that item difficulty is denoted by the usual choice b while regression coefficients are denoted by β, which is the usual choice in linear regression models. These parameters should not be confused).

In a Bayesian framework, the parameters in the model defined by (3.4), (3.5), and (3.6) are viewed as random variables. Inferences about the parameters are made in terms of their posterior distribution. However, as will be shown below, the simultaneous posterior distribution of all model parameters is quite complicated. Therefore, the complete set of parameters is split up into a number of subsets in such a way that the conditional posterior distribution of every subset given all other parameters has a tractable form and can be easily sampled. An MCMC procedure will be used for drawing samples from the conditional posterior distributions. The MCMC chains will be constructed using the Gibbs sampler.

To implement the Gibbs sampler for the normal ogive model, Albert (1992) augments the data by introducing independent random variables Z_{ijk}, which are assumed to be normally distributed with mean $a_k\theta_{ij} - b_k$ and variance equal to one. It is assumed that $Y_{ijk} = 1$ if $Z_{ijk} > 0$ and $Y_{ijk} = 0$ otherwise. Let $\mathbf{Z} = (Z_{111}, \ldots, Z_{n_J JK})$ with realization $\mathbf{z} = (z_{111}, \ldots, z_{n_J JK})$ and let $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ be the vectors of all person and item parameters, respectively. Though the joint distribution of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi})$ has an intractable form, the fully conditional distribution of each of the three parameters are easy to simulate. Each iteration m consists of three steps: (1) draw $\mathbf{Z}^{m+1}$ from its distribution given $\boldsymbol{\xi}^m$ and $\boldsymbol{\theta}^m$, (2) draw $\boldsymbol{\theta}^{m+1}$ from its distribution given $\mathbf{Z}^{m+1}$ and $\boldsymbol{\xi}^m$, and (3) draw $\boldsymbol{\xi}^{m+1}$ from its distribution given $\mathbf{Z}^{m+1}$ and $\boldsymbol{\theta}^{m+1}$. In the next section, it will be shown that this idea can be extended to estimation of the posterior distribution of all parameters in the multilevel IRT model.

3.1 Estimation of the Multilevel IRT Model using Gibbs Sampling

In the present case, the data consist of the item responses $\mathbf{Y}$, and the values of the Level 1 and 2 explanatory variables, denoted by $\mathbf{X}$ and $\mathbf{W}$, respectively. Besides the parameters $\mathbf{Z}, \boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the model has as parameters the Level 1 regression coefficients $\boldsymbol{\beta}$, the Level 2 coefficients $\boldsymbol{\gamma}$, and the variance components σ^2 and $\mathbf{T}$. As a result, the full posterior

distribution of the parameters given the data is given by

$$p\left(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma},\mathbf{T} \mid \mathbf{y},\mathbf{X},\mathbf{W}\right) \propto \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left[\left[\prod_{k=1}^{K} p\left(z_{ijk} \mid \theta_{ij},\boldsymbol{\xi}_k,y_{ijk}\right)\right] p\left(\theta_{ij} \mid \boldsymbol{\beta}_j,\sigma^2,\mathbf{X}_j\right)\right] p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma},\mathbf{T},\mathbf{W}_j\right) p\left(\boldsymbol{\gamma} \mid \mathbf{T}\right) p\left(\boldsymbol{\xi}\right) p\left(\sigma^2\right) p\left(\mathbf{T}\right),$$

with $\boldsymbol{\beta}_j$, $\mathbf{X}_j$ and $\mathbf{W}_j$ the Level 1 regression coefficients and the Level 1 and 2 explanatory variables of group j, respectively. The exact definition of $\mathbf{X}_j$ and $\mathbf{W}_j$ as matrices will be returned to below. From the definition of Z_{ijk} it follows that

$$p\left(z_{ijk} \mid \theta_{ij},\boldsymbol{\xi}_k,y_{ijk}\right) \propto \phi\left(z_{ijk};a_k\theta_{ij}-b_k,1\right)\left[I\left(z_{ijk}>0\right)I\left(y_{ijk}=1\right) + I\left(z_{ijk}\leq 0\right)I\left(y_{ijk}=0\right)\right],$$

where $\phi(.;a_k\theta_{ij}-b_k,1)$ stands for the normal density with a mean equal to $a_k\theta_{ij}-b_k$ and a variance equal to one, and $I\left(.\right)$ is an indicator variable taking the value one if its argument is true, and taking the value zero otherwise.

As with the basic two-parameter IRT model (see, for instance, Bock & Aitkin, 1981) the model must be identified by fixing the origin and scale of the latent dimension. Usually, this can be done by fixing the mean and the variance of the ability distribution to zero and one, respectively. An alternative is imposing the identifying restrictions on the item parameters. Since imposing $\prod_k a_k = 1$ and $\sum_k b_k = 0$ would require rescaling all drawn values in every iteration, a convenient way is to fix one discrimination parameter to one, and one difficulty to zero.

Assuming independence between the item difficulty and discrimination parameter simplifies the choice of the prior, because independent sets of parameters may be considered separately. A noninformative prior for the difficulty and discrimination parameter, which insures that each item will have a positive discrimination index, leads to the simultaneous noninformative prior $p\left(\boldsymbol{\xi}\right) = p\left(\mathbf{a}\right)p\left(\mathbf{b}\right) \propto \prod_{k=1}^{K} I\left(a_k>0\right)$. The other priors will be discussed below. The full posterior distribution has an intractable form and will be very difficult to simulate. Therefore, a Gibbs sampling algorithm will be used where the three steps of the original algorithm by Albert (1992) are extended to seven steps. Each step consist of sampling from the posterior of one of the seven parameter vectors $\mathbf{Z},\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma},\mathbf{T}$ conditionally on all other parameters. These fully conditional distributions are each tractable and easy to simulate. So the remaining problem is finding the conditional distributions of $\mathbf{Z},\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2$ and $\mathbf{T}$, respectively.

Step 1: Sampling $\mathbf{Z}$. Given the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the variables Z_{ijk} are independent, and

$$Z_{ijk} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y} \sim N\left(a_k \theta_{ij} - b_k, 1\right) \tag{3.7}$$

truncated at the left by 0 if $Y_{ijk} = 1$, truncated at the right by 0 if $Y_{ijk} = 0$.

Step 2: Sampling $\boldsymbol{\theta}$. The ability parameters are independent given $\mathbf{Z}, \boldsymbol{\xi}, \boldsymbol{\beta}$ and σ^2. Using equation (3.5) and (3.7) it follows that

$$\begin{aligned} p\left(\theta_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2\right) &\propto p\left(\mathbf{z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right) p\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2\right) \\ &\propto \exp\left(\frac{-1}{2} \sum_{k=1}^{K} \left(z_{ijk} + b_k - a_k \theta_{ij}\right)^2\right) \\ &\quad \exp\left(\frac{-1}{2\sigma^2} \left(\theta_{ij} - \mathbf{X}_{ij} \boldsymbol{\beta}_j\right)^2\right), \end{aligned} \tag{3.8}$$

where $\mathbf{X}_{ij}$ is a matrix of the explanatory variables of person i of group j, that is, $\mathbf{X}_{ij} = \left(X_{0ij}, \ldots, X_{Qij}\right)^t$.

Inspection shows that (3.8) is a normal model for the regression of $Z_{ijk} + b_k$ on a_k with θ_{ij} as a regression coefficient, where θ_{ij}, has a normal prior parameterized by $\boldsymbol{\beta}_j$ and σ^2 (e.g., see, Box & Tiao, 1973, pp. 74-75; Lindley & Smith, 1972). So the fully conditional posterior density of θ_{ij} is given by

$$\theta_{ij} \mid \mathbf{Z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2 \sim N\left(\frac{\widehat{\theta}_{ij}/v + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/v + 1/\sigma^2}, \frac{1}{1/v + 1/\sigma^2}\right), \tag{3.9}$$

with

$$\widehat{\theta}_{ij} = \frac{\sum_{k=1}^{K} a_k \left(z_{ijk} + b_k\right)}{\sum_{k=1}^{K} a_k^2},$$

and $v = \left(\sum_{k=1}^{K} a_k^2\right)^{-1}$.

Step 3: Sampling $\boldsymbol{\xi}$. Conditional on $\boldsymbol{\theta}$, $\mathbf{Z}_k = \left(Z_{11k}, \ldots, Z_{n_J Jk}\right)^t$ satisfies the linear model

$$\mathbf{Z}_k = \begin{bmatrix} \boldsymbol{\theta} & -\mathbf{1} \end{bmatrix} \boldsymbol{\xi}_k + \varepsilon_k, \tag{3.10}$$

where $\boldsymbol{\varepsilon}_k = (\varepsilon_{11k}, \ldots, \varepsilon_{n_J Jk})^t$ is a random sample from $N(\mathbf{0}, \mathbf{I}_N)$. Combining (3.10) with the prior for $\boldsymbol{\xi}$, it follows that

$$\begin{aligned} p(\boldsymbol{\xi}_k \mid \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) &\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(z_{ijk}; a_k\theta_{ij} - b_k, 1)\, p(\boldsymbol{\xi}_k) \\ &\propto \exp\left(\frac{-1}{2}(\mathbf{z}_k - \mathbf{H}\boldsymbol{\xi}_k)^t (\mathbf{z}_k - \mathbf{H}\boldsymbol{\xi}_k)\right) p(\boldsymbol{\xi}_k) \end{aligned}$$

with $\mathbf{H} = \begin{bmatrix} \boldsymbol{\theta} & -\mathbf{1} \end{bmatrix}$. Therefore,

$$\boldsymbol{\xi}_k \mid \boldsymbol{\theta}, \mathbf{Z}_k \sim N\left(\widehat{\boldsymbol{\xi}}_k, \left(\mathbf{H}^t\mathbf{H}\right)^{-1}\right) I(a_k > 0), \tag{3.11}$$

where $\widehat{\boldsymbol{\xi}}_k$ is the usual least squares estimator based on (3.10).

Step 4: Sampling $\boldsymbol{\beta}$. Define $\mathbf{X}_j = \left(\mathbf{X}_{1j}, \ldots, \mathbf{X}_{ij}, ..., \mathbf{X}_{n_j j}\right)^t$, with $\mathbf{X}_{ij}$ as defined in Step 2. Further, $\mathbf{W}_j$ is the direct product of $\mathbf{W}_{qj} = (W_{0qj}, \ldots, W_{Sqj})^t$ and a $(Q+1)$ identity matrix, that is, $\mathbf{W}_j = \{\mathbf{W}_{qj}\} \otimes \mathbf{I}_{Q+1}$ (the direct product is also known as tensor product or Kronecker product). Then the fully conditional posterior density of $\boldsymbol{\beta}_j$ is given by

$$\begin{aligned} p(\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}) &\propto p(\boldsymbol{\theta}_j \mid \boldsymbol{\beta}_j, \sigma^2)\, p(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}) \\ &\propto \exp\left(\frac{-1}{2\sigma^2}\left(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right)^t \mathbf{X}_j^t\mathbf{X}_j \left(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right)\right) \\ &\quad \exp\left(\frac{-1}{2}(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})^t \mathbf{T}^{-1}(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})\right) \end{aligned}$$

with $\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1} \mathbf{X}_j^t\boldsymbol{\theta}_j$. Notice that the fully conditional posterior of $\boldsymbol{\beta}_j$ entails a model for the regression of $\boldsymbol{\theta}_j$ on $\mathbf{X}_j$, with $\boldsymbol{\beta}_j$ as regression coefficients, where the regression coefficients have a normal prior induced by the Level 2 model (3.4), that is, the regression of $\boldsymbol{\beta}_j$ on $\mathbf{W}_j$. Define $\Sigma_j = \sigma^2\left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1}$, $\mathbf{d} = \Sigma_j^{-1}\widehat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1}\mathbf{W}_j\boldsymbol{\gamma}$ and $\mathbf{D} = \left(\Sigma_j^{-1} + \mathbf{T}^{-1}\right)^{-1}$. Then it follows that

$$\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T} \sim N(\mathbf{Dd}, \mathbf{D}). \tag{3.12}$$

Step 5: Sampling $\boldsymbol{\gamma}$. The matrix $\boldsymbol{\gamma}$ is the matrix of regression coefficients for the regression of $\boldsymbol{\beta}_j$ on $\mathbf{W}_j$. The unbiased estimator for $\boldsymbol{\gamma}$ will

be the generalized least squares estimator. Because

$$p\left(\boldsymbol{\gamma} \mid \boldsymbol{\beta}_j, \mathbf{T}\right) \propto \prod_{j=1}^{J} p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right) p\left(\boldsymbol{\gamma} \mid \mathbf{T}\right)$$
$$\propto \exp\left(\frac{-1}{2} \sum_{j=1}^{J} \left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1} \left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)\right),$$

using an improper noninformative prior density for $\boldsymbol{\gamma}$ results in

$$\boldsymbol{\gamma} \mid \boldsymbol{\beta}_j, \mathbf{T} \sim N(\widetilde{\boldsymbol{\gamma}}, \Omega) \tag{3.13}$$

where

$$\widetilde{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \mathbf{W}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \boldsymbol{\beta}_j,$$

as the generalized least squares estimator for $\boldsymbol{\gamma}$ and

$$\Omega = \left(\sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \mathbf{W}_j\right)^{-1}$$

is the conditional posterior variance.

Step 6: Sampling σ^2. The conjugate prior density for the variance σ^2 is the $Inv-\chi^2\left(v_0, \sigma_0^2\right)$. Upon setting $v_0 = 0$, it follows that the noninformative prior density for the variance is $p\left(\sigma^2\right) \propto \sigma^{-2}$. Then the conditional posterior distribution for σ^2 is given by

$$p\left(\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) \propto p\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \sigma^2\right) p\left(\sigma^2\right)$$
$$\propto \left(\sigma^2\right)^{-\left(\frac{N}{2}+1\right)} \exp\left(\frac{-N}{2\sigma^2} S^2\right),$$

with $S^2 = \frac{1}{N}\sum_{j=1}^{J} \left(\boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\beta}_j\right)^t \left(\boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\beta}_j\right)$. Thus, the posterior distribution of σ^2 given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is an inverse-chi-square distribution, that is,

$$\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta} \sim Inv-\chi^2\left(N, S^2\right). \tag{3.14}$$

The prior density for the variance σ^2 is improper, but yields a proper conditional posterior density for σ^2.

Step 7: Sampling $\mathbf{T}$. Above, $\mathbf{W}_j$ and $\boldsymbol{\beta}_j$ are defined as the matrix of explanatory variables and the vector of regression coefficients for Level

2 unit j, respectively. The Level 2 model for this unit can be written as $\boldsymbol{\beta}_j = \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{u}_j$,with $E(\mathbf{u}_j) = 0$, $E\left(\mathbf{u}_j\mathbf{u}_j^t\right) = \mathbf{T}$. Therefore,

$$\begin{aligned} p\left(\mathbf{T} \mid \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) &\propto p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right) p\left(\mathbf{T}\right) \\ &\propto |\mathbf{T}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)\right) p\left(\mathbf{T}\right). \end{aligned}$$

Subsequently, define $\mathbf{S} = \sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)^t$ and assume a non-informative prior for $\mathbf{T}$. Aggregating over Level 2 units results in

$$\begin{aligned} p\left(\mathbf{T} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}\right) &\propto |\mathbf{T}|^{-\frac{J}{2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)\right) p\left(\mathbf{T}\right) \\ &= |\mathbf{T}|^{-\frac{J}{2}} \exp\left(-\frac{1}{2} tr\left(\mathbf{S}\mathbf{T}^{-1}\right)\right) p\left(\mathbf{T}\right) \\ &= |\mathbf{T}|^{-\left(\frac{J}{2}+1\right)} \exp\left(-\frac{1}{2} tr\left(\mathbf{S}\mathbf{T}^{-1}\right)\right), \end{aligned}$$

and the posterior distribution of $\mathbf{T}$ given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is an inverse-Wishart distribution, that is,

$$\mathbf{T} \mid \boldsymbol{\beta}, \boldsymbol{\gamma} \sim inv - Wishart\left(J, \mathbf{S}^{-1}\right). \tag{3.15}$$

With initial values $\boldsymbol{\theta}^{(0)}$,$\boldsymbol{\xi}^{(0)}$,$\boldsymbol{\beta}^{(0)}$, $\sigma^{2(0)}$, $\boldsymbol{\gamma}^{(0)}$,and $\mathbf{T}^{(0)}$, the Gibbs sampler iteratively samples $\mathbf{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ^2 and $\mathbf{T}$ from the distributions (3.7), (3.9), (3.11), (3.12), (3.13), (3.14) and (3.15). The components are updated in the order given by steps 1-7 above. Roberts and Sahu (1997) showed that a different updating strategy can affect the speed of convergence. Furthermore, they show that in case of a hierarchically structured problem the strategy of iteratively updating the components in the fixed ordering is the best.

The values of the initial estimates are also important for the rate of convergence. When poor initial values are chosen, convergence will be very slow. Consider, for example, (3.9). When the parameters of the multilevel model are estimated conditional on poor estimates of $\boldsymbol{\theta}$, the poor estimates of the multilevel model parameters will subsequently produce poor estimates of the ability parameters. This is because, in Step 2 the prediction of $\boldsymbol{\theta}$ from the multilevel model will dominate the sampled values of $\boldsymbol{\theta}$ when the Level 1 residual variance σ^2 is smaller than the variance of $\widehat{\boldsymbol{\theta}}$, that is, v. So after some iterations, all the sampled values of the parameters are far away from the optimal parameter values, while σ^2 remains smaller than v. It will take a lot of iterations to alter this

pattern. Therefore, the following procedure can be used to obtain better initial estimates. First, MML estimates of the item parameters are made under the usual assumption that $\boldsymbol{\theta}$ is normally distributed with $\mu = 0$ and $\sigma = 1$ (see, Bock & Aitkin, 1981; Mislevy, 1986). Another suggestion might be to compute initial values using a distinct ability distribution for every subgroup j. These estimates can be computed using the program Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Then, using draws from the normal approximation of the standard errors of the parameter estimates of Bilog-MG as starting values, the MCMC procedure by Albert (1992) for estimating the normal ogive model can be run. That is, with the assumption that $\boldsymbol{\theta}$ is standard normal distributed formula (3.9) becomes

$$\theta_{ij} \mid Z_{ijk}, \boldsymbol{\xi} \sim N\left(\frac{\widehat{\theta}_{ij}/v}{1/v+1}, \frac{1}{1/v+1}\right), \tag{3.16}$$

and $\mathbf{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ can be sampled from the distributions (3.7), (3.16) and (3.11). As the Gibbs sampler has reached convergence, the means of the sampled values of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi})$ are computed to start sampling from the distributions (3.12), (3.13), (3.14) and (3.15). After convergence, means of the sampled values of $\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T}\right)$ are used as initial estimates. It is also possible to use an EM algorithm for estimating $\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T}\right)$ with the $\widehat{\boldsymbol{\theta}}$ (see, for instance, Bryk & Raudenbush, 1992). Once all initial values are estimated, equation (3.16) can be replaced by (3.9), and the complete seven-step MCMC procedure can be started for an estimation of $\left(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T}\right)$.

4. Simulated and Real-Data Examples

In this section, a simulated data set and a data set from a Dutch primary school mathematics test are analyzed. The simulated data set will be used to illustrate the parameter recovery with the Gibbs sampler. The Dutch primary school mathematics test will be used to illustrate the practical impact of the proposed multilevel IRT model.

4.1 A Numerical Example

To illustrate parameter recovery, data were simulated using a multilevel model with one explanatory variable on both levels. The model is given by

$$\begin{aligned} \theta_{ij} &= \beta_{0j} + \beta_{1j} X_{1ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_{10j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_{11j} + u_{1j}, \end{aligned} \tag{3.17}$$

with $e_{ij} \sim N\left(0, \sigma^2\right)$ and $u_{qj} \sim N\left(0, \tau_q^2\right)$. Response patterns were generated according to a normal ogive IRT model for a test of $K = 20$ dichotomous items. The generating values of the item parameters are shown under the label Generated in Table 3.1. The ability parameters of $2,000$ students were divided over $J = 10$ groups of $n_j = 200$ students each, and generated with the multilevel model given by (3.17). The true values for the fixed effects $\boldsymbol{\gamma}$ and the variance components τ_0, τ_1 and σ are shown under the label Generated in Table 3.2. The explanatory variables $\mathbf{X}$ and $\mathbf{W}$ were drawn from $N(0,1)$ and $N(1/2,1)$, respectively.

With Bilog-MG estimates as starting values, the normal ogive model was estimated with the MCMC procedure of Albert (1992). Subsequently, the parameters of the multilevel model were sampled, given the parameters of the normal ogive model. In the simulation study, 500 iterations were needed to estimate the normal ogive model and another 500 iterations were needed to compute the parameters of the multilevel model. After that, $20,000$ iterations were made to estimate the parameters of the multilevel IRT model[1]. The convergence of the Gibbs sampler was checked by monitoring the expected a posteriori estimate of each parameter and its posterior standard deviation for several consecutive sequences of $1,000$ iterations. The Gibbs sampler has reached convergence if differences are small. The sample variance of the individual draws was used as an estimator for the posterior variance (see, for instance, Patz & Junker, 1999b).

In Table 3.1, the estimates of the item parameters issued from the Gibbs sampler are given under the label Gibbs Sampler. The item parameter estimates are the means of the generated posterior distributions. The reported standard deviations are the estimated posterior standard deviations. In the Bayesian framework, credibility intervals are calculated as confidence regions for the parameters and they are given in the column labeled CI. These credibility intervals are the 95%-equal-tailed-intervals whose endpoints are the 2.5 and 97.5 percentiles of the marginal posterior distribution of the parameters.

Figure 3.1 presents the posterior densities of a_k for four specific items. In each plot of Figure 3.1, two lines are plotted representing the density estimates based on 500 and $20,000$ simulated values, respectively. It can be seen that the first 500 values, which were produced with the Gibbs sampler to get initial estimates, were quite removed from the final estimates.

[1] On a Pentium II 400mHz computer, $20,000$ iterations took about 10 hours. The S-Plus (Mathsoft, 1999) code can be downloaded from http://users.edte.utwente.nl/fox.

Table 3.1. Item parameter estimates of the normal ogive IRT model using the Gibbs sampler.

	Generated		Gibbs Sampler					
Item	a_k	b_k	a_k	s.d.	CI	b_k	s.d.	CI
1	.640	.004	.689	.056	[.587, .809]	0	0	[0, 0]
2	1.013	−.019	.982	.072	[.852, 1.137]	−.012	.054	[−.124, .085]
3	.939	−.508	.954	.072	[.826, 1.107]	−.511	.055	[−.626, −.411]
4	.780	−.066	.746	.058	[.638, .869]	−.117	.045	[−.208, −.031]
5	.824	−.180	.896	.067	[.776, 1.038]	−.212	.050	[−.316, −.123]
6	.772	−.017	.832	.063	[.717, .964]	−.016	.048	[−.113, .075]
7	.903	−.942	.848	.068	[.726, .991]	−.891	.053	[−1.002, −.793]
8	.789	.168	.823	.063	[.710, .955]	.108	.047	[.011, .194]
9	.915	.000	.877	.066	[.758, 1.021]	−.002	.049	[−.104, .088]
10	.967	.603	.998	.075	[.860, 1.156]	.563	.054	[.450, .663]
11	1.087	−.010	1.093	.078	[.951, 1.261]	−.032	.057	[−.152, .074]
12	.980	−.506	1.047	.077	[.909, 1.212]	−.549	.057	[−.667, −.441]
13	1.124	.458	1.111	.080	[.963, 1.281]	.413	.059	[.290, .520]
14	.945	−.691	.938	.071	[.814, 1.093]	−.679	.054	[−.791, −.580]
15	1.039	−.235	1.012	.072	[.880, 1.167]	−.263	.055	[−.378, −.164]
16	1.002	−.402	1	0	[1, 1]	−.371	.053	[−.479, −.271]
17	.676	.451	.602	.052	[.506, .713]	.467	.040	[.386, .544]
18	.845	−.578	.824	.064	[.709, .961]	−.588	.050	[−.691, −.496]
19	.796	.052	.943	.069	[.818, 1.092]	.046	.051	[−.060, .142]
20	.722	.115	.799	.061	[.689, .931]	.106	.046	[.012, .191]

Table 3.2 presents the results of the estimation of the fixed effects and the variance components of the model. Notice that the conventional multilevel terminology is still used although all parameters were treated as random in the estimation procedure. The posterior means and standard deviations estimates computed with the Gibbs sampler are given under the label Gibbs Sampler. It can be seen that the true parameter values are well within the computed credibility intervals except for γ_{10} and γ_{11}. As an additional check on the procedure, the fixed effects and variance components were also estimated from the true ability parameters $\boldsymbol{\theta}$ using HLM for Windows (Bryk, Raudenbush, & Congdon, 1996). In practice, these ability parameters are, of course, unknown. Inspection shows that the estimates issued by the two methods were quite close. That is, the parameter values from HLM are well within the computed credibility intervals. The estimates resulting from HLM are based on the true ability parameter, which results in more accurate estimates. It seems that a fully Bayesian method which includes all the uncertainty in

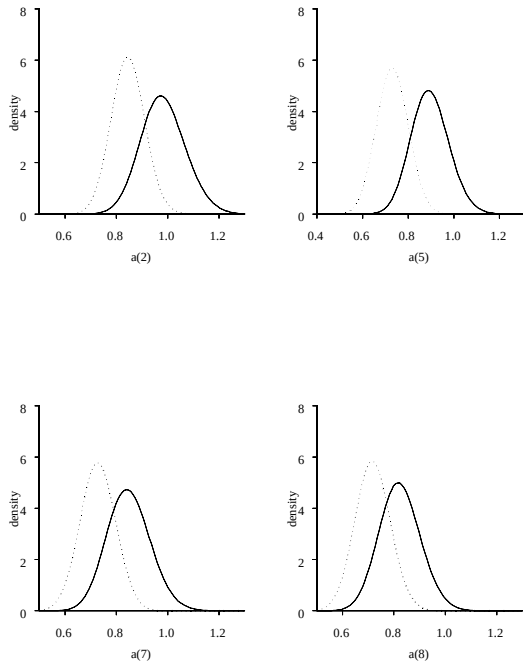

Figure 3.1. Posterior densities of a_k for items 2, 5, 7 and 8. Dotted line is an estimate of density after 500 values, and solid line is an estimate of density after 20,000 values.

the problem needs larger sample sizes to make adequate inferences. On the other hand, comparing MML and fully Bayesian estimates of an IRT model for responses to testlets, Glas, Wainer, and Bradlow (2000) argue that the smaller size of the frequentist confidence intervals is related to the fact that they are based on an asymptotic approximation that does not take the skewness into account. Obviously, more research comparing the two approaches needs to be done.

Finally, it is of interest to evaluate whether the multilevel IRT model was an improvement over the usual multilevel model on the observed scores. The linear model on the observed scores is less complex than the multilevel IRT model, but it was expected that using observed scores instead of latent scores as dependent variables will result in less precision in parameter recovery. For comparative purposes, the unweighted sums

Table 3.2. Parameter estimates of the multilevel model, with the Gibbs sampler and HLM for Windows.

	Generated	HLM		Gibbs Sampler		
Fixed Effects	Coefficient	Coefficient	s.e.	Coefficient	s.d.	CI
γ_{00}	$-.30$	$-.366$	.116	$-.319$	.182	$[-.681, .041]$
γ_{01}	.15	.291	.150	.209	.238	$[-.270, .690]$
γ_{10}	.35	.411	.042	.478	.061	$[.361, .601]$
γ_{11}	1	.929	.081	.728	.123	$[.486, .971]$
Random Effects	Variance Components	Variance Components		Variance Components	s.d.	CI
τ_0	.1	.131		.150	.018	$[.085, .262]$
τ_1	.1	.091		.097	.007	$[.051, .168]$
σ	.2	.199		.178	.006	$[.136, .205]$

of the item responses were rescaled to a standard normal distribution. These rescaled scores will be called Z-scores. Table 3.3 gives the results of the estimation with HLM for Windows using the true standardized ability parameters and Z-scores. From Table 3.2 and 3.3, it can be verified that the estimates computed using Z-scores differ substantially from the analogous estimates computed under a linear model on the true ability parameters and under a multilevel IRT model. The difference in the estimates of the variance components also had consequences for the estimates of the intraclass correlation coefficient. This coefficient expresses the proportion of variance in ability accounted for by group-membership, after controlling for the Level 1 predictor variable, that is,

$$\widehat{\rho}_0 = \frac{\widehat{\tau}_0}{\widehat{\tau}_0 + \widehat{\sigma}^2}.$$

From the results of Table 3.2, it can be verified that using the HLM estimates based on the true ability parameters resulted in $\widehat{\rho}_0 = .397$, while using the estimates from Gibbs sampler resulted in $\widehat{\rho}_0 = .457$. Notice that the same intraclass correlation coefficient is obtained using the variance components of the true standardized ability parameters as shown in Table 3.3. This shows that this measure is scale-independent.

Table 3.3. Parameter recovery of the multilevel model with standardized true latent scores and Z-scores as dependent variables.

	HLM		HLM (sum scores)	
Fixed Effects	Coefficient	s.e.	Coefficient	s.e.
γ_{00}	−.241	.133	−.191	.140
γ_{01}	.336	.173	.261	.184
γ_{10}	.474	.049	.555	.049
γ_{11}	1.071	.093	.704	.098
Random Effects	Variance Components		Variance Components	
τ_0	.151		.144	
τ_1	.105		.097	
σ	.229		.462	

From the results of Table 3.3, it can be verified that using the Z-scores resulted in $\widehat{\rho}_0 = .238$. So the conclusions drawn from a multilevel IRT model can be quite different from the conclusions drawn from a more traditional multilevel analysis.

4.2 A Dutch Primary School Mathematics Test

This section concerns a study of a primary school leaving test. A multilevel IRT model and an hierarchical linear model using observed scores were estimated and compared. One of the research questions in the study was whether schools that participate on a regular basis in the central primary school leaving test in the Netherlands perform better than schools that do not participate on a regular basis. To investigate this research question, the students of 97 schools were given a mathematics test for grade 8 students. The test consisted of 18 mathematics items taken from the school leaving examination developed by the National Institute for Educational Measurement (Cito). Of the 97 schools sampled, 72 schools regularly participated in the school leaving examination; in the sequel, these schools will be called the Cito schools. The remaining 25 schools will be called the non-Cito schools. The total number of students for which data were available was 2156.

Three students' characteristics were used as a predictor for the students' achievement: socio-economic status (SES), non-verbal intelligence

test (ISI) and Gender. SES was based on four indicators: the education and occupation level of both parents (if present). The non-verbal intelligence test was measured in grade 7 by three parts of an intelligence test. Predictors SES and ISI were normally standardized. The dichotomous predictor Gender is an indicator variable equal to 0 for males and equal to 1 for females. Finally, a predictor variable labeled End equaled 1 if the school participates in the school leaving test, and equals 0 if this is not the case. A complete description of the data can be found in (Doolaard, 1999, pp. 57).

The structural model used in the analysis is given by

$$\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_1 \text{ISI}_{ij} + \beta_{2j} \text{SES}_{ij} + \beta_3 \text{Gender}_{ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} \text{End}_j + u_{0j} \\
\beta_1 &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} + u_{2j} \\
\beta_3 &= \gamma_{30}
\end{aligned} \tag{3.18}$$

where $e_{ij} \sim N\left(0, \sigma^2\right)$, $u_{0j} \sim N\left(0, \tau_0^2\right)$ and $u_{2j} \sim N\left(0, \tau_2^2\right)$. Further, u_{0j} and u_{2j} are assumed independent. Notice that SES is modeled as a random effect, that is, its regression coefficient varies over schools. The two-parameter normal ogive model is used as the measurements model.

The fully conditional decomposition of Gibbs sampling was run for $25,000$ iterations, with a burn-in period of $5,000$ iterations[2]. $25,000$ iterations were "enough" in the sense that a substantial increase in the number of iterations did not perturb values of ergodic averages, that is, the average of the parameter draws over the iterations after the burn-in period.

The multilevel IRT analysis was compared to an analyses with an hierarchical model on observed scores. The score distribution of the mathematics test had a "ceiling", that is, a third of the students scored 15 or more, with a maximum of 18. A standard procedure for dealing with such skewed distributions is to transform the data to normality. This was done by assigning normal order statistics to the ranked scores (Goldstein, 1995, pp. 49). So these so-called N-scores had a standard normal distribution. For comparative purposes, a second transformation was applied to transform these N-scores to the same scale as the latent abilities. This was accomplished by transforming the N-scores such that their mean and variance were equal to the mean and variance of the posterior estimates of the ability parameters, respectively.

[2]Also the S-Plus code for this example can be downloaded from http://users.edte.utwente.nl/fox.

Table 3.4. Parameter estimates of the multilevel model with the Gibbs sampler and HLM using N-scores and rescaled N-scores as dependent variables.

	Gibbs Sampler			HLM (N-scores)		HLM (Rescaled N-scores)	
Fixed Effects	Coeff.	s.d.	CI	Coeff.	s.e.	Coeff.	s.e.
γ_{00}	$-.172$	.214	$[-.589, .242]$	$-.287$	.078	$-.125$	.068
γ_{01}	.467	.242	$[-.006, .943]$	.441	.087	.389	.077
γ_{10}	.445	.034	$[.384, .516]$	.415	.017	.367	.016
γ_{20}	.236	.111	$[.020, .456]$	.213	.023	.188	.020
γ_{30}	$-.181$	.040	$[-.262, -.102]$	$-.167$	.034	$-.148$	.030
Random Effects	Var. Comp.	s.d.	CI	Var. Comp.		Var. Comp.	
τ_0	.410	.041	$[.322, .514]$	.326		.288	
τ_2	.228	.021	$[.153, .324]$	.112		.099	
σ	.644	.056	$[.563, .729]$	.757		.669	

The results of the analyses are displayed in Table 3.4. The remark with respect to the difference in the standard errors made above also applies in the present case.

The main result of the analysis was that conditionally on SES, ISI and Gender, the Cito schools performed better than the non-Cito schools. This can be deduced from the estimate of the fixed effect γ_{01}, which models the contribution of participating in the school leaving exam to the ability level of the students via its influence on the intercept β_{0j}. This intercept β_{0j} is defined as the expected achievement of a male-student in school j when controlling for SES and ISI. There is a highly significant association between the Level 1 predictors ISI and SES and the ability of the students. Obviously, students with high ISI and SES scores performed better than students with lower scores. The effect of Gender on mathematics achievement was also significant and negatively related to achievement. This means that controlling for End, ISI and SES, boys outperformed girls on the mathematics test.

The residual variance for the school-level, τ_0, is the variance of the achievement of male-students in school j, β_{0j}, around the grand mean, γ_{00}, when controlling for SES and ISI. Apparently, a substantial propor-

tion of the variation in the outcome at the student level was between the schools, which justifies the use of a multilevel model.

There were some important differences between the estimates from the multilevel IRT model and the estimates from the HLM model via transformed N-scores. Firstly, the magnitude of the estimate of γ_{01} was greatest in the multilevel IRT analysis, so this approach discriminated more between Cito schools and non-Cito schools. Also the magnitude of the estimate of the variance τ_0^2 was greatest in the multilevel IRT analysis, which indicated more variability in the means in schools of the students' math achievement. Thus, the effect of grouping was greater in the multilevel IRT analysis. Notice that, again, the Bayesian multilevel IRT estimates had larger posterior standard deviations. So the remarks with respect to differences between frequentist and Bayesian credibility intervals made above also applies here.

In the HLM analyses, the variance τ_2^2 did not differ significantly from zero, so the SES-math regression slope did not vary from school to school. This is contrary to the multilevel IRT analysis, where the relationship between SES and math achievement within schools varied significantly across schools. Figure 3.2 displays the predicted abilities of the students in a Cito and a non-Cito school as a function of SES. The points are the expected posterior estimates of the students' abilities.

For the same students as in Figure 3.2, Figure 3.3 shows the predicted transformed N-scores as a function of SES. The points are the transformed N-scores. The abilities and the transformed N-scores in the two plots are corrected for the effects of ISI and Gender. The upper line represents the outcomes of students in a Cito school, which illustrates that students in Cito schools performed better than students in non-Cito schools. Furthermore, the differences between the two lines is greater in Figure 3.2 which illustrates that the subdivision in Cito and non-Cito schools was greater in the estimates resulting from the multilevel IRT analysis. Moreover, Figure 3.2 shows a sharper distinction between schools which indicates a greater school-level effect.

The differences between the estimates can be explained by the fact that the sum scores discriminate less between students' outcomes than the complete response patterns, which is further amplified by the "ceiling" effect which suppresses the variance in the dependent variable. Therefore, the multilevel IRT analysis gauges a greater variance between students' achievements which results in a greater school-level effect whereas the variance at Level 1 is almost the same. In conclusion, the multilevel IRT model reveals a sharper distinction in students' outcomes across schools.

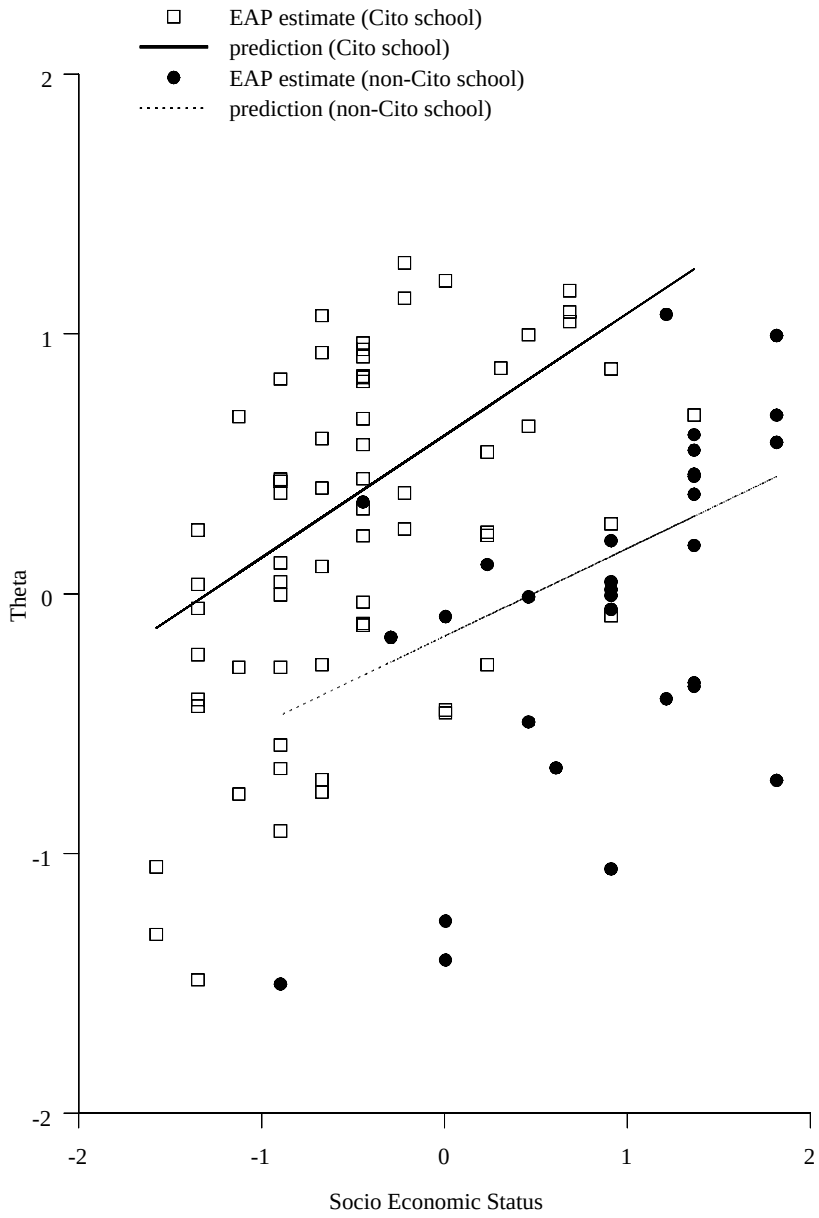

Figure 3.2. Expected posterior estimate and prediction of students' abilities in a Cito and non-Cito school as a function of SES, controlling for ISI and Gender.

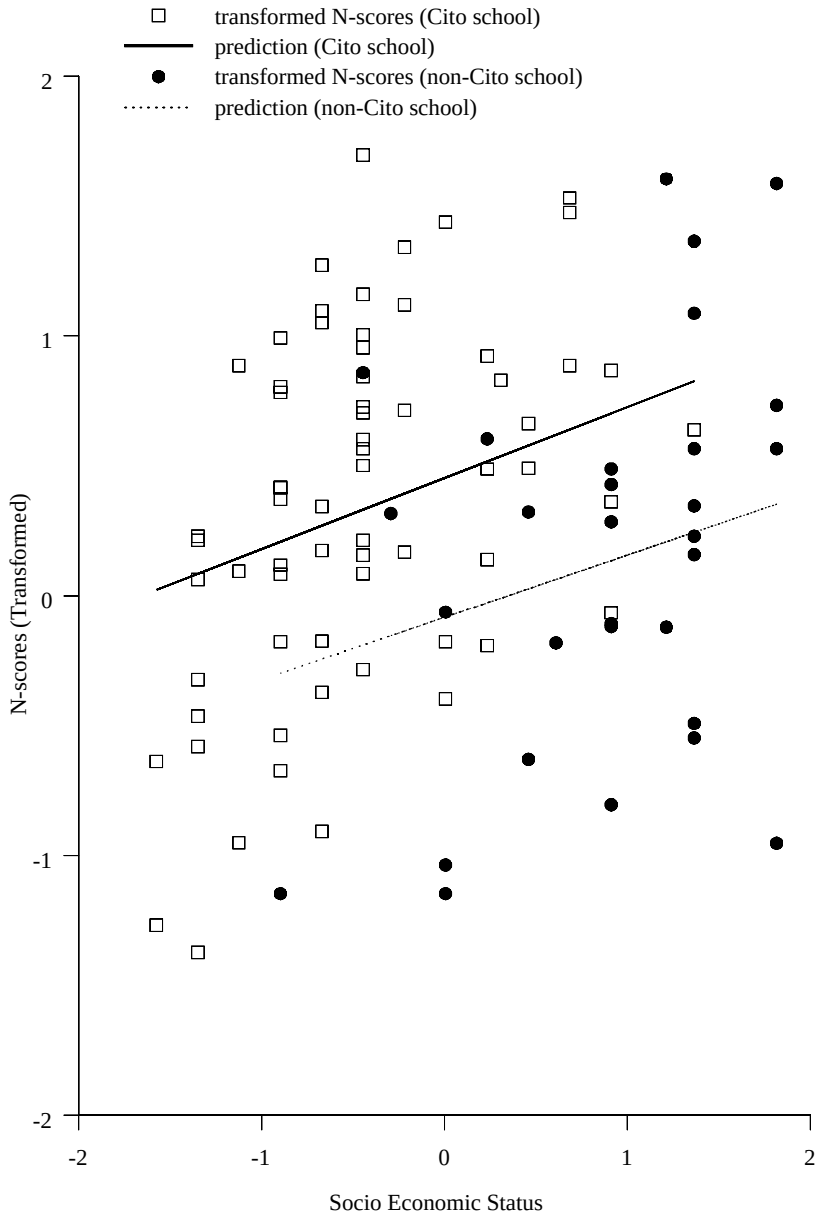

Figure 3.3. Students' N-scores and predicted N-scores in a Cito and non-Cito school as a function of SES, controlling for ISI and Gender.

5. Discussion

In this chapter, a two-level regression model is imposed on the ability parameters of the two-parameter normal ogive model. The advantage of using latent rather than observed scores is that it offers a more realistic way of modeling uncertainty in the dependent variable. Further, latent scores are test-independent, which offers the possibility of entering results from different tests in one analysis.

It was shown that the Gibbs sampler can be used to concurrently estimate all the parameters of the multilevel IRT model. The method presented is very powerful because there are no limitations to the number of parameters or the number of explanatory variables. Although good initial values will speed up convergence, there are still many iterations necessary for producing acceptable estimates. Further research will concentrate on the use of a Monte Carlo EM (MCEM) algorithm to limit the amount of iterations (Wei & Tanner, 1990).

It is easy to incorporate different types of prior beliefs about the item parameters $\boldsymbol{\xi}$. The numerical example illustrates that the posterior distribution of the item discrimination parameters were skewed to the right. Therefore, it could be interesting to use a log-normal prior for the discrimination parameters (Mislevy, 1986). It is also possible to incorporate different priors for $\boldsymbol{\gamma}$, σ^2 or $\mathbf{T}$. In this paper, Jeffreys' prior is used for the variance components, that is, $p\left(\sigma^2\right) \propto \sigma^{-2}$, $p\left(\tau\right) \propto \tau^{-1}$. However, Jeffreys' prior for τ is potentially a problem in cases where J is small (Morris, 1983; Rubin, 1981). Other possible choices of priors for σ^2 and τ are an uniform prior and an inverse-chi-square prior with small degrees of freedom (see, for instance, Seltzer et al., 1996). The inverse-chi-square distribution has the property that, in contrast to the uniform prior, the prior probabilities gradually decrease when values of the variance become arbitrarily large. Analogously, an alternative prior for $\mathbf{T}$ is an inverse-Wishart distribution with small degrees of freedom. Another possibility would be a more informative inverse-chi-square prior or inverse-Wishart prior with mode and spread specified in accordance with previous research. Using nonconjugate prior distributions has the disadvantage that sampling from the fully conditional distributions can be very complicated. In that case, approximations can be used from which sampling is possible. The Metropolis-Hastings algorithm can be used to compensate for the approximation (Gelman et al., 1995, pp. 329).

In this chapter, the focus was on inferences assuming that the model is correct. The problem of model checking using Bayes factors is rather difficult, especially when prior information is weak (O'Hagan, 1995). Posterior predictive data can be used to judge the fit of the Bayesian model

to the observed data. Tail-area probabilities, or posterior p-values, can be calculated under the posited model to quantify the extremeness of the observed value of a selected discrepancy (e.g., differences between observations and predictions). The predictive data are easily sampled via Monte Carlo simulation (see, for example, Gelman, Meng, & Stern, 1996). The Gibbs sampling formulation presented in this chapter can be extended to settings in which the fixed effects are distributed with heavy tails (Seltzer, 1993) to study the extent to which posterior means and intervals change as the degree of heavy-tailedness assumed increases.

Another remark concerns alternative modes of estimation. The first approach might be to use a logit-link in combination with a procedure to estimate a linear multilevel model, such as, for instance, HLM. Applying the logit transformation to the two-parameter logistic model, results in

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = a_k\theta_{ij} - b_k + \varepsilon_{ijk}, \tag{3.19}$$

were p_{ijk} stands for the probability of a correct response and ε_{ijk} is a normally distributed error variable. A linear multilevel model can then be imposed in θ_{ij}. The problem here is that the item discrimination parameters a_k are multiplicative with the ability parameter θ_{ij}, and there is no way to concurrently estimate the item parameters using a package for linear multilevel models. A solution might be to estimate the item parameters using Bilog-MG and impute them into the multilevel logit analysis. However, there are two problems with this approach. First, the uncertainty with respect to the imputed parameters is very difficult to model in the logit analysis. Second, in Bilog-MG the item parameters are estimated under the assumption that the ability parameters are normally distributed. However, the model imposed by (3.4) and (3.5) does not imply a normal distribution of θ_{ij}, and this miss-specification will cause bias in the parameters when the multilevel IRT model holds. The severity of this bias, however, is unknown, and to opt for this approach certainly more research needs to be done.

Another approach to estimating the parameters in the multilevel IRT model might be an MML or Bayes modal procedure. To study this approach in some detail, consider the one-way ANOVA model given in the first section of this chapter. The impact of the dependency structure (3.3) on an MML or Bayes modal estimation procedure can be assessed by inspection of a likelihood function marginalized over all random ef-

fects. This likelihood function can be written as

$$L(\gamma, \sigma^2, \tau, \boldsymbol{\xi}; \mathbf{y}) = \prod_j \int \left[\prod_{i|j} \int p(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}) g(\theta_{ij} \mid \beta_j, \sigma^2) d\theta_{ij} \right] h(\beta_j \mid \gamma, \tau) d\beta_j,$$

where $p(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi})$ is the IRT model specifying the probability of observing response pattern $\mathbf{y}_{ij}$ as a function of the ability parameter θ_{ij} and the item parameters $\boldsymbol{\xi}$, $g(\theta_{ij} \mid \beta_j, \sigma^2)$ is the density of θ_{ij} and $h(\beta_j \mid \gamma, \tau)$ is the density of β_j. It can be seen that the dependency structure results in nesting of integrations that might complicate an MML estimation procedure. Notice that the marginal likelihood entails a multiple integral over θ_{ij} and β_j. Hence there is no need to compute high-dimensional integrals: Computation of two-dimensional integrals suffices. In this respect, this approach to estimation is related to the bi-factor full-information factor analysis model by Gibbons and Hedeker (1992) who show that numerical integration by Gauss-Hermite quadrature is feasible in these problems. Therefore, MML and Bayes modal estimation are still options that deserve further investigation.

Chapter 4

Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory

Abstract It is shown that measurement error in predictor variables can be modeled using item response theory (IRT). Measurement error is modeled by treating the predictors as unobserved latent variables and using the normal ogive model to describe the relation between the latent variables and their observed indicator variables. The predictor variables can be defined at any level of an hierarchical regression model. The predictor variables are latent but can be measured indirectly by using tests or questionnaires. The observed responses on itemized instruments are related to the latent predictors by an item response theory model. It is shown that the multilevel model with measurement error in the predictor variables can be estimated in a Bayesian framework using Gibbs sampling. In this chapter, handling measurement error via the normal ogive model is compared with alternative approaches using the classical true score model. Examples using real data are given.

Keywords: classical test theory, Gibbs sampler, item response theory, Hierarchical Linear Models (HLM), Markov Chain Monte Carlo, measurement error, multilevel model, two-parameter normal ogive model.

1. Introduction

In much research areas, and especially in social sciences, studies may involve variables that cannot be observed directly or are observed subject to error. For example, a person's mathematical ability cannot be measured directly, only the performance on a number of mathematic test items. In general, data collected from respondents contain measurement error. This includes response variation due to the unreliability

of measurement instruments. Further, many forms of human response behavior are inherently stochastic in nature, and also variation stemming from stochastic response behavior will be categorized under the heading measurement error. In this context, Lord and Novick (1968, chapter 2) adhere to the so-called stochastic subject view in which it is assumed that responses of the subjects depend on small variations in the circumstances in which the response is generated. Accordingly, response variance is the variation in responses to the same question repeatedly administered to the same person. The use of unreliable explanatory variables leads to biased estimation of the regression coefficients and the resulting statistical inference can be very misleading unless careful adjustments are made (see, for example, Carroll, Ruppert, & Stefanski, 1995; Cook & Campbell, 1979; Fuller, 1987).

There has been a continuing interest in the study of regression models wherein the independent variables are measured with error. These models are commonly known as measurement error models. The enormous amount of literature on this topic in linear regression is summarized by Fuller (1987) and in this framework, measurement error is handled by the classical additive measurement error model. An example is the classical test theory model used in educational measurement. Goldstein (1995) extended some of the techniques to handle measurement errors in the independent variables in linear models to the multilevel model.

In the present paper, attention is focused on another way of handling response error in the independent variables in a multilevel model. The measurement error in the observed predictor variables is modeled by an item response theory (IRT) model. Modeling measurement error by an IRT model has several advantages. First, measurement error is defined conditionally on the value of the latent ability. That is, measurement error can be defined locally, for instance, as the posterior variance of the ability parameter given a response pattern. This local definition of measurement error results in heteroscedasticity: In the Rasch model, for instance, the posterior variance of the ability parameter given an extreme score is greater than the posterior variance of the ability parameter given an intermediate score (see, for instance, Hoijtink & Boomsma, 1995, pp. 59, Table 4.1). Second, the fact that reliability can be defined conditionally on the value of the latent variable, IRT can separates the influence of item difficulty and ability level, which supports the use of incomplete test administration designs, optimal test assembly, computer adaptive testing and test equating.

Besides IRT, another theme of this chapter will be Bayesian data analysis. The formulation of measurement-error problems in the framework of Bayesian analysis has recently been developed (Carroll et al., 1995;

Richardson, 1996; Zellner, 1971). It provides a natural way of taking into account all sources of uncertainty in the estimation of the parameters. Computing the posterior distributions involves high-dimensional numerical integration but these can be carried out straightforwardly by Gibbs sampling (Gelfand, Hills, Racine-Poon, & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 1995). Furthermore, the Bayesian formulation supports a straightforward model identification. That is, the model is identified in a natural way by fixing the latent ability scale, without needing prior knowledge about the variances of the measurement errors.

This chapter consists of eight sections. The next section presents a general multilevel model with covariates observed subject to error. In the following section, a classical test theory model and an item response theory model as measurement error models will be discussed. Then, a Markov Chain Monte Carlo (MCMC) estimation procedure will be described for estimating the parameters of a multilevel model with measurement error in covariates on both levels. In the following section, measurement error in correlated predictors will be discussed. Then, a small simulation study and some real-data examples will be given. The last section contains a discussion and suggestions for further research.

2. The Structural Multilevel Model

In social research, data structures often consist of observations measured at different levels. Examples of this nested structure include data from surveys where respondents are nested under an interviewer, test data of students within schools and data of multiple observations gathered over time. As an example, consider school effectiveness research, where interest is focused on the effects of school-variables on the educational achievement of the students. To evaluate school effectiveness, information is needed at both the level of students and the school-level. The heterogeneity in student and school characteristics requires a statistical model that takes the variation and relationships at each of the levels into account. Multilevel models support these requirements. A number of investigators have examined the issue of multilevel modeling of educational data (Bryk & Raudenbush, 1992; de Leeuw & Kreft, 1986; Goldstein, 1995; Raudenbush, 1988, Snijders & Bosker, 1999).

The hierarchical model is commonly used for continuous outcomes is a two-level formulation with Level 1 regression parameters multivariate normally distributed across Level 2 units. Suppose that students (Level 1), indexed ij $(i = 1, \ldots, n_j, j = 1, \ldots, J)$, are nested within schools (Level 2), indexed j $(j = 1, \ldots, J)$. In its general form, Level 1 consists of a regression model, for each of J nesting Level 2 groups $(j = 1, \ldots, J)$, in which the observations are modeled as a function of Q predictor vari-

ables $\Lambda_{1j}, \ldots, \Lambda_{Qj}$, that is,

$$y_{ij} = \beta_{0j} + \beta_{1j}\Lambda_{1ij} + \ldots + \beta_{qj}\Lambda_{qij} + \ldots + \beta_{Qj}\Lambda_{Qij} + e_{ij}, \qquad (4.1)$$

where $\mathbf{e}_j$ is an $(n_j \times 1)$ vector of normally distributed residuals with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_{n_j}$. The regression parameters are treated as outcomes in a Level 2 model given by

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}\Gamma_{1qj} + \ldots + \gamma_{qs}\Gamma_{sqj} + \ldots + \gamma_{qS}\Gamma_{Sqj} + u_{qj}, \qquad (4.2)$$

for $q = 0, \ldots, Q$, where the Level 2 error terms u_{qj}, $q = 0, \ldots, Q$, have a multivariate normal distribution with mean zero and covariance matrix $\mathbf{T}$,γ_{qs} and Γ_{sqj} are Level 2 regression coefficients (fixed effects) and predictor variables, respectively. Although the coefficients of all the predictors in the Level 1 model could be treated as random, it may be desirable to constrain the variation in one or more of the regression parameters to zero. This will be returned to below.

The explanatory variables at Level 1 comprise students' characteristics, such as, gender or age. Level 1 explanatory variables can also be latent, such as, socio-economic status, intelligence, community loyalty, social consciousness, managerial ability or willingness to adopt new practices. Explanatory variables as region, school-funding or gender are observed without an error. Latent variables can not be observed directly and have to be estimated, often with an error. Below, an example will be given of an analysis where students' abilities, regarding mathematics, are estimated as scores, on Level 1, obtained using an IQ test and, on Level 2, obtained using an adaptive instruction test taken by teachers. Both explanatory variables are measured with an error due to the limited number of items in the tests and the response variance. In predicting students' abilities, an increase in precision (i.e. reduction in σ^2) could be obtained by using student pretest scores as a covariate in the Level 1 model but errors in the predictor variables cause bias in estimated regression coefficients (Carroll et al., 1995, pp. 22).

The latent Level 1 covariates are denoted by $\boldsymbol{\theta}$ whereas the observed covariates without an error are denoted by Λ. Therefore, Level 1 of the structural model, formula (4.1), is reformulated as

$$y_{ij} = \beta_{0j} + \ldots + \beta_{qj}\theta_{qij} + \beta_{(q+1)j}\Lambda_{(q+1)ij} + \ldots + \beta_{Qj}\Lambda_{Qij} + e_{ij}, \quad (4.3)$$

where the first q predictors correspond to latent variables and the remaining $Q-q$ predictors correspond to observable variables. The regression coefficients are allowed to vary across Level 2 groups. This variation can be accounted for by treating the Level 1 regression coefficients as outcomes of Level 2 predictors. The explanatory variables at Level 2

consists of latent predictors denoted by $\boldsymbol{\zeta}$ and covariates observed without an error denoted by Γ. The Level 2 model in (4.2) is reformulated as

$$\beta_{qj} = \gamma_{q0} + \ldots + \gamma_{qs}\zeta_{sqj} + \gamma_{q(s+1)}\Gamma_{(s+1)qj} + \ldots + \gamma_{qS}\Gamma_{Sqj} + u_{qj}, \tag{4.4}$$

for $q = 0, \ldots, Q$, where the first s predictors correspond to latent variables and the remaining $S - s$ predictors to known fixed constants. The set of latent variables $\boldsymbol{\theta}$ is not observable but information about $\boldsymbol{\theta}$, denoted as $\mathbf{X}$, is available. $\mathbf{X}$ is called a surrogate for $\boldsymbol{\theta}$, that is, $\mathbf{X}$ has no information about $\mathbf{Y}$ other than what is available in $\boldsymbol{\theta}$. This is characteristic of nondifferential measurement error (Carroll et al., 1995, pp. 16-17). On Level 2, $\mathbf{W}$ is defined as a surrogate for $\boldsymbol{\zeta}$. The surrogates $\mathbf{X}$ and $\mathbf{W}$ are also called manifest variables or proxies. The effects from disregarding measurement error can range from biased parameter estimates to situations where real effects are hidden and signs of the estimated coefficients are reversed relative to the case with no measurement error (Carroll et al., 1995, pp 21-23).

3. Measurement Error Models

This section focuses on two parametric models for the response: the well-known classical true score model and the normal ogive model.

3.1 The Classical True Score Model

In the classical true score model (Lord & Novick, 1968), the individual's score on a particular test form, the observed score, is considered to be a chance variable with some, usually unknown, distribution. This distribution is generally known as the propensity distribution. The mean (expected value) of this distribution is interpreted as the true score. The error of measurement is the discrepancy between the observed scores and the true score. Since, by definition, the expected value of the observed scores is the true score, the expectation of the errors of measurement or error scores is zero. It is assumed that the corresponding true scores and error scores are uncorrelated and that error scores on different measurements are also uncorrelated. Denote X_{ijk} as the measurement associated with individual ij, let θ_{ij} be the mean of the response distribution and let ε_{ijk} the sampling deviation for the k-th response obtained from the k-th individual's response distribution, that is,

$$\varepsilon_{ijk} = X_{ijk} - \theta_{ij}. \tag{4.5}$$

The true score θ_{ij} of a person indexed ij is defined as the expected value of the observed score where the expectation is taken with respect to the

response distribution. This response distribution is hypothetical because in psychology and other subject areas it is usually not possible to obtain more than one independent observation. This model coincides mathematically with the classical additive measurement error model (Fuller, 1987, equation 1.1.2), where a normal distribution of the error variable is assumed. Let X_{ij} be the observed score of person ij, as the sum over item scores, given the responses to a set of items. Further on, the observed scores are modeled to handle measurement error and to estimate the true scores.

It is not strictly necessary to assume that the response distribution variances are equal for different persons. Some persons' responses may be measured more accurately than others. But error variances for individual examinees are usually subject to large sampling fluctuations. In the sequel, the group specific error variance is used as an approximation to the individual error variances of which it is the average. The group specific error variance is denoted as φ, where the group contains all examinees. This group specific error variance is the variance over the examinees of the errors of measurement, which is equal to the specific error variance averaged over the total number of examinees (Lord & Novick, 1968, pp. 155).

The classical true score model is based on assumptions that may not always be realistic. Measurement error is supposed to be independent of the predictor variables. Further, the variance of measurement errors is assumed to be equal conditional on different values of the dependent variable, say, the score level of the test taker in educational measurement. Another problem is that the reliability of measures is not easily assessed. The error variance could be estimated from repeated measurements to obtain an estimate of the error variance. However, besides the practical difficulties, it is not realistic to assume that the repeated measures are independent. To overcome these problems it is assumed that the variances and covariances of the measurement errors are known in advance, or suitable estimates exist (Goldstein, 1995, pp. 142). However, the estimates of the response variance are generally imprecise. In case of the usual maximum likelihood approach, the ratio of the error terms' variances or alternatively one or both of the variances ought to be known to identify the model (Fuller, 1987, pp. 9-11).

3.2 The Normal Ogive Model

For dichotomous items, the item response function (traceline, item characteristic curve) is the probability of a correct response as a function of ability. In this section, the normal ogive model is considered as a measurement error model (see Lord, 1980, pp. 27-41). The probabil-

ity of a correct response of a person indexed ij on an item indexed k $(k = 1, \ldots, K)$, $X_{ijk} = 1$, is given by

$$P\left(X_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right) = \Phi\left(a_k\theta_{ij} - b_k\right), \tag{4.6}$$

where Φ denotes the standard normal cumulative distribution function, and a_k and b_k are the discrimination and difficulty parameter of item k, respectively. The parameters of item k are also denoted by $\boldsymbol{\xi}_k = (a_k, b_k)$. An IRT model provides the frequency distribution of test scores for an examinee ij having a specified level θ_{ij} of ability or skill. The variance, $\sigma^2_{\mathbf{x}_{ij}|\theta_{ij}}$, of this conditional distribution of number right-score $\mathbf{X}_{ij}$ is

$$\begin{aligned}\sigma^2_{\mathbf{x}_{ij}|\theta_{ij}} &= \sum_{k=1}^{K} P\left(X_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right)\left[1 - P\left(X_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right)\right] \\ &= \sum_{k=1}^{K} \Phi\left(a_k\theta_{ij} - b_k\right)\Phi\left(b_k - a_k\theta_{ij}\right). \end{aligned} \tag{4.7}$$

Notice that this implies response variance given $\boldsymbol{\theta}$. The posterior distribution of θ_{ij} given $\mathbf{x}_{ij}$, $p\left(\theta_{ij} \mid \mathbf{x}_{ij}\right)$, is proportional to the distribution of $\mathbf{x}_{ij}$ given the ability level θ_{ij}, $p\left(\mathbf{x}_{ij} \mid \theta_{ij}\right)$, multiplied by the standard normal distribution. Therefore, the posterior variance of $p\left(\theta_{ij} \mid \mathbf{x}_{ij}\right)$ or local reliability, $\sigma^2_{\theta_{ij}|\mathbf{x}_{ij}}$, is closely related to response variance $\sigma^2_{\mathbf{x}_{ij}|\theta_{ij}}$, and this implies the possibility of heteroscedasticity. Furthermore, the measurement scale is independent of the items in the test. This is in contrast to classical test theory, where the true score depends on the items in the test and homoscedasticity is assumed.

4. An MCMC Estimation Procedure for a Multilevel Model with Measurement Error

The response error in the observed predictor variables of a structural multilevel model is modeled by an item response theory model and a classical true score model. The structural multilevel model combined with an IRT model is called a multilevel IRT model, and the structural multilevel model combined with a classical true score model is called a multilevel true score model. The estimation procedure for both models will be outlined concurrently.

Bayesian analysis of parametric models requires the specification of a likelihood and prior. Often a non-informative prior is used. The posterior distribution, derived from the joint density of the data and parameters according to Bayes formula, summarizes all of the information about the values of the parameters. Interest is focused on the expected a posteriori values of the parameters and posterior standard errors. In general,

complex models, such as the proposed multilevel model with measurement error in the covariates, require sophisticated numerical analytical methods to obtain estimates of the parameters of interest. However, Markov chain Monte Carlo algorithms (MCMC), in specific the Gibbs sampler, have proven potential for estimating complex models (Bernardo & Smith, 1994; Gelfand & Smith, 1990; Geman & Geman, 1984; Robert & Casella, 1999). Gibbs sampling succeeds because it reduces the problem of dealing simultaneously with missing data and a large number of related unknown parameters into a much simpler problem of dealing with one unknown quantity at a time by sampling each from its full conditional distribution. This sampling-based method is conceptually simple and easily implemented. The Gibbs sampler generates a Markov chain which converges in distribution to the joint posterior distribution of the parameters of interest (Tierney, 1994). That is, a Markov chain is constructed in such a way that its stationary distribution, also denoted limiting distribution, is the joint posterior distribution of the model parameters.

First, the implementation of the Gibbs sampler is considered for a multilevel model with a normal ogive model for the predictor variables. In this implementation the predictor variables are assumed to be uncorrelated. Second, the implementation of the Gibbs sampler is described with the classical true score model as measurement model. Correlated predictors with measurement error will be discussed in the next section.

4.1 Estimation using Gibbs Sampling

Evaluation of the model for the observed data is complicated by the fact that some elements are missing. The $\boldsymbol{\theta}$'s and $\boldsymbol{\zeta}$'s are treated as unobserved random parameters. Let $\boldsymbol{\theta}_{ij}$ be the first q explanatory variables on Level 1, which are latent, as in formula (4.3). The set of explanatory variables on Level 1 for predicting Y_{ij} is defined as $\Omega_{ij} = (\boldsymbol{\theta}_{ij}, \Lambda_{ij})$, where Λ_{ij} consists of the remaining $Q - q$ observable covariates on Level 1. Further, let $\boldsymbol{\zeta}_{qj}$ be the first s latent explanatory variables predicting β_{qj} on Level 2, as in formula (4.4). To complete the description of the covariates on Level 2, let $\Psi_{qj} = (\boldsymbol{\zeta}_{qj}, \Gamma_{qj})$ be the set of explanatory variables for β_{qj}, where Γ_{qj} are the remaining $S - s$ directly observable variables, also according to formula (4.4).

The MCMC algorithm is straightforwardly implemented by introducing a continuous latent variable that underlies each binary response. This approach follows the procedure of Albert (1992), which builds on the Data Augmentation algorithm of Tanner and Wong (1987), and has been extensively used in other missing data problems (see, for example, Béguin, 2000; Fox & Glas, 2001; Johnson & Albert, 1999, pp. 194-202;

Robert & Casella, 1999, pp. 414-438). Assume that the latent variables θ_{qij} are related to the observed responses X_{qijk} of person ij on an item k. This observation X_{qijk} indicates whether a continuous variable $Z^{(x)}_{qijk}$ with normal density is positive or negative. The superscript indicates the observed response variable $\mathbf{X}$. Further, $X_{qijk} = 1$ if $Z^{(x)}_{qijk} > 0$ and $X_{qijk} = 0$ otherwise. It follows that

$$\begin{aligned} p\left(z_{qijk} \mid \theta_{qij}, \boldsymbol{\xi}_k, x_{qijk}\right) \quad \propto \quad & f\left(z_{qijk}; a_k\theta_{qij} - b_k, 1\right)\left[I\left(z_{qijk} > 0\right)\right. \\ & \left. I\left(x_{qijk} = 1\right) + I\left(z_{qijk} \leq 0\right) I\left(x_{qijk} = 0\right)\right], \end{aligned}$$

where $f(.; a_k\theta_{qij} - b_k, 1)$ stands for the normal density with mean equal to $a_k\theta_{qij} - b_k$ and unit variance, and $I(.)$ is an indicator variable taking the value one if its argument is true, and zero otherwise. Further, θ_{qij} and $\boldsymbol{\xi}^{(x)}_k$ are the person and item parameters for person ij and item k, respectively. The matrix $\mathbf{Z}^{(x)}$ serves to simplify calculations and the value of $\mathbf{Z}^{(x)}$ does not affect the value of the estimator, that is, $\mathbf{Z}^{(x)}$ is only a useful device.

Let W_{sqjk} be a dichotomous response variable of a Level 2 unit, indexed j, on an item, indexed k, related to the s^{th} Level 2 latent variable, $\boldsymbol{\zeta}_{sqj}$, for predicting β_{qj}. For example, $\boldsymbol{\zeta}_{sqj}$ might be the pedagogical climate of school j measured using a questionnaire with dichotomously scored questions administered to a teacher or principal of school j. In the same way as for Level 1, complete data are formed; the augmented data will be denoted with $Z^{(w)}_{sqjk}$.

The Gibbs sampler arranges the sampling from one of the parameters conditionally on all other parameters in a number of steps. The entire procedure constitutes of stepwise draws from the conditional posterior distributions of the components $\mathbf{Z}^{(x)}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \mathbf{Z}^{(w)}, \boldsymbol{\xi}^{(w)}$ and $\boldsymbol{\zeta}$. The procedure consists of 10 steps:

1. Draw $\mathbf{Z}^{(x)}$ conditional on $\boldsymbol{\theta}, \boldsymbol{\xi}^{(x)}$ and $\mathbf{X}$.
2. Draw $\boldsymbol{\xi}^{(x)}$ conditional on $\boldsymbol{\theta}$ and $\mathbf{Z}^{(x)}$.
3. Draw $\boldsymbol{\theta}$ conditional on $\mathbf{Z}^{(x)}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\beta}, \sigma^2, \Omega$, and $\mathbf{Y}$.
4. Draw $\boldsymbol{\beta}$ conditional on $\Omega, \Psi, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}$ and $\mathbf{y}$.
5. Draw $\boldsymbol{\gamma}$ conditional on $\boldsymbol{\beta}, \Psi$ and $\mathbf{T}$.
6. Draw σ^2 conditional on $\boldsymbol{\beta}, \Omega$ and $\mathbf{y}$.
7. Draw $\mathbf{T}$ conditional on $\boldsymbol{\beta}, \Psi$ and $\boldsymbol{\gamma}$.
8. Draw $\mathbf{Z}^{(w)}$ conditional on $\boldsymbol{\zeta}, \boldsymbol{\xi}^{(w)}$ and $\mathbf{W}$.

9. Draw $\boldsymbol{\xi}^{(w)}$ conditional on $\boldsymbol{\zeta}$ and $\mathbf{Z}^{(w)}$.
10. Draw $\boldsymbol{\zeta}$ conditional on $\mathbf{Z}^{(w)}, \boldsymbol{\xi}^{(w)}, \boldsymbol{\beta}, \Psi$ and $\boldsymbol{\gamma}$.

Step 1-2. Sampling augmented data, $\mathbf{Z}^{(x)}$, and sampling the item parameters, $\boldsymbol{\xi}^{(x)}$, is described by Albert (1992) and Fox and Glas (2001).

Step 3. The variables, $\theta_{1ij}, \ldots, \theta_{qij}$, can be sampled individually because they are uncorrelated. Given $\mathbf{Z}^{(x)}_{qij}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\beta}_j$ and σ^2 are they independent and distributed as a mixture of normal distributions. That is, the augmented data $\mathbf{Z}^{(x)}_{qij}$ and the observed data Y_{ij} are normally distributed with, among others, parameter θ_{qij}, which is a priori normally distributed. The two-parameter normal ogive model is identified by fixing the origin and scale of the latent dimension. Therefore, the mean and variance of the ability distribution are fixed to zero and one, respectively. According to formula (4.3), the definition of the augmented data and the prior for θ_{qij} it follows that

$$p\left(\theta_{qij} \mid \mathbf{z}^{(x)}_{qij}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\beta}_j, \sigma^2, \Omega^-_{ij}, y_{ij}\right) \propto p\left(\mathbf{z}^{(x)}_{qij} \mid \theta_{qij}, \boldsymbol{\xi}^{(x)}\right) p\left(y_{ij} \mid \theta_{qij}, \boldsymbol{\beta}_j, \sigma^2, \Omega^-_{ij}\right) p(\theta_{qij}) \quad (4.8)$$

where Ω^-_{ij} are the set of explanatory variables for person ij on Level 1 without θ_{qij}. Split the regression coefficients on Level 1, $\boldsymbol{\beta}_j$, into β_{qj} and $\boldsymbol{\beta}^{(\Omega)}_j$, to distinguish the regression coefficient of explanatory variable θ_{qij} from the regression coefficients of the other explanatory variables Ω^-_{ij}. Formula (4.8) is the product of a normal model for the regression of $Z^{(x)}_{qijk} + b_k$ on a_k with θ_{qij} as a regression coefficient, a normal model for the regression of $Y_{ij} - \boldsymbol{\beta}^{(\Omega)}_j \Omega^-_{ij}$ on β_{qj} with θ_{qij} as a regression coefficient and a standard normal prior for θ_{qij}. Due to standard properties of normal distributions (e.g., see, Box & Tiao, 1973; Lindley & Smith, 1972) the fully conditional posterior density of θ_{qij} is also normally distributed,

$$\theta_{qij} \mid \mathbf{Z}^{(x)}_{qij}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\beta}_j, \sigma^2, \Omega^-_{ij}, Y_{ij} \sim N\left(\frac{\frac{\widehat{\theta}_{qij}}{v} + \frac{\widetilde{\theta}_{qij}}{\phi}}{\frac{1}{v} + \frac{1}{\phi} + 1}, \frac{1}{\frac{1}{v} + \frac{1}{\phi} + 1}\right), \quad (4.9)$$

where the posterior expectation constitutes of $\widehat{\theta}_{qij} = \left(\sum_{k=1}^K a_k^2\right)^{-1} \sum_{k=1}^K a_k \left(z_{qijk} + b_k\right)$, and $\widetilde{\theta}_{qij} = \beta_{qj}^{-1}\left(y_{ij} - \boldsymbol{\beta}^{(\Omega)}_j \Omega^-_{ij}\right)$, and the posterior variances of $v = \left(\sum_{k=1}^K a_k^2\right)^{-1}$ and $\phi = \beta_{qj}^{-2}\sigma^2$. The posterior expectation, formula (4.9), is the well-known composite or shrinkage estimator.

The estimate of θ_{qij} is a combination of two estimates, $\widehat{\theta}_{qij}$ and $\widetilde{\theta}_{qij}$, with the weights proportional to the precisions.

Step 4-7. The modification of the multilevel model to handle measurement error in the covariates causes minimal change in the complete conditional distributions of the parameters of the multilevel model, $\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T}\right)$. The full conditionals of the multilevel model parameters, required in the estimation procedure, can be found in Fox and Glas (2001) and Seltzer (1993) and Seltzer, Wong, & Bryk (1996).

Step 8-9. Measurement error in the predictor variables on Level 2 are treated in the same way as on Level 1, with a normal ogive model as measurement model. Therefore, augmented data denoted as $\mathbf{Z}^{(w)}$, in relation to the observed data $\mathbf{W}$, item parameters $\boldsymbol{\xi}^{(w)}$ and $\boldsymbol{\zeta}$ have to be sampled. An adapted complete conditional of $\mathbf{Z}^{(w)}$ given $\boldsymbol{\zeta}, \boldsymbol{\xi}_k^{(w)}$can be found in Albert (1992) and Fox and Glas (2001). Also an adapted complete conditional distribution of the item parameters can be found therein.

Step 10. Split the regression coefficients $\boldsymbol{\gamma}_q$ on Level 2 in γ_{qs} and $\boldsymbol{\gamma}_q^{(\Psi)}$, relating to the predictor ζ_{sqj} and remaining Level 2 covariates Ψ_{qj}^-, respectively, where Ψ_{qj}^- is the set of explanatory variables for β_{qj} on Level 2 without ζ_{sqj}. The latent predictor variables $\zeta_{1qj}, \ldots, \zeta_{sqj}$ can be sampled individually, because they are independent. The Level 2 model, formula (4.4), is reformulated as,

$$\beta_{qj} - \boldsymbol{\gamma}_q^{(\Psi)} \Psi_{qj}^- = \gamma_{qs} \zeta_{sqj} + u_{qj}, \tag{4.10}$$

where $u_{qj} \sim N\left(0, \tau_{qq}^2\right)$ and τ_{qq}^2 is the q^{th} diagonal element of $\mathbf{T}$. From formula (4.10) the least squares estimator $\widetilde{\zeta}_{sqj} = \gamma_{qs}^{-1}\left(\beta_{qj} - \boldsymbol{\gamma}_q^{(\Psi)} \Psi_{qj}^-\right)$ can be obtained. The parameters ζ_{sqj} given augmented data $\mathbf{Z}_{sqj}^{(w)}$ and parameters $\boldsymbol{\xi}^{(w)}, \beta_{qj}, \Psi_{qj}^-$ and $\boldsymbol{\gamma}_q$ are independent and distributed as a mixture of normal distributions. That is, augmented data, $\mathbf{Z}_{sqj}^{(w)}$, and regression coefficient, β_{qj}, are normally distributed with, among others, parameter ζ_{sqj} which is a priori normally distributed. Therefore, it follows that

$$\begin{aligned} p\left(\zeta_{sqj} \mid \mathbf{z}_{sqj}^{(w)}, \boldsymbol{\xi}^{(w)}, \beta_{qj}, \Psi_{qj}^-, \boldsymbol{\gamma}_q\right) \propto\ & p\left(\mathbf{z}_{sqj}^{(w)} \mid \zeta_{sqj}, \boldsymbol{\xi}^{(w)}\right) \\ & p\left(\beta_{qj} \mid \zeta_{sqj}, \Psi_{qj}^-, \boldsymbol{\gamma}_q\right) p(\zeta_{sqj}). \end{aligned} \tag{4.11}$$

For identification of the model the prior for ζ_{sqj} is the standard normal distribution. Hence, the fully conditional posterior density of ζ_{sqj} is

given by

$$\zeta_{sqj} \mid \mathbf{Z}^{(w)}_{sqj}, \boldsymbol{\xi}^{(w)}, \beta_{qj}, \Psi^{-}_{qj}, \boldsymbol{\gamma}_q \sim N\left(\frac{\frac{\widehat{\zeta}_{sqj}}{\kappa} + \frac{\widetilde{\zeta}_{sqj}}{\psi}}{\frac{1}{\kappa} + \frac{1}{\psi} + 1}, \frac{1}{\frac{1}{\kappa} + \frac{1}{\psi} + 1}\right), \quad (4.12)$$

where $\widehat{\zeta}_{sqj}$ is the least squares estimator following from the regression of $z^{(w)}_{sqjk} + b'_k$ on a'_k and κ the variance of $\widehat{\zeta}_{sqj}$, as in Step 3. The item parameters $\boldsymbol{\xi}^{(w)}_k = (a'_k, b'_k)$ are sampled in Step 9. Finally, $\widetilde{\zeta}_{sqj}$ is the least squares estimator for ζ_{sqj}, formula (4.10) , with variance $\psi = 1/\gamma^2_{qs}$.

This implementation of the Gibbs sampler is easily changed into a procedure for estimating the parameters of the structural (multilevel) model with the classical true score model as measurement error model. It is assumed that the variance structure, φ, is known and given by formula (4.5) . This is also necessary for identification of the model. The surrogates $\mathbf{X}$ and $\mathbf{W}$ provide a sum score or observed score X_{ij} of the examinee indexed ij on Level 1 and a sum score, W_j, observed in school j. Thus, in this case the classical true score model, instead of the normal ogive model, is used as measurement error model on Level 1 and Level 2. Augmented data and item parameters do not have to be sampled. Therefore, Step 1, 2, 8 and Step 9 can be left out. Step 3 and Step 10 changes into the following two steps.

Step 3′. Let X_{qij} denote the observed score of a person, indexed ij, in relation to θ_{qij}, the q^{th} latent covariate on Level 1 in predicting Y_{ij}. Again, the latent predictors on Level 1 can be sampled separately because they are independent. Further, X_{qij} is a random variable taking on values from independent repeated measurements, which is normally distributed with mean θ_{qij} and variance φ. The complete conditional of θ_{qij} follows from the regression of X_{qij} on θ_{qij} and the regression of Y_{ij} on Ω_{ij}, formula (4.3) . It follows that

$$p\left(\theta_{qij} \mid \Omega^{-}_{ij}, \boldsymbol{\beta}_j, \sigma^2, \varphi, x_{qij}, y_{ij}\right) \propto p\left(x_{qij} \mid \theta_{qij}, \varphi\right) p\left(y_{ij} \mid \theta_{qij}, \Omega^{-}_{ij}, \boldsymbol{\beta}_j, \sigma^2\right)$$

The prior information for θ_{qij} is incorporated into the measurement error model, where the distribution and variance structure of the true score is determined. It follows that the fully conditional posterior density of θ_{qij} is given by

$$\theta_{qij} \mid \Omega^{-}_{ij}, \boldsymbol{\beta}_j, \sigma^2, \varphi, X_{qij}, Y_{ij} \sim N\left(\frac{\frac{x_{qij}}{\varphi} + \frac{\widetilde{\theta}_{qij}}{\phi}}{\frac{1}{\varphi} + \frac{1}{\phi}}, \frac{1}{\frac{1}{\varphi} + \frac{1}{\phi}}\right), \quad (4.13)$$

with $\widetilde{\theta}_{ij}$ and ϕ as in formula (4.9) .

The classical true score model can also be used for modeling the measurement error in the predictor variables on Level 2. Let ζ_{sqj} be the expected value of the observed score, W_{sqj}, where the expectation is taken with respect to the normal distribution, the assumed response distribution. Further, define κ as the variance, a priori known, over parallel observations of W_{sqj}. It follows that ζ_{sqj} can be sampled in the same way as in Step 3′. That is, Step 10′, draw ζ_{sqj} conditional on $W_{sqj}, \kappa, \beta_{qj}, \Psi_{qj}^{-}$ and γ_q.

In formula (4.3) it is assumed that every regression coefficient varies across Level 2 groups. In certain applications, it can be desirable to constrain the effect of one or more of the Level 1 predictors to be identical across Level 2 units. An implementation of the Gibbs sampler, where regression coefficients are treated as non-varying across Level 2 groups, needs a further division of regression components. This calls for a division in regression coefficients related to observed predictors and latent predictors, with a further subdivision of both parts into components treated as random and components treated as non-random across Level 2 groups. Finally, the complete conditional distribution of each subset, given the other parameters and the data, must be specified (see, for example, Seltzer et al., 1996).

The presented 10 steps define the Gibbs sampler for estimation of the parameters of the multilevel model with measurement error in the predictor variables, where the normal ogive model or the classical true score model is used as measurement error model. With initial values for the parameters, the Gibbs sampler repeatedly samples from the full conditional distributions with systematic scan, that is, the sampler updates the components in the natural ordering. A different strategy of updating the components can affect the speed of convergence (Roberts & Sahu, 1997). The values of the initial parameters are important for the rate of convergence. Initial estimates can be obtained by estimating the normal ogive model using Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), subsequently, the parameters of the multilevel model can be estimated with HLM for Windows (Raudenbush, Bryk, Cheong, & Congdon, 2000) given the parameter estimates of the normal ogive model.

Convergence can be evaluated by comparing the between and within variance of generated multiple Markov chains from different starting points (see, for instance, Robert & Casella, 1999, pp. 366). Another method is to generate a single Markov chain and to evaluate convergence by dividing the chain into subchains and comparing the between- and within-sub-chain variance. A single run is less wasteful in the num-

ber of iterations needed. Additionally, a unique chain and a slow rate of convergence is more likely to get closer to the stationary distribution than several shorter chains. In the example given below, the full Gibbs sample was used in estimating all parameters instead of subsampling from this sample. The latter procedure leads to losses in efficiency (MacEachern & Berliner, 1994). Finally, after the Gibbs sampler has reached convergence and "enough" samples are drawn, posterior means of all parameters of interest are estimated with the mixture estimator to reduce the sampling error attributable to the Gibbs sampler (Liu, Wong, & Kong, 1994). The posterior standard deviations and highest posterior density intervals can be estimated from the sampled values obtained from the Gibbs sampler (Chen & Shao, 1999).

5. Measurement Error in Correlated Predictor Variables

In this section, measurement error in explanatory variables on Level 1 will be modeled by an IRT model for the item responses related to these explanatory variables. Because it is often not realistic to assume that the predictor variables are independent, a multivariate IRT model will be used as measurement error model. The same procedure can be applied to measurement errors in correlated explanatory variables on Level 2. It is assumed that there exists a manifest variable for every unobserved predictor variable and every manifest variable consists of a set of item responses.

Assume that the latent variables θ_{qij} are related to observable variables $\mathbf{X}_{qij}$, $(q = 1, \ldots, Q)$ via a normal ogive IRT measurement model. Let $\mathbf{X}_{qij} = \left(X_{qij1}, \ldots, X_{qijK_q}\right)^t$, with realization $\left(x_{qij1}, \ldots, x_{qijK_q}\right)^t$, denote a response vector on a test with K_q items. Before the actual parameters $\boldsymbol{\theta}$ will be identified, consider a parametrization $\boldsymbol{\theta}^*$. Let $\boldsymbol{\theta}^*_{ij}$ be the vector of latent predictor variables for a person indexed ij, that is, $\boldsymbol{\theta}^*_{ij}$ has elements θ^*_{qij}. Further, suppose that for every predictor a two-parameter compensatory normal ogive model holds, that is, $P\left(X_{qijk} = 1 \mid \theta^*_{qij}, a^*_{qk}, b^*_{qk}\right) = \Phi\left(a^*_{qk}\theta^*_{qij} - b^*_{qk}\right)$, where a^*_{qk} and b^*_{qk} are item parameters of an item of predictor q. Because the predictor variables θ^*_{qij} are considered dependent, it will be assumed that $\boldsymbol{\theta}^*_{ij}$ has a multivariate normal distribution with mean zero and covariance matrix Σ^*. However, the parametrization $\boldsymbol{\theta}^*$ can be transformed such that $\boldsymbol{\theta}$ has a multivariate normal distribution with mean zero and covariance matrix $\mathbf{I}$, that is, the variables θ_{qij} become independent. Under this transformation, the normal ogive model transforms to

$$P\left(X_{qijk} = 1 \mid \boldsymbol{\theta}_{ij}, \mathbf{a}_{qk}, b_{qk}\right) = \Phi\left(\mathbf{a}^t_{qk}\boldsymbol{\theta}_{ij} - b_{qk}\right),$$

where $\mathbf{a}_{qk}$ is a vector of discrimination-parameters or factor loadings (see, for instance, McDonald, 1967, 1982, 1997). Notice that every item response now depends on all latent dimensions. This gives rise to the following procedure.

Analogous with the above procedure, see Step 1 to 3 above, a random vector $\mathbf{Z}_{ij} = \left(Z_{1ij1}, \ldots, Z_{QijK_Q}\right)^t$ is introduced, where $Z_{qijk} \sim N\left(\mathbf{a}_{qk}^t \boldsymbol{\theta}_{ij} - b_{qk}, 1\right)$, and it is supposed that $X_{qijk} = 1$ when $Z_{qijk} > 0$ and $X_{qijk} = 0$ otherwise. After deriving the fully conditional distributions, the Gibbs sampler can again be used to estimate the posterior distributions of all parameters.

Step 1: Sampling $\mathbf{Z}$. Given the parameters $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\xi}_{qk}$, the variables Z_{qijk} are independent and

$$Z_{qijk} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\xi}_{qk}, X_{qijk} \sim N\left(\mathbf{a}_{qk}^t \boldsymbol{\theta}_{ij} - b_{qk}, 1\right), \tag{4.14}$$

truncated at the left by 0 if $X_{qijk} = 1$ and truncated at the right by 0 if $X_{qijk} = 0$.

Step 2: Sampling $\boldsymbol{\theta}_{ij}$. Let $\boldsymbol{\theta}_{ij}$ be the vector with Q predictor variables for a person indexed ij. These are the regression coefficients in the normal linear model

$$\mathbf{Z}_{ij} + \mathbf{b} = \mathbf{A}\boldsymbol{\theta}_{ij} + \boldsymbol{\varepsilon}_{ij},$$

where $\mathbf{b} = \left(b_{11}, \ldots, b_{1K_1}, b_{21}, \ldots, b_{QK_Q}\right)^t$, $\boldsymbol{\theta}_{ij} = (\theta_{1ij}, \ldots, \theta_{Qij})^t$ and $\mathbf{A}$ is a $\left(\sum_q K_q \times Q\right)$ matrix with row vectors $\mathbf{a}_{qk}^t$, for items $k = 1, \ldots, K_q$ and predictors $q = 1, \ldots, Q$. Furthermore, the vector $\boldsymbol{\varepsilon}_{ij}$ has elements ε_{qijk}, which are independent and standard normally distributed. It is assumed that all Level 1 predictors are unobserved and their regression coefficients are treated as varying across Level 2 groups. For identification of the model, $\boldsymbol{\theta}_{ij}$ has a multivariate standard normal prior, and it follows that

$$p\left(\boldsymbol{\theta}_{ij} \mid \mathbf{z}_{ij}, y_{ij}, \boldsymbol{\xi}_{qk}, \boldsymbol{\beta}_j, \sigma^2\right) \propto p\left(\mathbf{z}_{ij} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\xi}_{qk}\right) p\left(y_{ij} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2\right) f\left(\boldsymbol{\theta}_{ij}; \mathbf{0}, \mathbf{I}_Q\right).$$

As in the unidimensional case, the mixture of multivariate normal distributions results in a multivariate normal distribution with a shrinkage estimator as expectation,

$$\boldsymbol{\theta}_{ij} \mid \mathbf{Z}_{ij}, Y_{ij}, \boldsymbol{\xi}_{qk}, \boldsymbol{\beta}_j, \sigma^2 \sim \mathbf{N}\left(\frac{\Upsilon^{-1}\widehat{\boldsymbol{\theta}}_{ij} + \Phi^{-1}\widetilde{\boldsymbol{\theta}}_{ij}}{\Upsilon^{-1} + \Phi^{-1} + \mathbf{I}_Q}, \left(\Upsilon^{-1} + \Phi^{-1} + \mathbf{I}_Q\right)^{-1}\right), \tag{4.15}$$

where $\widehat{\boldsymbol{\theta}}_{ij} = \left(\mathbf{A}^t\mathbf{A}\right)^{-1}\mathbf{A}^t(\mathbf{z}_{ij}+\mathbf{b})$ and $\widetilde{\boldsymbol{\theta}}_{ij} = \left(\boldsymbol{\beta}_{-j}^t\boldsymbol{\beta}_{-j}\right)^{-1}\boldsymbol{\beta}_{-j}^t\left(y_{ij}-\beta_{0j}\right)$, with $\boldsymbol{\beta}_{-j} = \left(\beta_{1j},\ldots,\beta_{Qj}\right)$ and the corresponding variances are $\Upsilon = \left(\mathbf{A}^t\mathbf{A}\right)^{-1}$ and $\Phi = \sigma^2\left(\boldsymbol{\beta}_{-j}^t\boldsymbol{\beta}_{-j}\right)^{-1}$.

Step 3: Sampling $\boldsymbol{\xi}_{qk}$. Let $\boldsymbol{\xi}_{qk} = (\mathbf{a}_{qk}, b_{qk})^t$, $k = 1,\ldots,K_q$ and $q = 1,\ldots,Q$, which represent the item-parameters of item k of a test relating to predictor q. Further, define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_Q)$ with $\boldsymbol{\theta}_q = (\theta_{q11},\ldots,\theta_{qn_JJ})^t$. Given $\boldsymbol{\theta}$, the $\mathbf{Z}_{qk} = (Z_{q11k},\ldots,Z_{qn_JJk})^t$ satisfy the linear model

$$\mathbf{Z}_{qk} = \begin{bmatrix} \boldsymbol{\theta} & -\mathbf{1} \end{bmatrix}\boldsymbol{\xi}_{qk} + \boldsymbol{\varepsilon}_{qk} \tag{4.16}$$

where $\boldsymbol{\varepsilon}_{qk} = (\varepsilon_{q11k},\ldots,\varepsilon_{qn_JJk})^t$ are standard normally distributed. Combining the prior for $p\left(\boldsymbol{\xi}_{qk}\right) = \prod_{q=1}^{Q} I\left(\mathbf{a}_{qk} > \mathbf{0}\right)$ with equation (4.16) gives

$$\boldsymbol{\xi}_{qk} \mid \boldsymbol{\theta}, \mathbf{Z}_{qk} \sim N\left(\widehat{\boldsymbol{\xi}}_{qk}, \left(\mathbf{H}^t\mathbf{H}\right)^{-1}\right)\prod_{q=0}^{Q} I\left(\mathbf{a}_{qk} > \mathbf{0}\right),$$

where $\mathbf{H} = \begin{bmatrix} \boldsymbol{\theta} & -\mathbf{1} \end{bmatrix}$ and $\widehat{\boldsymbol{\xi}}_{qk}$ is the least squares estimator based on (4.16).

Again, this procedure could be extended to handle observed and non-observed explanatory variables with regression coefficients varying or fixed across Level 2 units. Notice that the steps for sampling the other parameters of the structural model, described in the previous section, remain the same. Modeling measurement error in correlated predictor variables with the classical true score model needs a lot of prior information. The group specific error variance regarding all tests has to be known a priori, that is, the covariance matrix of Q explanatory variables of person ij has to be known in advance. The covariance matrix of the correlated latent predictor variables also identifies the model, in case of the classical true score model as measurement error model. Then, the conditional distribution of $\boldsymbol{\theta}_{ij}$ becomes

$$\boldsymbol{\theta}_{ij} \mid \mathbf{X}_{ij}, Y_{ij}, \boldsymbol{\beta}_j, \sigma^2, \Upsilon \sim \mathbf{N}\left(\frac{\Upsilon^{-1}\mathbf{x}_{ij} + \Phi^{-1}\widetilde{\boldsymbol{\theta}}_{ij}}{\Upsilon^{-1}+\Phi^{-1}}, \left(\Upsilon^{-1}+\Phi^{-1}\right)^{-1}\right),$$

where $\mathbf{x}_{ij} = (x_{1ij},\ldots,x_{Qij})$ and x_{qij} is the sum score of person ij on a test related to predictor q. Further, Υ is the a priori known covariance matrix of the sum scores of person ij. In most cases, the covariance matrix is population dependent and fixed over persons taking the tests to get a reliable estimate.

6. A Simulation Study

In this section, a numerical example was analyzed to illustrate parameter recovery with the Gibbs sampler. Data were simulated using a multilevel model with two latent predictors. The model is given by

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}\theta_{1ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_{10j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j}, \end{aligned} \tag{4.17}$$

where $e_{ij} \sim N\left(0, \sigma^2\right)$ and $\mathbf{u}_j \sim N\left(0, \mathbf{T}\right)$. Furthermore, it was assumed that the observed variables $\mathbf{X}$ and $\mathbf{W}$ were related to the latent predictors $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ through a normal ogive model. Response patterns $\mathbf{X}$ and $\mathbf{W}$ were generated according to a normal ogive model for a test of 20 items and a test of 40 items, respectively. For the test relating to the latent covariate $\boldsymbol{\theta}$ at Level 1, 4,000 response patterns were generated which were divided over $J = 200$ groups of 20 students each. The generating values of the item parameters are shown under the label Generated in Table 4.1. Accordingly, for the test relating to the latent covariate $\boldsymbol{\zeta}$ at Level 2, 200 response patterns were generated. The true values of the fixed and random effects, $\boldsymbol{\gamma}, \sigma^2$ and $\mathbf{T}$, are shown under the label Generated in Table 4.2.

The normal ogive models were estimated with Bilog-MG. Next, the initial parameters of the multilevel model were computed with HLM given the parameters of the normal ogive models. In the simulation study, 500 iterations were needed to estimate the measurement error models and another 500 iterations were needed to compute the parameters of the multilevel model. Subsequently, 20,000 iterations were made to estimate the parameters of the multilevel IRT model. The convergence of the Gibbs sampler was checked by examining the plots of sampled parameter values. It was concluded that a burn-in period of 1,000 iterations was sufficient. The location of the unobserved predictors can be fixed by transforming each sample during the Gibbs sampling process. Grand mean or group-mean centering of an unobserved explanatory variable is obtained by subtracting the grand mean or group-means from each sample drawn in each step of the Gibbs sampler. The model was identified by fixing the scale of the latent variables to the true scale of the generated latent variables. This way, the estimated parameters were directly comparable to the true parameter values. The model could also be identified by restricting the sum of the difficulty parameters to zero and the product of the discrimination parameters to one. Accordingly, the estimated parameters should be rescaled to compare them to the

Table 4.1. Item parameter estimates of the normal ogive IRT model at Level 1.

	Generated		Gibbs Sampler					
Item	a_k	b_k	a_k	s.d.	HPD	b_k	s.d.	HPD
1	.821	−.180	.792	.029	[.736, .852]	−.165	.022	[−.208, −.122]
2	1.058	−.418	.981	.035	[.913, 1.051]	−.399	.024	[−.449, −.353]
3	1.810	.366	1.810	.062	[1.685, 1.929]	.482	.030	[.422, .539]
4	1.690	−.254	1.645	.053	[1.541, 1.751]	−.128	.028	[−.182, −.074]
5	.804	−.096	.777	.029	[.720, .835]	−.108	.022	[−.152, −.066]
6	1.409	.550	1.441	.048	[1.346, 1.537]	.653	.029	[.595, .710]
7	1.461	.077	1.470	.049	[1.376, 1.569]	.119	.026	[.067, .170]
8	.932	1.026	.954	.036	[.886, 1.025]	1.093	.031	[1.030, 1.152]
9	.599	.675	.623	.027	[.569, .676]	.748	.024	[.702, .795]
10	1.788	.091	1.764	.061	[1.646, 1.885]	.216	.028	[.160, .270]
11	.777	−.455	.710	.030	[.651, .768]	−.474	.023	[−.520, −.430]
12	.299	.980	.319	.025	[.270, .367]	.991	.025	[.942, 1.039]
13	.829	.594	.833	.031	[.773, .891]	.630	.024	[.582, .678]
14	2.806	−1.024	2.836	.130	[2.594, 3.107]	−.964	.055	[−1.079, −.862]
15	.876	−.287	.818	.030	[.758, .876]	−.266	.023	[−.311, −.222]
16	1.814	.093	1.743	.059	[1.627, 1.856]	.177	.028	[.121, .230]
17	.773	.184	.734	.028	[.679, .788]	.194	.022	[.152, .236]
18	1.539	.648	1.500	.051	[1.400, 1.597]	.692	.030	[.634, .751]
19	1.166	.141	1.064	.035	[.994, 1.132]	.201	.024	[.153, .246]
20	.994	.425	1.012	.034	[.946, 1.097]	.469	.024	[.421, .516]

true parameter values. To illustrate the parameter recovery, the model was identified such that the parameters were directly comparable.

In case of the multilevel true score model 500 iterations were used as a burn-in period and another 20,000 iterations were used to compute the parameters. Initial values of the multilevel parameters were obtained by HLM using the observed scores as explanatory variables. The model was identified by specifying the group specific error variances in advance. The group specific error variances, relating to test **X** and **W**, denoted as φ_1 and φ_2, were .156 and .089, respectively. The estimates of the group specific error variances were obtained by averaging the unbiased estimates for the error variances of individual examinees (Lord & Novick, 1968, pp. 155). The scale of the latent variables was fixed to the true scale of the generated latent variables. The latent variables in the different models were equally scaled, therefore, the generated parameters, the estimated parameters of the multilevel IRT model, and the estimated parameters of the multilevel true score model were directly comparable.

Table 4.2. Parameter estimates of the multilevel model with measurement error in the covariates.

	Generated	IRT Model			Classical True Score Model $\varphi_1 = .156, \varphi_2 = .089$		
Fixed Effects	Coeff.	Coeff.	s.d.	HPD	Coeff.	s.d.	HPD
γ_{00}	2	2.010	.041	$[1.932, 2.085]$	2.011	.042	$[1.927, 2.091]$
γ_{01}	1	.970	.031	$[.907, 1.029]$	.967	.031	$[.902, 1.028]$
γ_{10}	1	.928	.036	$[.857, .997]$	.936	.037	$[.864, 1.007]$
Random Effects	Var. Comp.	Var. Comp.	s.d.	HPD	Var. Comp.	s.d.	HPD
σ^2	.5	.487	.014	$[.461, .515]$	.472	.014	$[.445, .499]$
τ_0^2	.2	.311	.040	$[.237, .391]$	.332	.045	$[.249, .422]$
τ_1^2	.2	.221	.027	$[.170, .273]$	.238	.028	$[.185, .293]$
τ_{01}^2	.1	.202	.027	$[.151, .256]$	.221	.030	$[.165, .281]$
		$E\left[L^2\right]$	s.d.		$E\left[L^2\right]$	s.d.	
		.625	.014		.747	.018	

In Table 4.1, the estimates of the item parameters resulting from the Gibbs sampler, associated with the measurement error model for $\boldsymbol{\theta}$, are given under the label Gibbs Sampler. The reported standard deviations are the posterior standard deviations. Highest posterior density intervals were calculated as confidence regions for the parameters and they are given in the column labeled HPD. These highest posterior density intervals are the 95%-intervals. Most of the true parameter values were well within the computed intervals. The estimates of the item parameters, from the test relating to $\boldsymbol{\zeta}$, and the true parameter values were also quite close but contained larger standard deviations due to the small number of groups.

Table 4.2 presents the results of estimating the parameters of the multilevel model. The estimates of the fixed and random effects using the classical true score model are given under the label Classical True Score Model. The estimates of the fixed and random effects using the normal ogive model are given under the label IRT Model. It was remarkable that in both models the parameter estimates of the variances on Level 2 and covariance between the Level 2 residuals were too high. The

parameter estimates of the random coefficients using the classical true score model differed more from the true parameter values. By estimating the random coefficients given the true generated latent variables, $(\boldsymbol{\theta}, \boldsymbol{\zeta})$, the estimates of the random coefficients τ_0^2, τ_1^2 and τ_{01}^2 were .280, .205 and .205, respectively. It was verified that the estimated parameters obtained using the classical true score model, instead of the normal ogive model, differed more from these parameter estimates.

The models were compared using posterior predictive data, $\mathbf{Y}^{rep}, \mathbf{X}^{rep}$, and $\mathbf{W}^{rep}$ under the different models (Carlin & Louis, 1996; Gelman et al., 1995; Gelman, Meng, & Stern, 1996). Let $\mathbf{Y}^{rep}$ denote replicate observations, given the underlying model parameters. Analogously, let $\mathbf{X}^{rep}$ and $\mathbf{W}^{rep}$ denote replicated observations, given $\mathbf{X}$ and $\mathbf{W}$, respectively, and given the underlying model parameters.

Define L_{1j} as the distance from $\mathbf{Y}_j^{rep}$ to $\mathbf{Y}_j$ given model M and data $(\mathbf{X}_j, \mathbf{W}_j)$, so

$$E\left[L_{1j}^2 \mid M, \mathbf{y}_j\right] = \int\int\int \prod_{i|j} \left(y_{ij} - y_{ij}^{rep}\right)^2 p\left(y_{ij}^{rep} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2\right) p\left(\boldsymbol{\theta}_{ij}, \sigma^2 \mid \mathbf{x}_{ij}, \mathbf{y}\right) dy_{ij}^{rep} d\boldsymbol{\theta}_{ij} d\sigma^2. \quad (4.18)$$

Aggregating over schools results in

$$\begin{aligned} E\left[L_1^2 \mid M, \mathbf{y}\right] &= E\left[(\mathbf{y} - \mathbf{y}^{rep})^2 \mid M, \mathbf{y}\right] \\ &= \prod_j \int\int E\left[L_{1j}^2 \mid M, \mathbf{y}_j\right] p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\zeta}_j, \mathbf{y}_j\right) p\left(\boldsymbol{\zeta}_j \mid \mathbf{w}_j, \mathbf{y}_j\right) d\boldsymbol{\beta}_j d\boldsymbol{\zeta}_j, \end{aligned} \quad (4.19)$$

where $p\left(y_{ij}^{rep} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2\right)$ is the probability of replicated data given the parameters, $p\left(\boldsymbol{\theta}_{ij}, \sigma^2 \mid \mathbf{x}_{ij}, \mathbf{y}\right)$ and $p\left(\boldsymbol{\zeta}_j \mid \mathbf{w}_j, \mathbf{y}_j\right)$ are the joint posterior density of the unobserved explanatory variables and variance at Level 1 and the posterior density of the unobserved explanatory variables at Level 2, respectively. In the same way, define L_2 as the distance from $\mathbf{X}^{rep}$ to $\mathbf{X}$ given model M and data $(\mathbf{Y}_j, \mathbf{W}_j)$. This results in

$$E\left[L_2^2 \mid M, \mathbf{x}\right] = E\left[(\mathbf{x}. - \mathbf{x}.^{rep})^2 \mid M, \mathbf{x}\right], \quad (4.20)$$

where $\mathbf{x}.$ and $\mathbf{x}.^{rep}$ denote the observed sum scores and the replicated sum scores, respectively. Accordingly, let L_3 be the distance from $\mathbf{W}^{rep}$ to $\mathbf{W}$ given model M and data $(\mathbf{Y}_j, \mathbf{X}_j)$, leading to the statistic

$$E\left[L_3^2 \mid M, \mathbf{w}\right] = E\left[(\mathbf{w}. - \mathbf{w}.^{rep})^2 \mid M, \mathbf{w}\right], \quad (4.21)$$

where $\mathbf{w}.$ denotes the observed sum scores and $\mathbf{w}.^{rep}$ denotes the replicated sum scores. Each statistic summarizes the information concerning the predictive data given the observed data. Besides, each statistic is the sum of the variance of the replicated data plus the square of the bias of the replicated data with respect to the observed data. Together these three predictive criterions given some model M reflect the quality of prediction of a replicate of the observed data. It is a natural way to evaluate model performance by comparing what it predict with what has been observed (Bernardo & Smith, 1994, pp. 397). If the model fits, $E\left[L_1^2 \mid M, \mathbf{y}\right]$, $E\left[L_2^2 \mid M, \mathbf{x}\right]$, and $E\left[L_3^2 \mid M, \mathbf{w}\right]$ should be small. The sum of the three statistics, formula (4.19), (4.20), and (4.21), summarizes the information concerning the general fit of the model. It will be denoted as

$$E\left[L^2 \mid M, \mathbf{y}, \mathbf{x}, \mathbf{w}\right]. \tag{4.22}$$

In Table 4.2 the values of $E\left[L^2 \mid M, \mathbf{y}, \mathbf{x}, \mathbf{w}\right]$ per Level 1 unit and corresponding posterior standard deviations of both models are given. The smaller value for the multilevel IRT model indicated that this model predicted the observed data better than the multilevel true score model. Figure 4.1 presents $20,000$ values of the statistic computed at every iteration of the Gibbs sampler. It can be seen that the statistic indicated a preference for the IRT model. Besides, the plot also illustrates the good convergence of the Gibbs sampler.

Figure 4.2 presents the distributions of the generated data and the posterior predictive distributions of the data using the multilevel true score model and the multilevel IRT model. The top figure shows that both models predict the dependent data $\mathbf{y}$ very well. The multilevel true score model is less restrictive than the multilevel IRT model and this resulted in a slightly better prediction. But multilevel IRT entails a more realistic way of modeling the independent observed data $(\mathbf{x}, \mathbf{w})$. This can be seen in the middle and bottom figure. Another important point is that the normal ogive model predicted the skewed distribution of the test scores $\mathbf{x}$ at Level 1 better. The classical true score model discriminated less between students' outcomes because it is based on sum scores instead of complete response patterns. Further, the variance in the explanatory variable $\boldsymbol{\theta}$ was suppressed by a "ceiling" effect in the observed sum scores in the classical true score model. The more flexible multilevel true score model resulted in a better prediction of $\mathbf{y}$, but the price to pay was a far less precise prediction of $\mathbf{x}$ and $\mathbf{w}$. In general, the multilevel IRT model predicted all observed data very well, and this resulted in a much better fit to the data.

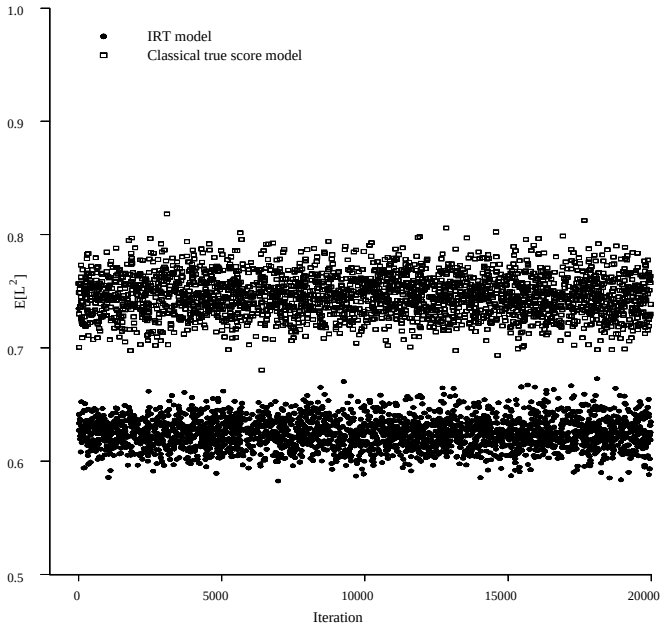

Figure 4.1. The $E\left[L^2\right]$ for the multilevel IRT model and the multilevel true score model.

7. An Illustrative Example of Measurement Error in Hierarchical Models

The multilevel IRT and the true score models were used in an analysis of a mathematics test, administered to 3713 pupils of grade 4 in 198 regular primary schools (Bosker, Blatchford, & Meijnen, 1999; Hofman & Bosker, 1999). Among other things, interest was focused on the relation between achievement in mathematics and educational provisions at the school level and adaptive instruction by teachers. A test measuring the willingness, and capability to introduce educational program changes was taken by teachers. This test, denoted as $\mathbf{X}$, consisted of 23 dichotomously scored items to measure adaptive instruction, denoted as AI.

By posing the following Level 1 model, the nested structure of the data was taken into account. For each school j $(j = 1, \ldots, J)$,

$$y_{ij} = \beta_{0j} + \beta_{1j} IQ_{ij} + e_{ij}, \tag{4.23}$$

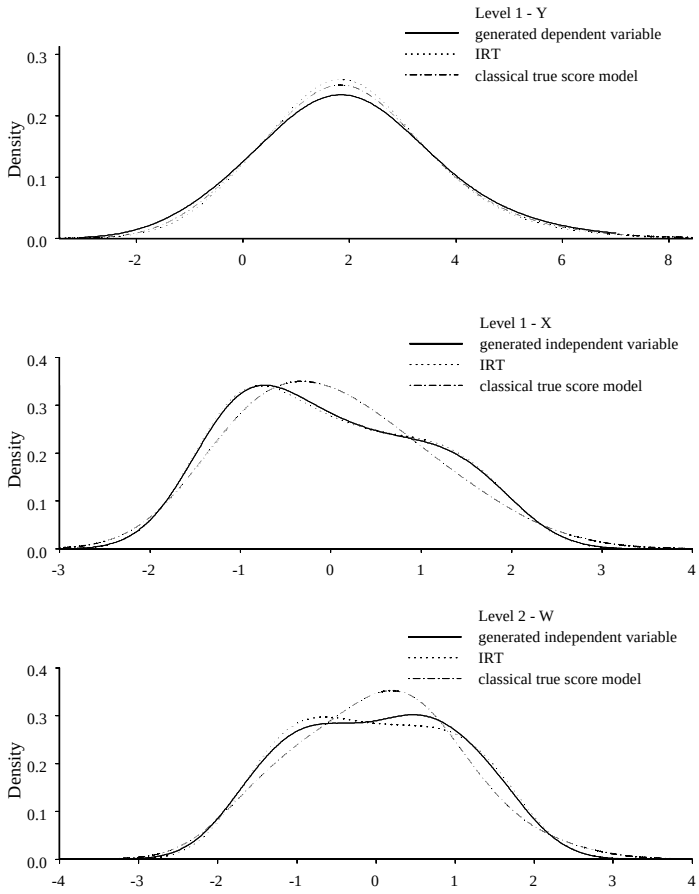

Figure 4.2. Density plots of the observed and replicated data using the normal ogive model and the classical true score model.

where y_{ij} was the score of the mathematics test and IQ_{ij} was an unobserved predictor representing the intelligence of a person indexed ij. IQ was measured by an intelligence test of 37 items, denoted as $\mathbf{W}$. The response patterns of 3713 pupils were available. The e_{ij} were assumed normally distributed with mean zero and variance σ^2.

First, it was assumed that the intercept was group-dependent and varies randomly from school to school. Furthermore, the sum scores measuring adaptive instruction were group level variables that express relevant attributes of the schools. They were supposed to have an influ-

ence in the diversity in mathematics scores. Therefore, the variability in β_{0j} was modeled as

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} AI_j + u_{0j} \\ \beta_{1j} &= \gamma_{10}, \end{aligned} \tag{4.24}$$

where u_{0j} were assumed normally distributed with variance τ_0^2.

For each analysis, from examining the plots of sampled parameter values, it was concluded that a burn-in period of 500 iterations was sufficient. Then an additional 20,000 Gibbs cycles, from which parameters of the posterior distribution were estimated, were run.

Table 4.3 presents the parameter estimates of Model 1, formula (4.23) , where a measurement error model was applied to the unobserved explanatory variable representing the IQ values of the examinees. The estimated group specific error variance, φ, was .39. For the moment, the mean observed score from the AI test was used, neglecting its error component. The main result of the analysis was that, conditionally on IQ, adaptive instruction for teachers seemed to have a small positive effect on mathematics achievements of students, but this effect did not differ significantly from zero. Furthermore, individuals with high IQ values scored high on the mathematics test. The use of multilevel model was justified, because a substantial proportion of the variation in the outcome at the student level was between schools. This is the variance of the achievements of students in school j controlling for IQ, around the grand mean, γ_{00}, which did not differ significantly from zero.

There were only small differences between the parameter estimates from the multilevel IRT model and the multilevel true score model, with $\varphi = .39$, denoted by M_1 and M_{c1}, respectively. The parameter estimates in Table 4.3 are comparable because the IQ predictors in both models were scaled to the standard normal distribution. The variance at Level 1 was slightly smaller for the multilevel true score model. The differences in handling the response error in the explanatory variable at Level 1 were evaluated using the posterior predictive data. Table 4.3 presents the $E\left[L^2\right]$ and corresponding standard deviations for both models. Model M_1 performed slightly better than model M_{c1}. Both models resulted in a better model fit in terms of minimization of $E\left[L_1^2\right]$ in comparison to the standard hierarchical model treating the AI and IQ variables as observed.

Next, a measurement error model was introduced for Level 2. The response variance of the AI test was modeled using (4.24). Table 4.4 presents the parameter estimates of the multilevel IRT model and the multilevel true score model with response error in IQ and AI. The model labeled M_2, modeled both unobserved predictors with a normal ogive

Table 4.3. Parameter estimates of the multilevel model with the normal ogive and the classical true score model as measurement error models.

	IRT Model M_1			Classical True Score Model $M_{c1}, \varphi = .39$		
Fixed Effects	Coefficient	s.d.	HPD	Coefficient	s.d.	HPD
γ_{00}	$-.018$	.075	$[-.164, .126]$	$-.017$	.074	$[-.162, .126]$
γ_{01}	.059	.075	$[-.089, .207]$	.052	.075	$[-.095, .198]$
γ_{10}	.397	.017	$[.364, .430]$	.487	.017	$[.453, .521]$
Random Effects	Variance Components	s.d.	HPD	Variance Components	s.d.	HPD
σ	.845	.028	$[.825, .865]$	.801	.028	$[.780, .824]$
τ_0	.349	.011	$[.296, .403]$	.338	.011	$[.287, .394]$
	$E\left[L^2\right]$	s.d.		$E\left[L^2\right]$	s.d.	
	1.873	.035		1.978	.037	

model, Model M_{c2} used the classical true score model as measurement model for both predictors with $\varphi_1 = .39$ and $\varphi_2 = .43$ as the estimated response variance for the *IQ* and *AI* test, respectively. The results from both models showed that adaptive instruction for teachers still had no significant effect on the mathematics achievements of students. Further, students with high IQ scores still performed better than students with lower scores. The proportion of variance in mathematics scores accounted for by group-membership, controlling for IQ scores, was .148 using model M_2 and .146 using model M_{c2}. This emphasized the small differences between the parameter estimates of both models.

Model M_2 and M_{c2} considered response error in all predictors. The $E\left[L_1^2\right]$ was reduced for both models in comparison to model M_1 and M_{c1} but the $E\left[L^2\right]$ increased due to the extra error term $E\left[L_3^2\right]$. The variability in the predictors induced larger variances of the parameter estimates and decreased the distance between the replicated data and the observed data. Correcting for bias resulted in more variable estimates but also in a better prediction of the data. The lowest value of $E\left[L^2\right]$ was obtained with model M_2. This means that the predicted data corresponded to the observed data at best with model M_2. In case of model M_{c2}, the estimated variance at Level 1 was lower and the esti-

Table 4.4. Parameter estimates of the multilevel model with the normal ogive and the classical true score model as measurement error models on both levels.

	IRT Model M_2			Classical True Score Model $\varphi_1 = .39, \varphi_2 = .43$, M_{c2}		
Fixed Effects	Coefficient	s.d.	HPD	Coefficient	s.d.	HPD
γ_{00}	$-.017$	.087	$[-.188, .153]$	$-.018$	.086	$[-.191, .147]$
γ_{01}	.055	.089	$[-.120, .231]$	.094	.097	$[-.100, .279]$
γ_{10}	.410	.019	$[.373, .448]$	.447	.021	$[.404, .485]$
Random Effects	Variance Components	s.d.	HPD	Variance Components	s.d.	HPD
σ	.854	.013	$[.830, .879]$	.837	.013	$[.811, .862]$
τ_0	.357	.034	$[.292, .422]$	.345	.035	$[.283, .418]$
	$E\left[L^2\right]$	s.d.		$E\left[L^2\right]$	s.d.	
	2.453	.087		2.735	.098	

mates of the fixed effects were somewhat larger resulting in a slightly better prediction of the dependent variable. But the inferior predictions of the observed sum scores related to the IQ and AI test resulted in a higher value of the statistic $E\left[L^2\right]$. In general, model M_2 fitted the observed data $(\mathbf{y}, \mathbf{x}, \mathbf{w})$ best.

Overall, it can be concluded that correcting for measurement error with the normal ogive model on both levels resulted in more variance of the parameter estimates but less bias and the model fit is better. In general, the use of a measurement error model led to a reduction in bias and variance of the replicated data in relation to the observed data in all cases.

A weak point of the classical true score model is that the measurement error variance has to be imputed. The Gibbs sampler was used to estimate the multilevel true score model and the corresponding $E\left[L^2\right]$ for various values of φ_1 and φ_2. Varying φ will lead to different predictions with respect to the observed data. Figure 4.3 displays the $E\left[L^2\right]$ and $E\left[L_1^2\right]$ for various values of the a priori established error variance on Level 1 and Level 2. It can be seen that $E\left[L_1^2\right]$ decreased when the variance in the predictor variable IQ increased. This follows directly from formula (4.13). The posterior mean of $\boldsymbol{\theta}$ is based on the values of the ob-

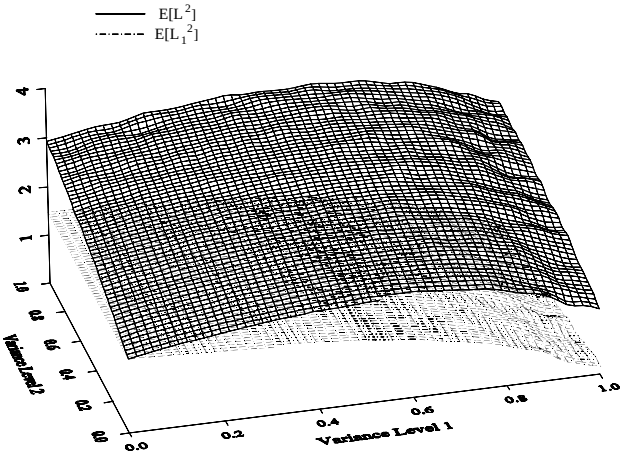

Figure 4.3. The $E\left[L^2\right]$ and $E\left[L_1^2\right]$ for different values of the error variance to model the latent predictor variables on Level 1 and Level 2 with the classical true score model.

served data $\mathbf{y}$ if the variance in the observed data $\mathbf{x}$ is high. As a result, the predictions $\mathbf{y}^{rep}$ resemble the observed data $\mathbf{y}$ more. It follows that the discrepancy between the observed data $\mathbf{y}$ and the predicted data $\mathbf{y}^{rep}$ enlarges when the response variance decreases. Then the posterior mean of $\boldsymbol{\theta}$ is based largely on the observed sum scores relating to the intelligence test, instead of on $\mathbf{y}$. The $E\left[L^2\right]$ increased when one or both of the response variances increased because the distance between predicted IQ-scores and AI-scores deviated more from the observed sum scores, partly due to the enlarged response variances. High response variance in the IQ test led to better results of the statistics $E\left[L^2\right]$ and $E\left[L_1^2\right]$. Generally, the prior information about the group specific error variance highly influenced the results.

8. Discussion

In this chapter, a normal ogive model is imposed on the unobserved explanatory variables in a multilevel model. In the social sciences, it is rarely possible to measure all relevant covariates directly and accurately. Correcting for measurement error is dependent on knowledge of the mea-

surement error process. Here, the normal ogive model describes the link between the observed data and the unobserved variables. This is compared with the classical true score model as measurement error model. Appropriate methods for correcting for the effects of measurement error depend on the measurement error distribution (Carroll et al., 1995). It is shown that both measurement error models reduce the bias in the estimates with an increase of the variance. This bias versus variance trade-off works well in both cases. Better results are obtained with the multilevel IRT model in terms of the expected square distance between all observed and predicted data. The multilevel true score model requires information about the group specific error variance and depends highly on this prior information. This leads to a certain degree of arbitrariness. Moreover, the variance structure of the errors in the predictor variables is difficult to estimate. The multilevel IRT model amounts to a more realistic way of modeling measurement error in the predictor variables, because it does not depend on any arbitrary assumption on the error variance structure.

It is possible to use other IRT models as a measurement error model. Examples are the three-parameter item response model and models for polytomously scored items. These models can be estimated within the Bayesian framework using the Gibbs sampler (Béguin, 2000; Johnson & Albert, 1999). If the conditional distribution of some parameters is difficult to sample from, then a Metropolis-Hastings step within Gibbs sampler can be used to obtain samples from the posterior distribution of the specific parameters (Chib & Greenberg, 1995).

The test statistic discussed above only focuses on the extent to which the observed data are reproduced by the model. Other posterior predictive checks can be developed to judge the fit and assumptions of the model, such as local independence and homoscedasticity, but this is beyond the scope of the present chapter.

In the present chapter, the response variable, $\mathbf{Y}$, is treated as observed without measurement error. It is possible to extend the procedure and to model this variable with an IRT model also. This more complex problem, where both the response and some of the predictors are measured with error, deserves further research. The basic structure of this more complex model is related to the multilevel IRT model, Chapter 2 and 3, or the generic hierarchical IRT model (Patz & Junker, 1999b) with background variables measured with an error.This whole framework is also strongly related to the framework of structural equation modeling, where there is a measurement part and a structural part. The measurement part of the model consists of the response variable and observed predictor surrogates and latent variables, and the structural part is de-

fined in terms of the latent variables regressed on each other and some observed background variables. In MIMIC modeling (see, for example, Bollen, 1989; Muthén, 1989), one or more latent variables intervene between the observed background variables predicting a set of observed response variables and surrogates. The main difference between these approaches and the one presented here is the use of an IRT model as a measurement error model, and integration of these various approaches remains a point of further study.

Chapter 5

Bayesian Model Checking and Residual Analysis

1. Introduction

School effectiveness research is a major topic in education, especially in light of the concern for evaluation of differences in achievement and accountability. Main interest is put in identifying the characteristics of effective schools and criteria for measuring effectiveness. This research is characterized by a variety of methods and designs. Therefore, issues as sample sizes, what variables should be measured, and at what level the data should be analyzed need to be tackled. In this chapter, attention is focused on model choice and goodness of fit.

The methods of measuring school effectiveness have been changed radically with the development of multilevel analysis. The hierarchical structure of educational systems emphasizes the necessity of multilevel modeling. Multilevel analysis enables that the data are treated in an appropriate manner, instead of being reduced to a single level. The differences between classes and schools can be taken into account properly, rather than aggregated arbitrarily. In this framework, most of the variance is explained by student background variables, such as intelligence and socio-economic status, other parts of the variance can be explained by class or school factors. Applications of multilevel models to educational data can, for example, be found in Bock (1989) and Goldstein (1995).

A major component of any school effectiveness assessment is the use of achievement scores as a measure of effectiveness. Most often, schools are compared in terms of the achievements of the pupils, and sum scores are used to represent these achievements. In Chapter 3, a measurement error model, that is, an item response theory model, is proposed to specify the

relationship between latent abilities and observed responses of students. Together with the structural multilevel model, this resulted in a multilevel IRT model. This model is evaluated in a fully Bayesian framework, which has the advantage that all parameters can be estimated concurrently with the Gibbs sampler (Gelfand & Smith, 1990). Furthermore, a fully Bayesian framework supports definition of a full probability model for quantifying uncertainty. One step further in statistical inference is the assessment of the plausibility of the posited model or of some of its specific assumptions.

In this chapter, several Bayesian checks are proposed that can be used to judge the fit and assumptions of a multilevel IRT model. The binary outcomes on item-level are supposed to have an underlying normal regression structure on latent continuous data. This assumption results in a analysis of Bayesian latent residuals. It will be shown that the Bayesian latent residuals have continuous-valued posterior distributions and are easily estimated with the Gibbs sampler. This in contrast to the classical residuals that are difficult to define and interpret due to the discrete nature of the response variable (Albert & Chib, 1995). Further, Bayesian residuals have different posterior variances but the Bayesian latent residuals are identically distributed. An unbiased estimator of the Bayesian latent residuals and its variance will be proposed using its conditional expectation given a sufficient statistic.

The posterior distributions of the random errors are used to detect outliers in the multilevel IRT model. An outlier is defined as an observation with a large random error, generated by the model under consideration (Chaloner & Brant, 1988). The posterior distributions can be used to calculate the posterior probability that an observation is an outlier. These posterior probabilities of an observation being an outlier are calculated with the Gibbs sampler. Other Bayesian approaches to outlier detection can be found in, for example, Box and Tiao (1973) and Zellner (1975).

Hypotheses can be tested using highest posterior density intervals. According to the usual form of a hypothesis that a parameter value or a function of parameter values is zero, HPD intervals will show, in most cases, the difference (Box & Tiao, 1973). This concept is used to check heteroscedasticity at Level 1, that is, to check whether grouped Level 1 residuals have the same posterior distribution. The parametric forms of the marginal posterior distributions are unknown, but samples of the distributions are available through the Gibbs sampler. These samples are utilized to check the homoscedasticity assumption at Level 1.

Further, the sensitivity of inferences to reasonable changes in the prior distribution will be examined. The need to study alternative prior dis-

tributional assumptions for the variance components arises when Bayes factors are used for model comparison. The Gibbs sampler will fail when prior distributions are specified that become infinite at zero (Pauler, Wakefield, & Kass, 1999). The discussion of improper priors in relation with Bayes factors can be found in, e.g., Lavine and Schervish (1999) and O'Hagan (1995).

In the first section, Bayesian residual analysis and estimation methods are described. Next, a method to detect outliers by examining the posterior distribution of the residuals using Gibbs sampler is discussed. Then, tests based on highest posterior density intervals, are described to test the homoscedasticity at Level 1. Further, the prior sensitivity of the parameter estimates will be discussed. Then, examples of the procedure will be given by analyzing the data used in Chapter 3. Finally, the last section contains a discussion and suggestions for further research.

2. Bayesian Residual Analysis

The multilevel IRT model assessment includes a check whether the assumptions made to specify the model are justified. This can be done by examining the regression residuals to check such assumptions as normality, conditional independence of observations and homoscedasticity of variance. Furthermore, there is interest in the magnitudes of the errors that actually occurred. The realized errors are not observed. They need to be estimated from the data together with the uncertainties associated with these estimates. In the present chapter, realized residuals are viewed as random parameters with unknown values. Posterior distributions for realized errors need to be calculated and can be used to make posterior probability statements about the values of the realized errors (see, for example, Box & Tiao, 1973; Zellner, 1975).

The residuals are defined as

$$r_{ijk} = y_{ijk} - \Phi\left(a_k\theta_{ij} - b_k\right).$$

In the classical residual analysis, the most common way toward analyzing residuals is to transfer the residuals to a scale where they are approximately normally distributed. The most common normalizing transformations lead to Pearson, deviance, and adjusted deviance residuals. But in case of Bernoulli observations such transformations result in poor approximations of the distributions of the Pearson, deviance and adjusted deviance residuals by the Gaussian distribution. A fully Bayesian residual analysis does not suffer from this problem. In the Bayesian residual analysis attention is focused on the posterior distribution of each residual. Bayesian residuals have continuous-valued posterior distributions which can also be used to detect outliers.

In Chapter 3 and 4, an MCMC estimation method has been proposed to estimate all parameters of the multilevel IRT model. The proposed Gibbs sampler can be used to estimate the posterior distribution of the residuals. Denote an MCMC sample from the posterior distribution of the parameters (θ_{ij}, a_k, b_k) by $\left(\theta_{ij}^{(m)}, a_k^{(m)}, b_k^{(m)}\right)$, $m = 1, \ldots, M$. It follows that sampled values from the residual posterior distribution corresponding to observation ijk are defined by

$$r_{ijk}^{(m)} = y_{ijk} - \Phi\left(a_k^{(m)}\theta_{ij}^{(m)} - b_k^{(m)}\right), \; m = 1, \ldots, M. \tag{5.1}$$

To check that these residuals are normally distributed, the ordered sampled values can be compared to the expected order statistics of the normal distribution in a quantile-quantile plot. Further, interest is focused on identifying residuals whose distribution is concentrated on an interval not containing zero. Checking if a residual r_{ijk} is unusually large can be done by plotting the quantiles of the posterior distribution of r_{ijk} against the posterior mean of the probability $p_{ijk} = \Phi\left(a_k\theta_{ij} - b_k\right)$, (Albert & Chib, 1995). A drawback is that the marginal distributions of the ordered residuals differ. For example, the distribution of the smallest residual is different from that of the median residual. The posterior variances of the residuals differ and are not directly comparable. Therefore, it is difficult to assess how extreme each distribution is. These problems can be averted by using Bayesian latent residuals as an alternative to the Bayesian residuals.

2.1 Computation of Bayesian Latent Residuals

The observation Y_{ijk} can be interpreted as an indicator variable that a continuous variable with normal density is above or below zero. This latent continuous score is defined as Z_{ijk}, where $Z_{ijk} > 0$ if $Y_{ijk} = 1$ and $Z_{ijk} \leq 0$ if $Y_{ijk} = 0$. Complete data, consisting of augmented data $\mathbf{Z}$ and observed data $\mathbf{Y}$, are formed to simplify calculations. From the definition of Z_{ijk} it follows that

$$p\left(z_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k, y_{ijk}\right) \propto \phi\left(z_{ijk}; a_k\theta_{ij} - b_k, 1\right)\left[I\left(z_{ijk} > 0\right) I\left(y_{ijk} = 1\right) + I\left(z_{ijk} \leq 0\right) I\left(y_{ijk} = 0\right)\right], \tag{5.2}$$

where $\phi(.; a_k\theta_{ij} - b_k, 1)$ stands for the normal density with a mean equal to $a_k\theta_{ij} - b_k$ and a variance equal to one, and $I\left(.\right)$ is an indicator variable taking the value one if its argument is true, and the value zero otherwise. Implementations of the Gibbs sampler with the use of augmented data can be found in Albert (1992), Johnson and Albert (1999) and, in particular, for the multilevel IRT model in Chapter 3 and 4.

The Bayesian latent residuals corresponding to observations Y_{ijk} are defined as

$$\varepsilon_{ijk} = Z_{ijk} - a_k\theta_{ij} + b_k. \tag{5.3}$$

From the definition of the augmented data it follows that given a_k, b_k and θ_{ij}, the latent residuals ε_{ijk} are standard normally distributed. These latent residuals are easily estimated as a by-product of the Gibbs sampler. That is, MCMC samples from Z_{ijk}, $\boldsymbol{\xi}_k$ and θ_{ij} produce samples ε_{ijk} from its posterior distribution. Accordingly, posterior means and standard deviations of the latent residuals can be computed from the sampled values. A more efficient estimator is the conditional expectation given a sufficient statistic, called a Rao-Blackwellised estimator (Gelfand & Smith, 1990). That is, the sampling error attributable to the Gibbs sampler is reduced to obtain a more efficient estimate of the posterior means. The unbiased character of the Monte Carlo estimator remains while reducing its variance.

The conditional expectation of the latent residuals needs to be calculated given a sufficient statistic. Suppose that $Y_{ijk} = 1$, it follows that

$$\begin{aligned} E\left(\varepsilon_{ijk} \mid Y_{ijk} = 1, \theta_{ij}, \boldsymbol{\xi}_k\right) &= \int_0^\infty E\left(\varepsilon_{ijk} \mid z_{ijk}, Y_{ijk} = 1, \theta_{ij}, \boldsymbol{\xi}_k\right) \\ &\qquad \frac{f\left(z_{ijk}, Y_{ijk} = 1 \mid \theta_{ij}, \boldsymbol{\xi}_k\right)}{f\left(Y_{ijk} = 1 \mid \theta_{ij}, \boldsymbol{\xi}_k\right)} dz_{ijk} \\ &= \int_0^\infty \varepsilon_{ijk} \frac{\phi\left(z_{ijk} - a_k\theta_{ij} + b_k\right)}{\Phi\left(a_k\theta_{ij} - b_k\right)} dz_{ijk} \\ &= \frac{\phi\left(b_k - a_k\theta_{ij}\right)}{\Phi\left(a_k\theta_{ij} - b_k\right)}, \end{aligned} \tag{5.4}$$

where ϕ represents the density of the standard normal distribution. Likewise, it follows for $Y_{ijk} = 0$ that

$$E\left(\varepsilon_{ijk} \mid Y_{ijk} = 0, \theta_{ij}, \boldsymbol{\xi}_k\right) = \frac{-\phi\left(b_k - a_k\theta_{ij}\right)}{\Phi\left(b_k - a_k\theta_{ij}\right)}. \tag{5.5}$$

The expected value of ε_{ijk} depends on the value of Y_{ijk} and the sign of $a_k\theta_{ij} - b_k$. Some elementary calculations need to be done to find an expression for the variance. Define $\mu_{ijk} = E\left(\varepsilon_{ijk} \mid Y_{ijk}, \theta_{ij}, \boldsymbol{\xi}_k\right)$ and

assume that $Y_{ijk} = 1$. Skipping the elementary steps it follows that

$$\begin{aligned} Var\left(\varepsilon_{ijk} \mid y_{ijk}, \theta_{ij}, \boldsymbol{\xi}_k\right) &= \int_0^\infty \left(\varepsilon_{ijk} - \mu_{ijk}\right)^2 \frac{f\left(z_{ijk}, y_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k\right)}{f\left(y_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k\right)} dz_{ijk} \\ &= \int_0^\infty \left(\varepsilon_{ijk} - \mu_{ijk}\right)^2 \frac{\phi\left(z_{ijk} - a_k\theta_{ij} + b_k\right)}{\Phi\left(a_k\theta_{ij} - b_k\right)} dz_{ijk} \\ &= 1 - \frac{\phi\left(-\eta_{ijk}\right)}{\Phi\left(\eta_{ijk}\right)} \left(\eta_{ijk} + \frac{\phi\left(-\eta_{ijk}\right)}{\Phi\left(\eta_{ijk}\right)}\right), \end{aligned} \tag{5.6}$$

where $\eta_{ijk} = a_k\theta_{ij} - b_k$. In the same way in case $Y_{ijk} = 0$, the variance of estimate (5.5) is

$$1 - \frac{\phi\left(-\eta_{ijk}\right)}{\Phi\left(-\eta_{ijk}\right)} \left(-\eta_{ijk} + \frac{\phi\left(-\eta_{ijk}\right)}{\Phi\left(-\eta_{ijk}\right)}\right). \tag{5.7}$$

The estimates of the Bayesian latent residuals are easily implemented in a Gibbs sampler. Then it can be checked if the latent residuals are normally distributed given the observations by a quantile-quantile plot.

The realized residuals at Level 1 are viewed as parameters with unknown values. The estimates of these latent residuals follow directly from their definition. That is,

$$\begin{aligned} E\left(e_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\theta}_j, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right) = E\left(\theta_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2\right) \\ - \mathbf{X}_{ij} E\left(\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right), \end{aligned}$$

where MCMC samples are needed of the parameters on which they are conditioned. Both expectations on the right-hand side are easily derived due to the fact that the conditional distributions of the parameters consists of a product of two normal densities, see Chapter 3. The variance of the residuals at Level 1 can be estimated in the same way. Posterior probability statements about the values of the realized errors are made via the derivation of the posterior distributions from the Gibbs sampler. Collecting the raw residuals $e_{ij}^{(m)} = \theta_{ij}^{(m)} - \mathbf{X}_{ij}\boldsymbol{\beta}_j^{(m)}$ for the $m = 1, \ldots, M$ iterations of the Gibbs sampler results in samples from the posterior distribution of the latent residuals at Level 1. Also posterior moments, measures of skewness and kurtosis can be derived from the Gibbs sampler. Characterizing the properties of the realized residuals can help to discover the nature of possible departures from the assumptions of the multilevel IRT model. Finally, latent residuals at Level 2 are estimated as

$$E\left(\mathbf{u}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right) = E\left(\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right) - \mathbf{W}_j E\left(\boldsymbol{\gamma} \mid \boldsymbol{\beta}_j, \mathbf{T}\right). \tag{5.8}$$

The conditional distributions of the parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}$ are both normally distributed, and, for that reason, the expectations are easily derived. The rough residuals $\mathbf{u}_j^{(m)} = \boldsymbol{\beta}_j^{(m)} - \mathbf{W}_j\boldsymbol{\gamma}^{(m)}$ for $m = 1, \ldots, M$ are samples from the marginal posterior distribution. Among other things, the samples can be used to estimate the posterior variance. A quantile-quantile plot can be made to check if the latent residuals are normally distributed given the observations.

It must be remarked that the residuals in the measurement model and at Level 1 and Level 2 of the multilevel model are not considered separately in this manner. This means that analyzing residuals of the measurement model is based on specifications of the multilevel model and, analogously, analyzing residuals of the multilevel model is based on specifications of the measurement model. Obviously, it is possible to check the residuals of the measurement model separately by estimating the model independent of the multilevel model but then residuals are analyzed from a completely different model. The Level 1 residuals can be computed by ordinary least squares regressions within each group separately (see, Snijders & Bosker, 1999, pp. 128-132), given estimates of the latent variables at Level 1. But the estimates of the latent variables will also contain the influence of the residuals in the measurement model. The residuals at Level 1 will be analyzed unconfounded by the Level 2 residuals but the other residuals are analyzed according to the complete model.

3. Detection of Outliers

The outlier detection problem is addressed from a Bayesian perspective. As stated above, realized regression error terms are treated as unknown parameters, see Zellner (1975). The posterior distribution of these residuals can be used to calculate the posterior probability that an observation is an outlier. Outliers can be detected by examining the posterior distribution of the error terms. An observation can be considered to be outlying if the posterior distribution of the corresponding residual is located far from its mean (Albert & Chib, 1995). Here, the posterior distribution of the Bayesian latent residuals are examined to detect outliers among the observations. The Bayesian latent residuals are a function of unknown parameters and the posterior distributions are therefore straightforward to calculate.

Following Chaloner and Brant (1988), Johnson and Albert (1999) and Zellner (1975), the ijk^{th} observation is an outlier if the absolute value of the residual is larger as some pre-specified value l times the standard deviation. That is, observations Y_{ijk} with a high posterior probability, $P\left(|\varepsilon_{ijk}| > l \mid y_{ijk}\right)$, are marked as outliers, according to formula (5.3).

In fact the augmented continuous scores Z_{ijk} are marked as outliers but Z_{ijk} has a one-to-one correspondence with Y_{ijk}. The probability that an observation exceeds a pre-specified value is called the outlying probability. The outlying probabilities can be estimated with the Gibbs sampler.

First, consider the residuals at the IRT level. Suppose $Y_{ijk} = 1$, it follows that

$$P\left(|\varepsilon_{ijk}| > l \mid Y_{ijk} = 1, \theta_{ij}, \boldsymbol{\xi}_k\right) = \int_l^\infty \frac{f\left(z_{ijk}, Y_{ijk} = 1 \mid \theta_{ij}, \boldsymbol{\xi}_k\right)}{f\left(Y_{ijk} = 1 \mid \theta_{ij}, \boldsymbol{\xi}_k\right)} dz_{ijk}$$
$$= \frac{\Phi\left(-l\right)}{\Phi\left(a_k\theta_{ij} - b_k\right)} \tag{5.9}$$

and if $Y_{ijk} = 0$, then

$$P\left(|\varepsilon_{ijk}| > l \mid y_{ijk}, \theta_{ij}, \boldsymbol{\xi}_k\right) = \frac{\Phi\left(-l\right)}{1 - \Phi\left(a_k\theta_{ij} - b_k\right)}. \tag{5.10}$$

To obtain an estimate of the probability $P\left(|\varepsilon_{ijk}| > l \mid y_{ijk}\right)$ MCMC samples of the ability and item parameters are needed to calculate the mean of $P\left(|\varepsilon_{ijk}| > l \mid y_{ijk}, \theta_{ij}^{(m)}, \boldsymbol{\xi}_k^{(m)}\right)$ for $m = 1, \ldots, M$. As in Chapter 3 and 4, an improper simultaneous prior, $p\left(\mathbf{a}, \mathbf{b}\right) \propto \prod_{k=1}^K I\left(a_k > 0\right)$, is used in the Gibbs sampler for the item parameters to insure that each item will have a positive discrimination index. Other priors are possible; examples will be discussed below.

It is possible to find l such that the probability $P\left(|\varepsilon_{ijk}| > l \mid y_{ijk}\right)$ assumes a given percentage, say $\nu\%$. Therefore, in every Gibbs iteration l must be solved in the equation $P\left(|\varepsilon_{ijk}| > l \mid y_{ijk}\right) = \frac{\nu}{100}$. The mean of these values is an estimate of the unique root, that is, the l-percent value, or the probability that Z_{ijk} will deviate from its mean by more than l.

The choice of l is quite arbitrarily, but if the model under consideration is required to describe the data then $l = 2$ might be used to find observations that are not well described by the data. There is reason for concern if more than 5% of the residuals have high posterior probability of being greater than two standard deviations.

Notice that other complex posterior probabilities can be computed with the Gibbs sampler by keeping track of all the possible outcomes of the relevant probability statement. However, this method has the drawback that a lot of iterations are necessary to get a reliable estimate. It could be possible, for example, that in case of multiple outliers a test for a single outlier does not detect one outlier in the presence of another outlier. This so-called masking occurs when two posterior probabilities,

say $P(|\varepsilon_{ijk}| > l \mid y_{ijk})$ and $P(|\varepsilon_{sjk}| > l \mid y_{sjk})$, do not indicate any outliers but the posterior probability $P(|\varepsilon_{ijk}| > l \text{ and } |\varepsilon_{sjk}| > l \mid \mathbf{y})$ shows that ε_{ijk} and ε_{sjk} are both outliers. This simultaneous probability can be estimated by counting the events that both absolute values of the residuals are greater than l times the standard deviation divided by the total number of iterations.

4. Heteroscedasticity

In a standard linear multilevel model, the residuals at Level 1 and 2 are assumed to be homoscedastic. It is possible that the variances of the residuals are heteroscedastic when they depend on some explanatory variables. By modeling the variation as a function of the explanatory variables will return homoscedastic variances. Neglecting the heteroscedasticity may lead to incorrect inferences concerning the hypotheses tests for variables which are responsible for the heteroscedasticity (Snijders & Bosker, 1999).

In a Bayesian framework, complex variance structures can be defined as prior information. Here, Level 1 variation will be considered but the same principles apply to higher levels. General functions of more than one explanatory variable can be considered to model the variance at Level 1. Examples of complex variation modeling are given in, for example, Goldstein (1995, pp. 50) and Snijders & Bosker (1999, pp. 110-119). One common example is the case where variances are specific for subgroups. For example, the error variance could differ for male and female respondents. Besides modeling the variance conditional on the value of the explanatory variable, it is also possible to give the explanatory variable a random slope at Level 1. Further, the random slope variances can be made to depend on some variable. In case the variance parameter cannot be sampled directly from the full conditional distribution a Metropolis-Hastings-within-Gibbs step could be incorporated.

Without a specific connection to some explanatory variable heteroscedasticity is harder to detect. Therefore, it is sometimes useful to consider the possibility that the residual variance differs between groups. The Gibbs sampler can be used to generate samples of group specific residual variances. Subsequently, these draws can be used to check the assumption of between-group differences in the Level 1 residual variance.

4.1 Highest Posterior Density Intervals

Here, two tests for heteroscedasticity at Level 1 in case of two or more groups are considered that are easy to compute in combination with the use of a Gibbs sampler. Testing the equality of variances of two or more

grouped residuals at Level 1 coincides with the hypothesis that samples of residuals $(\mathbf{e}_1, \ldots, \mathbf{e}_L)$ have a common variance σ^2. That is, samples drawn from the marginal distributions of the group specific variances have the same location. The sampled values are used to estimate the posterior means of σ_l^2, $l = 1, \ldots, L$ but can also be used to test the hypothesis, $\sigma_1^2 = \ldots = \sigma_L^2$ against the alternative $\sigma_l^2 \neq \sigma_{l'}^2$ for at least one $l \neq l'$. Under the null-hypothesis the L samples are drawn from a common population.

In case of two groups at Level 1, the variances of two Normal distributions, denoted as σ_1^2 and σ_2^2, are compared. By looking at the highest posterior density interval of σ_2^2/σ_1^2 it can be judged if the residual variance of group 1 differ from group 2. Since

$$\frac{\sigma_2^2/s_2^2}{\sigma_1^2/s_1^2} \sim F\left(n_1 - 1, n_2 - 1\right) \tag{5.11}$$

where $s_i^2 = \sum_{i=1}^{n_i} \left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2$ for $i = 1, 2$, it follows that

$$\frac{\sigma_2^2}{\sigma_1^2} \sim \frac{s_2^2}{s_1^2} F\left(n_1 - 1, n_2 - 1\right), \tag{5.12}$$

see, Box and Tiao (1973, pp. 110-112). The mode of the distribution of F is 1, thus the mode of the posterior distribution of σ_2^2/σ_1^2 is s_2^2/s_1^2. Using the Gibbs sampler, the limits of the HPD interval are specified by the F distribution in combination with an estimate of s_2^2/s_1^2.

To insure that comparisons of L scale parameters $\left(\sigma_1^2, \ldots, \sigma_L^2\right)$ are unaffected by any linear recoding of the data, consider $(L-1)$ linearly independent contrasts in $\log \sigma_l^2$. So, let $\Delta_l = \log \sigma_l^2 - \log \sigma_L^2$. The point $\Delta_0 = \mathbf{0}$ is included in the highest posterior density region of content $(1 - \alpha)$ if and only if

$$P\left(p\left(\Delta \mid \mathbf{y}\right) > p\left(\Delta_0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) < 1 - \alpha.$$

The density function $p\left(\Delta \mid \mathbf{y}\right)$ is a monotonic decreasing function of a function with parameters σ_l^2 and s_l^2 which is asymptotically distributed as χ_{L-1}^2, as $n_l \to \infty$, $l = 1, \ldots, L$, where s_l^2 is the mean sum of squares in group l (Box & Tiao, 1973, pp. 133-135). In case of the hypothesis $\Delta_0 = \mathbf{0}$, which corresponds to the situation $\sigma_1^2 = \ldots = \sigma_L^2$, this function becomes

$$M_0 = -\sum_{l=1}^{L} n_l \left(\log s_l^2 - \log \overline{s}^2\right) \tag{5.13}$$

where $\overline{s}^2 = \frac{1}{N}\sum_{l=1}^{L} n_l s_l^2$. It follows that

$$\lim_{n_l \to \infty} P\left(p\left(\Delta \mid \mathbf{y}\right) > p\left(\Delta_0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) = P\left(\chi^2_{L-1} < M_0\right).$$

Hence, for large samples, the point $\Delta_0 = \mathbf{0}$ is included in the $(1-\alpha)$ highest posterior density region if

$$M_0 < \chi^2_{L-1,\alpha}. \tag{5.14}$$

For moderate sample sizes Bartlett's approximation can be used to approximate the distribution with greater accuracy (Box & Tiao, 1973, pp. 135-136). It follows that,

$$P\left[p\left(\Delta \mid \mathbf{y}\right) > p\left(\Delta_0 \mid \mathbf{y}\right) \mid \mathbf{y}\right] \doteq P\left(\chi^2_{L-1} < \frac{M_0}{1+A}\right), \tag{5.15}$$

where $A = \frac{1}{3(L-1)}\left(\sum_{l=1}^{L} n_l^{-1} - N^{-1}\right)$. The difficulty in practice with this test for equal variances is the extreme sensitivity to the assumption of normality.

The expression on the right of (5.13) is computed by taking the mean over the computed values of (5.13) in every iteration of the Gibbs sampler. Notice that it is not necessary to estimate the model with the assumption of heteroscedasticity on level 1, because the value of σ^2 can be passed on in the Gibbs sampler. That is, sample σ_l^2, $l = 1, \ldots, L$, from the conditional distribution given $(\boldsymbol{\theta}, \boldsymbol{\beta})$, but pass on the sampled σ^2, based on $(\boldsymbol{\theta}, \boldsymbol{\beta})$, to update the conditional values of the other model parameters in the Gibbs sampler. Group specific variances are sampled separately but the variances are equal to the overall sampling variance σ^2, in case the null-hypothesis of equal variances is true.

It is possible to compute the highest posterior density of $\left(\sigma_1^2, \ldots, \sigma_L^2\right)$ given the observed data by integrating over the random effects $(\boldsymbol{\theta}, \boldsymbol{\beta})$ and computing the probability density, in every iteration of the Gibbs sampler. The highest posterior density region should be constructed in such a way that the probability of every set of interior points is at least as large that of any set of exterior points. Further, the region should be such that for a given probability, it occupies the smallest possible volume in the parameter space. The obtained vectors of parameter values can be used to construct such a region. Accordingly, the equality of variances can be tested by checking if the vector $\left(\sigma_1^2, \ldots, \sigma_L^2\right) = \mathbf{0}$ lies within the highest posterior density region.

4.2 Normal Approximation to the Posterior Distribution

Another test of equality of variances is obtained by approximating the posterior distribution of the individual group specific variances by a normal distribution. If the posterior distributions are unimodal and roughly symmetric they can be approximated by a normal distribution centered at the mode (Bernardo & Smith, 1994, pp. 287-288; Gelman, Carlin, Stern, & Rubin, 1995, pp. 94-96). The approximation of the posterior distribution of $\log\left(\sigma_l^2\right)$ will turn out convenient since unknown parameters will enter only into the mean and not in the variance of the approximated distribution. Using a Taylor series expansion of $\log\left(\sigma_l^2\right)$ it follows that

$$p\left(\log\sigma_l^2 \mid \boldsymbol{\theta}^{(l)}, \boldsymbol{\beta}^{(l)}, \mathbf{y}\right) \approx N\left(\log\widehat{\sigma}_l^2, \left[I\left(\log\widehat{\sigma}_l^2\right)\right]^{-1}\right), \tag{5.16}$$

for $l = 1, \ldots, L$ where $\boldsymbol{\theta}^{(l)}$ and $\boldsymbol{\beta}^{(l)}$ denote the ability parameters and regression coefficients at Level 1 corresponding to group l. Further, $\log\widehat{\sigma}_l^2$ is the mode of the posterior distribution and $I\left(\log\widehat{\sigma}_l^2\right)$ is the observed information evaluated at the mode. With a noninformative prior locally uniform in $\log\sigma_l^2$ it follows that

$$p\left(\log\sigma_l^2 \mid \boldsymbol{\theta}^{(l)}, \boldsymbol{\beta}^{(l)}, \mathbf{y}\right) \approx N\left(\log s_l^2, \frac{2}{n_l}\right). \tag{5.17}$$

So the problem of testing $\sigma_1^2 = \ldots = \sigma_L^2$ is reduced to that of testing the equality of L means of independent normally distributed variables $s_l' = \log\left(s_l^2\right)$. This problem simplifies in the particular case that the number of observations per group are equal, that is, $n_l = n$. A test for testing the equality of the means of the L normal distributions is

$$\frac{\sum_{l=1}^{L}\left(s_l' - \overline{s}'\right)^2}{2/\left(n-1\right)} > C, \tag{5.18}$$

where $2/\left(n-1\right)$ is the common variance of the s_l' and where C is determined by

$$\int_C^{\infty} \chi_{L-1}^2\left(y\right) dy = \alpha. \tag{5.19}$$

If the observations per group differ then the transformation s_l'/λ_l, with $\lambda_l = 2/\left(n_l - 1\right)$, results in a test which rejects when

$$\sum_{l=1}^{L}\left(\frac{s_l'}{\lambda_l}\right)^2 - \frac{\left(\sum_{l=1}^{L} s_l'/\lambda_l^2\right)^2}{\sum_{l=1}^{L}\left(1/\lambda_l^2\right)} > C, \tag{5.20}$$

where C is determined by (5.19) see, Lehmann (1986, pp. 377). The Gibbs sampler is used to estimate the s_l' for every group l. That is, after a sufficient number of iterations, the test statistic is computed to test the homogeneity of variances.

Both parametric tests for equality of variances are highly sensitive to the assumption of normality and should be used with some carefulness. The assumption of normality can be checked by using the Student's t-distribution in place of the normal distribution to assess the sensitivity to the normal assumption by varying the degrees of freedom from large to small (see, e.g., Gelman et al., 1995, pp. 349; Seltzer, 1993).

5. Choice of Priors

Prior information about the parameters of the model is usually represented by an appropriately chosen probability distribution. A distinction can be made between two types of priors: data-based or informative priors, founded on information from past data and nondata-based, noninformative or vague priors, arising from, for example, theoretical considerations. Incorporating noninformative priors in the analysis is done in such a way that the mathematical operations can be done conveniently. The easiest way is to use natural conjugate priors which are useful in representing prior information. In the former chapters, noninformative conjugated priors were used to reflect vague ideas about the distribution of the parameters. In the fully Bayesian approach one must be aware that the obtained parameters may be sensitive to the choice of priors. Using various priors and comparing results can provide information on their impact. Below, prior choices and the importance of recalculating the marginal posteriors with alternative priors are discussed further.

Many noninformative priors are improper, that is, they do not integrate to a finite number. Kass and Wasserman (1996) discuss several problems caused by improper priors. Here, attention is focused on improper posteriors.The complexity of the multilevel IRT model makes it difficult to analytically check if the posteriors are proper. A more easy way is to use alternative priors and check the results for agreement.

Avoiding improper posterior distributions can be done by using proper conjugate priors that are diffuse (Carlin & Louis, 1996; Gelfand, Hills, Racine-Poon, & Smith, 1990). This means that its density is slowly varying over the region in which the likelihood function is concentrated. It is often possible to choose the spread in a proper prior suitable large to remain vague. An alternative to Jeffreys' prior for the Level 1 variance σ^2 is an inverse chi-square prior with degrees of freedom ν_1 small. With $1 < \nu_1 < 4$ the inverse chi-square distribution has an infinite variance, so the prior information is weak relative to the information provided

by the data. Furthermore, the prior density does not become infinitely large as σ^2 approaches zero. It follows that the Gibbs sampler includes the conditional distribution of σ^2 given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$,

$$\begin{aligned} p\left(\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) &\propto p\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \sigma^2\right) p\left(\sigma^2; \nu_1, S_0\right) \\ &\propto \left(\sigma^2\right)^{-\left(\frac{N+\nu_1}{2}+1\right)} \exp\left(\frac{-1}{2\sigma^2}\left(S^2 + S_0^2\right)\right), \end{aligned} \tag{5.21}$$

where S^2 is the sum of squares at Level 1, $N = \sum_{j=1}^{J} n_j$ and S_0 the scale parameter of the prior for σ^2. The conditional distribution of σ^2 given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is an inverse-gamma distribution with parameters $\left(\frac{N+\nu_1}{2}, \frac{S^2+S_0^2}{2}\right)$. In the Level 2 model, an inverse Wishart distribution with small degrees of freedom ν_2 can be used as a diffuse proper prior for $\mathbf{T}$. It follows that

$$\mathbf{T} \sim inv - Wishart\left(\nu_2 + J, (\mathbf{S} + \mathbf{S}_1)^{-1}\right) \tag{5.22}$$

where $\nu_2 \geq Q + 1$, the dimension of $\boldsymbol{\beta}_j$, $\mathbf{S}$ is the weighted sum of squares at Level 2 and $\mathbf{S}_1$ a positive definite scale matrix.

The difficulty with both priors lies in the specification of the scale parameters $(S_0, \mathbf{S}_1)$. Some idea about the values of the scale parameters could be obtained by estimating the multilevel parameters with ability parameters estimated from the program Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). This two-stage estimation procedure ignores the uncertainty concerning $\boldsymbol{\theta}$ in the estimation of the multilevel parameters, and, on the other hand, in the estimation of $\boldsymbol{\theta}$ the multilevel structure is ignored.

Conjugated proper prior information for $\boldsymbol{\gamma}$ can be expressed as a normal distribution, that is,

$$\boldsymbol{\gamma} \sim N\left(\boldsymbol{\gamma}_0, \Sigma_\gamma\right)$$

where both parameters, expectation $\boldsymbol{\gamma}_0$ and variance Σ_γ, must be specified. A diffuse proper prior is obtained by defining a very large spread. Another way to handle improper priors is to truncate the domain to a compact region, which will result in a proper prior. For example, Jeffreys' prior, $\boldsymbol{\gamma} \sim c$, defined on a compact space $\mathcal{A}$, results in a proper prior, and the full conditional of $\boldsymbol{\gamma}$ given $\boldsymbol{\beta}, \mathbf{T}$ becomes

$$p\left(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \mathbf{T}\right) \propto \prod_{j=1}^{J} p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right) I\left(\boldsymbol{\gamma} \in \mathcal{A}\right)$$

where $I(.)$ is an indicator function. However the impropriety of Jeffreys' prior, $\boldsymbol{\gamma} \sim c$, does not result in an improper posterior,

$$p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \mathbf{T}) = \frac{p(\boldsymbol{\beta}, \mathbf{T} \mid \boldsymbol{\gamma}) \cdot c}{c \cdot \int p(\boldsymbol{\beta}, \mathbf{T} \mid \boldsymbol{\gamma})\, d\boldsymbol{\gamma}}, \tag{5.23}$$

where c cancels. The technique of truncating the domain of an improper prior can also be used for improper priors for the variance components. Obviously, restricting the domain will result in a more informative prior.

The normal distribution can be used as an informative prior for the difficulty parameter. Specifying this distribution with a large spread results in a diffuse proper conjugated prior. A noninformative prior for the discrimination parameter must at least express positivity (Mislevy, 1986). This can be accomplished by assuming that the distribution of $a_k, k = 1, \ldots, K$, is lognormal. The lognormal prior is a non-conjugate prior which leads to difficulties in sampling from the full conditional distribution of the discrimination parameter. The Metropolis-Hastings algorithm can be used to sample from approximate full conditional distributions whilst maintaining the stationary distribution of the Markov chain (Gilks, 1996; Tierney, 1994). Therefore, sampling from the full conditional consists of the following three steps:

1. Sample a_k' from $p\left(a_k; \mu, \sigma^2\right)$
2. Sample u from $U(0,1)$
3. if $u \leq \min\left[1, \frac{p\left(\mathbf{y}_k \mid \boldsymbol{\theta}, a_k', b_k\right) p\left(a_k'; \mu, \sigma^2\right)}{p\left(\mathbf{y}_k \mid \boldsymbol{\theta}, a_k, b_k\right) p\left(a_k; \mu, \sigma^2\right)}\right]$ accept a_k' else holds the old value of a_k.

It is also possible to change the prior specifications, $\left(\mu, \sigma^2\right)$, during the estimation procedure, see, for example, Patz and Junker (1999b). This method, labeled Metropolis-Hastings-within-Gibbs, produces a different Markov chain but with the same stationary distribution. If a parameter cannot be sampled directly from its complete conditional distribution, a Metroplis step can be incorporated to sample from the full conditional. It must be mentioned that the convergence of the algorithm is depended on the proposal distribution in the Metropolis-Hastings step. Further, the parameters of the proposal distribution must be chosen carefully to establish an acceptable convergence rate for the Markov chain.

The specification of diffuse proper priors avoids, in most cases, possible problems as improper posteriors. Kass and Wasserman (1996) pointed out that possible problems can still occur if the prior dominates the data. In data dominated cases, that is, when the posterior

is dominated by a peaked likelihood, the use of improper and diffuse proper priors remains acceptable. Determining whether a posterior is data dominated is hard to establish. A possible solution is to use several noninformative priors and to check the results for agreement.

6. An Analysis of a Dutch Primary School Mathematics Test

This section is concerned with the study of a primary school leaving test. In Chapter 3, this dataset was analyzed to compare parameter estimates of a multilevel IRT model and an hierarchical linear model using observed scores. Here, the goodness of fit of the multilevel IRT model will be analyzed. Residuals at different levels are analyzed, outliers are identified and different models are compared. Further, heteroscedasticity at Level 1 is tested.

The dataset consisted of responses from 2156 grade 8 students, unequally spread over 97 schools, to 18 mathematics items taken from the school leaving examination developed by the National Institute for Educational Measurement (Cito). Of the 97 schools sampled, 72 schools regularly participated in the school leaving examination, denoted as Cito schools and the remaining 25 schools will be denoted as the non-Cito schools. Socio-economic status (SES), non-verbal intelligence test (ISI) and Gender were used as predictors for the students' achievement. SES was based on four indicators: the education and occupation level of both parents (if present). Predictors SES and ISI were normalized and standardized. The dichotomous predictor Gender was an indicator variable equal to 0 for males and equal to 1 for females. Finally, a predictor variable labeled End equaled 1 if the school participates in the school leaving test, and 0 if this was not the case.

Students were clustered over schools with a distinction between Cito and non-Cito schools. Consider the model M_1 given by

$$\begin{aligned} \theta_{ij} &= \beta_{0j} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\text{End}_j + u_{0j} \end{aligned} \tag{5.24}$$

where $e_{ij} \sim N\left(0, \sigma^2\right)$, $u_{0j} \sim N\left(0, \tau_0^2\right)$. The model contains random groups and random variation within groups. The dependent variable equals the sum of a general mean γ_{00}, a random effect at the school level, u_{0j}, and a random effect at the individual level, e_{ij}, corrected for the predictor End. The two-parameter normal ogive model is used as measurement model. In Table 5.1, the estimates of the parameters issued from the Gibbs sampler are given under the label IRT Model M_1. The reported standard deviations and HPD regions are the posterior

Table 5.1. Parameter estimates of a multilevel IRT model with explanatory variable End on Level 2.

	IRT Model M_1		
Fixed Effects	Coefficient	s.d.	HPD
γ_{00}	$-.273$	.210	$[-.621, .067]$
γ_{01}	.463	.240	$[.072, .854]$
Random Effects	Variance Component	s.d.	HPD
σ^2	.593	.071	$[.476, .707]$
τ_0^2	.204	.046	$[.130, .275]$

standard deviations and the 90% highest posterior density intervals, respectively.

The general mean achievement of the students was zero for those attending non-Cito schools and slightly higher for the students attending Cito schools. The intraclass-correlation coefficient was approximately .26, which is the proportion of variance accounted for by group membership given the explanatory variable End.

The behavior of the Bayesian latent residuals for this data set were considered. The Bayesian latent residuals, the probabilities of a correct response, and the outlying probabilities, that is, the probabilities that the residuals were larger than 2, were estimated using formulae (5.4) , (5.5) , (5.9) , and (5.10). In Figure 5.1, the Bayesian latent residuals, ε_{ijk}, are plotted against the probabilities of a correct response of person ij to item k, p_{ijk}, and the outlying probabilities. The residuals were grouped by the value of y_{ijk}. If the answer was correct, $y_{ijk} = 1$, the Bayesian residual, ε_{ijk}, was negative, otherwise, it was positive. In general, it was possible that a correct answer corresponded with a negative residual and the other way around. Successes, $y_{ijk} = 1$, with fitted probabilities close to one and failures, $y_{ijk} = 0$, with fitted probabilities close to zero corresponded to small absolute values of the residuals. In these cases, the fitted probabilities agreed with the observed data. The outlying probability increased if the value of the residuals increased. The points with low fitted probabilities corresponding to correct answers and high fitted probabilities corresponding to incorrect answers were marked as outliers. Obviously, Figure 5.1 shows that there are a lot of outliers so the model doesn't fit the data very well.

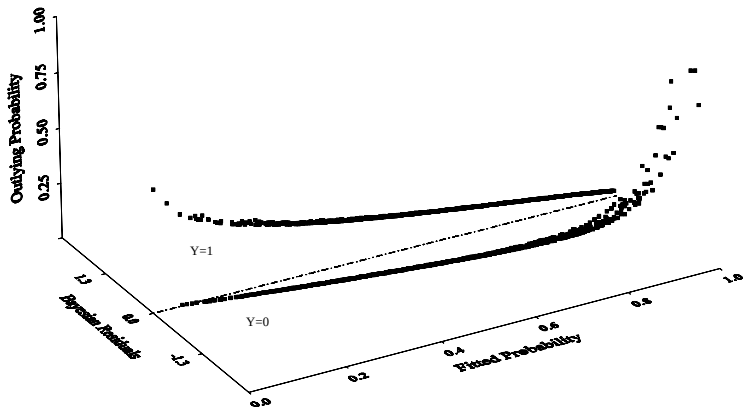

Figure 5.1. Bayesian latent residuals plotted against the probabilities of a correct response and the outlying probabilities.

The marginal prior of each residual, ε_{ijk}, was standard normal distributed. Fitted probabilities close to one corresponding to successes and fitted probabilities close to zero corresponding to failures had residuals that resembled the standard normal curves. However, the observations had large influence on the posterior distribution of the residuals when the fitted probabilities were in conflict with the observations. In Figure 5.2, posterior distributions of the residuals corresponding to Item 17 of the math test of several students are plotted. In this case, the Bayesian residuals were easily compared to each other because they were supposed to have the same spread. Some of the Bayesian residuals were marked as outliers because their posterior distributions differed from the standard normal distribution. That is, the conflict between the observations and the fitted probabilities was expressed in the nonzero location and the smaller standard deviation of the posterior distribution of these residuals. The outlying probability of the largest residual, in Figure 5.2, was .982. The corresponding response pattern showed that all items were scored correct except for Item 17 which was answered correctly by 88% of the students.

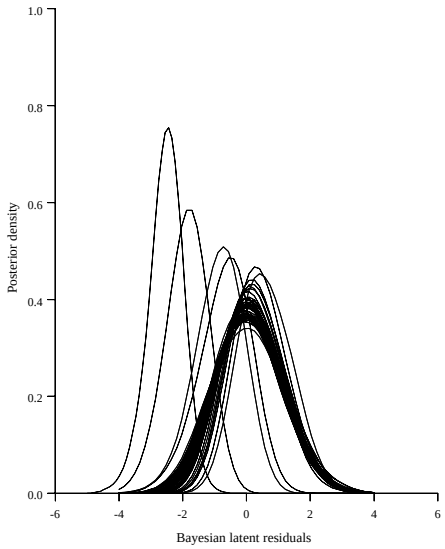

Figure 5.2. Posterior densities of the Bayesian latent residuals corresponding to Item 17 for a number of students.

It was assumed that the non-verbal intelligence test and the socio-economic status provide information about the achievements on a math test. These predictors should discriminate more between the students' achievements. Therefore, Model M_1, formula (5.24), was extended with these Level 1 predictors, that is,

$$\begin{aligned} \theta_{ij} &= \beta_{0j} + \beta_1 \mathrm{ISI}_{ij} + \beta_2 \mathrm{SES}_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} \mathrm{End}_j + u_{0j} \\ \beta_1 &= \gamma_{10} \\ \beta_2 &= \gamma_{20} \end{aligned} \tag{5.25}$$

where $e_{ij} \sim N\left(0, \sigma^2\right)$ and $u_{0j} \sim N\left(0, \tau_0^2\right)$. In the sequel, this model will be labeled M_2. Here, it was assumed that the effects of the scores of the intelligence test and the socio-economic status of the students did not differ per school, that is, the random regression coefficients were fixed over schools. The parameter estimates resulting from the Gibbs sampler are given in Table 5.2.

The residual variance at Level 1 was decreased due to the predictors at Level 1. The Level 1 residuals were easily estimated as a by-product

Table 5.2. Parameter estimates of a multilevel IRT model with explanatory variables ISI and SES on Level 1 and End on Level 2.

	IRT Model M_2		
Fixed Effects	Coefficient	s.d.	HPD
γ_{00}	$-.248$	.210	$[-.593, .094]$
γ_{01}	.348	.238	$[.047, .827]$
γ_{10}	.425	.030	$[.374, .471]$
γ_{20}	.225	.023	$[.187, .263]$
Random Effects	Variance Component	s.d.	HPD
σ^2	.380	.045	$[.294, .442]$
τ_0^2	.156	.038	$[.097, .212]$

of the Gibbs sampler, and were assumed to be normally distributed. The latent variable, $\boldsymbol{\theta}$, depends on the Level 1 and Level 2 residuals but also on the residuals, $\boldsymbol{\varepsilon}$, at the item level. It is impossible to consider these residuals separately. Except that the Level 1 residuals, $\mathbf{e}$, can be estimated such that they are unconfounded by the Level 2 residuals (Snijders & Bosker, 1999, pp. 128-132).

The Level 1 residuals were estimated within each group using only the Level 1 variables. This had the advantage that the estimates of the Level 1 residuals were no longer influenced by a Level 2 misspecification. In Figure 5.3, the posterior means of the standardized Level 1 residuals, $\mathbf{e}$, are plotted against the corresponding expected values of the standard normal distribution, according to the rank of $\mathbf{e}$. This was done with the estimated residuals at Level 1 confounded and unconfounded with the Level 2 residuals, denoted as posterior means with Level 2 and posterior means without Level 2, respectively. It was remarkable that the distributions of the residuals had smaller tails than the standard normal distribution. This indicated that the spread in the achievements of the students was rather small, although an item response model was used, instead of sum scores, to distinguish students' achievements better from each other.

6.1 Heteroscedasticity

The residuals at level 1 were assumed to have a constant variance, that is, they were assumed to be homoscedastic. It was investigated if

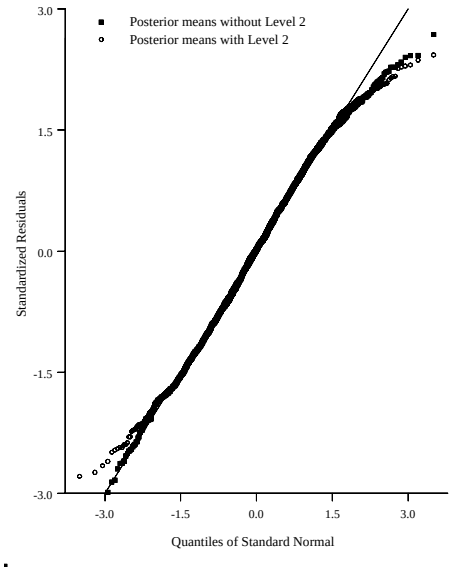

Figure 5.3. Normal probability plot of standardized residuals at Level 1.

the residual variance at Level 1 differed between male and female students. Both variances were randomly drawn under the assumption of equal variances. So, each group specific residual error variance was sampled during the parameter estimation of model (5.25), which assumed homoscedasticity at Level 1. In Figure 5.4, the top figure shows the posterior distribution of the group specific residual variance at Level 1 for both the male and the female group, respectively. It can be seen that the posterior means of the variances did not differ much. The 90% HPD region of the ratio of the two group specific residual variances was $[.84, 1.04]$. Thus the point of equal variances was included in the 90 per cent region. In Figure 5.4, the bottom figure shows the posterior distribution of the variance ratio and illustrates the 90% HPD region. This ratio consisted of the residual error variance within the male group divided by the residual variance within the female group. The posterior mean of the variance ratio was shifted towards the left of zero. Therefore, the residual variance within the female group was slightly, but not significantly, higher. The other test statistics, formula (5.13) and (5.20), were computed in every iteration of the Gibbs sampler. Both means of the computed test statistics corresponded with a p-value of .27. Therefore,

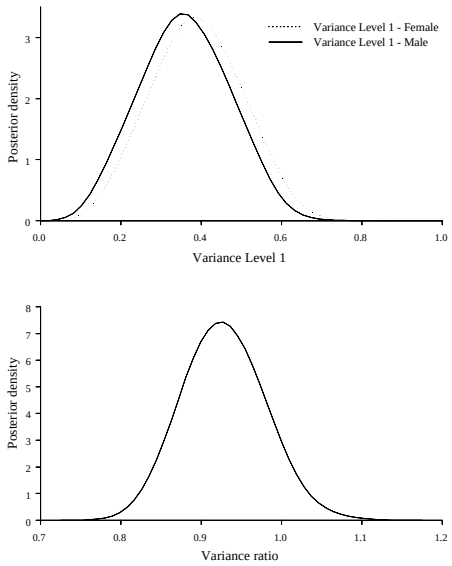

Figure 5.4. Testing heteroscedasticity at Level 1.

it can be concluded that there were no indications of residual variance differences between the male and female group at level 1.

6.2 The Choice of Priors

The use of noninformative prior distributions may lead to problems in estimating parameters and testing hypothesis. For example, testing hypotheses by a Bayes factor requires proper prior distributions on the parameters of interest (see, for example, Kass & Raftery, 1995). Further, the parameter estimates and test statistics can be sensitive to choice of priors. In estimation, the influence of the priors is small if the sample is fairly large. But in hypothesis testing the test statistic can be defined up to an undefined multiplicative constant leading to unjustified inference.

In this section, the sensitivity of the priors and the improperness of the posterior distributions are checked by computing the parameter estimates and the test statistics using proper informative priors. Proper inverse chi-square priors were used for both the variance components. To remain vague, the scale parameter and the degrees of freedom were both set equal to two. Further, a normal distribution with a large variance (standard deviation = 100) was used as prior for the difficulty parameters and the fixed effects. The prior distribution for the discrimination

Table 5.3. Parameter estimates of a multilevel IRT model with explanatory variables ISI and SES on Level 1 and End on Level 2 using proper informative priors.

	IRT Model M_3		
Fixed Effects	Coefficient	s.d.	HPD
γ_{00}	$-.305$	.216	$[-.669, .040]$
γ_{01}	.517	.242	$[.105, .902]$
γ_{10}	.500	.034	$[.442, .554]$
γ_{20}	.264	.027	$[.220, .307]$
Random Effects	Variance Component	s.d.	HPD
σ^2	.514	.063	$[.406, .605]$
τ_0^2	.211	.046	$[.140, .284]$

parameter was a log-normal distribution with mean zero and variance two. The Metropolis-Hastings algorithm was used to sample the discrimination parameters. Convergence of the Markov chain was established in the same way as in Chapter 3. In Table 5.3, the estimates of model M_3 are given, where model M_3 is equal to model M_2, but with proper priors. The scale of the latent variable $\boldsymbol{\theta}$ changed due to changes of the priors. Therefore, the parameter estimates in Table 5.3 differ from the estimates in Table 5.2, where the same model was estimated with noninformative priors. But the conclusions that can be drawn from both tables are much the same. The same parameters are significantly different from zero, and also the intraclass correlation coefficient remains almost the same. Also tests on heteroscedasticity led to the same conclusions. This indicates that, in this situation, the use of noninformative priors doesn't result to improper posteriors for the parameters of interest. The Bayes factor was well-defined using proper priors and resulted in a small favour of Model 3 in comparison to Model 1 with proper priors. Also, the logarithm of the marginal likelihood of the data, $f(\mathbf{y} \mid M_3)$, was increased, in comparison to Model 1 using proper priors. Finally, small changes in the prior specifications, that is, increasing the degrees of freedom, or changing the scale or mean of the prior distributions did not result in major differences in the parameter estimates or the Bayes factor.

7. Discussion

Above, methods for evaluation of the fit of a multilevel IRT model was discussed. It was shown that Bayesian latent residuals are easily estimated and particularly useful in case of dichotomous data. Estimates of these Bayesian latent residuals can be used to detect outliers. Moreover, outlying probabilities of the residuals are easily computed with the Gibbs sampler. Together, these estimates provide useful information regarding the fit of the model. One particular assumption of the multilevel IRT model is homoscedasticity at Level 1. Several tests are given to check this assumption. They can be computed as a by-product of the Gibbs sampler. Finally, the impact of prior distributions was discussed. The multilevel IRT model was easily estimated with informative priors and a Metropolis-hastings algorithm can be used to sample parameters from their conditional distributions. Estimating the model with proper priors justifies the use of the Bayes factor. This ratio can be used to test several hypothesis (Kass & Raftery, 1995; Pauler et al., 1999).

One class of tests to check the discrepancy between the model and the data are the so called posterior predictive checks, introduced by Rubin (1984). Posterior predictive checks consist of quantifying the extremeness of the observed value of a selected discrepancy. Several general discrepancies are developed but this can be any function of the data and the model parameters (Meng, 1994; Gelman et al., 1996). Obviously, these tests can be used to judge the fit of a multilevel IRT model. More research is required into the relation between the tests described in this chapter and posterior predictive checks.

The introduction of the latent data to connect the binary data to the continuous latent data has several advantages. The problem of estimating all parameters reduces to sampling from standard distributions, as can be seen in Chapter 3. The latent residuals provide information concerning the fit of the model and possible outliers are easily detected. These techniques can be extended to multilevel IRT models with latent variables in the dependent and independent variables, as described in Chapter 2. This simulation technique introduces extra randomness in the estimation procedure, therefore, establishing the convergence of the algorithm requires extra attention.

Chapter 6

A Stochastic EM Approach

Abstract An item response (IRT) model is used as a measurement error model for the dependent variable of a multilevel model. The dependent variable is latent but can be measured indirectly by using tests or questionnaires. The advantage of using latent scores as dependent variables of a multilevel model is that it offers the possibility of modeling response variation and measurement error and separating the influence of item difficulty and ability level. The two-parameter normal ogive model is used for the IRT model. It is shown that the stochastic EM (SEM) algorithm can be used to estimate the parameters which are close to the maximum likelihood estimates. This algorithm is easily implemented. The estimation procedure will be compared to an implementation of the Gibbs sampler in a Bayesian framework. Examples using real data are given.

Keywords: Bayes estimates, Data Augmentation, Gibbs sampler, item response theory, Markov chain Monte Carlo, multilevel model, stochastic EM, two-parameter normal ogive model.

1. Introduction

Many data in educational science have a hierarchical or clustered structure. For example, in schooling systems students are nested within schools. Information relevant to educational outcomes is inherently multilevel or hierarchical. In order to properly understand educational phenomena relevant to schooling, it is important to work with multilevel models that explicitly take this hierarchical organization into account. Therefore, multilevel analysis is a common way for properly analyzing

such data (Bryk & Raudenbush, 1992; Goldstein, 1995). Furthermore, multilevel analysis makes it possible to compare schools in terms of the achievements of their students and factors can be studied that explain school differences.

In Chapter 2, a multilevel IRT model is proposed to model such data and a latent variable is used as outcome in the multilevel analysis. This approach takes into account that, for example in school effectiveness research, the students' abilities are latent variables. A measurement error model is used to model this latent variable. Tests or questionnaires consisting of separate items are used to perform a measurement error analysis. This approach has the advantage that it is no longer assumed that the error component is independent of the outcome variable, i.e., the score of the test taker. Measurement error is defined locally as the variance of the ability parameter given a response pattern. This local definition of measurement error results in heteroscedasticity. An IRT approach to multilevel models gives a more realistic treatment of measurement error. Besides, contrary to observed scores, latent scores are test-independent, which offers the possibility of analyzing data from incomplete designs, such as, for instance, matrix-sampled educational assessments, where different (groups of) persons respond to different (sets of) items.

In the field of IRT models some applications of the multilevel model can be found. Adams, Wilson and Wu (1997) discuss the treatment of latent variables as outcomes in a regression analysis. They show that a regression model on latent proficiency variables can be viewed as a two-level model where the first level consists of the item response measurement model which serves as a within-student model and the second level consists of a model on the student population distribution, which serves as a between-students model. Further, Adams, Wilson and Wu (1997) show that this approach results in an appropriate treatment of measurement error in the dependent variable of the regression model. Raudenbush and Sampson (1999) embedded the Rasch model within a three-level hierarchical regression model, that is, the Level 1 model consists of the predictable and random variation among item responses within each group. Another application of multilevel modeling in the framework of IRT models was given by Mislevy and Bock (1989) where group-level and student-level effects are combined in an hierarchical IRT model. Finally, Patz and Junker (1999b) developed a generic hierarchical item response model which allow covariates on subjects and covariates on items.

In Chapter 3, a fully Bayesian estimation procedure is described, and within this procedure a Markov chain Monte Carlo method (Gibbs sam-

pler) is used for estimating all parameters. The fully conditional decomposition of Gelfand and Smith's (1990) Gibbs sampling produces an approximation for the posterior distributions of the parameters. That is, the Gibbs sampler is used to find the mode of the posterior distribution in a Bayesian framework, taking account of all sources of uncertainty in the estimation of the parameters. In the present paper, Bayes estimator will be compared to the maximum likelihood estimator which has attractive features, as good-large sample properties. More specific properties of maximum likelihood estimates can be found in, for example, Lehmann and Casella (1998) and Rao (1973). Besides, the likelihood of the sample of observations represented by the data is maximized without any prior knowledge regarding the parameters of interest.

The likelihood function is complex due to the absence of some part of the data. Maximizing the likelihood directly is often numerically infeasible. The idea is to associate with the given incomplete-data problem, a complete-data problem for which maximum likelihood estimation is feasible. That is, the problem of maximizing the likelihood is reformulated in such a way that the maximum likelihood estimates are more easily computed from a complete-data likelihood. The stochastic EM (SEM) algorithm is particularly appealing in situations where inference on complete-data is easy. The algorithm handles complex missing-data structures in which high-dimensional integrations may be involved. It imputes values for the missing data and then iteratively performs direct parametric inference based on the complete-data. This makes it attractive for estimating the multilevel IRT model with latent variables defined by a complex structural model. Moreover, the parameter estimates resulting from the algorithm are close to the maximum likelihood estimates. Further applications of the SEM algorithm can be found in, e.g., Celeux and Diebolt (1985), Celeux, Chauveau, & Diebolt, (1996), Diebolt and Ip (1996) and Ip (1994).

In the first section, the notation and a general multilevel IRT model is presented. Next, the principles of SEM and the implementation for estimating the parameters of a multilevel IRT model are described. Furthermore, a parallel will be drawn between parameter estimation with SEM and Markov chain Monte Carlo (Gibbs sampler). After that, a Dutch primary language test will be analyzed and the mentioned estimators will be compared. Finally, the last section contains a discussion and suggestions for further research.

2. A Multilevel IRT Model

This section contains the basic principles and formulae of a multilevel IRT model. For a detailed introduction of the model, see Chapter 2 and

3. In its general form, Level 1 of the two level multilevel model consists of a regression model, for each of J nesting Level 2 groups $(j = 1, \ldots, J)$, in which the $(n_j \times 1)$ ability vector $\boldsymbol{\theta}_j$ is modeled as a function of Q predictor variables $(\mathbf{X}_{1j}, \ldots, \mathbf{X}_{Qj})$:

$$\boldsymbol{\theta}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{e}_j, \tag{6.1}$$

where $\mathbf{e}_j$ is an $(n_j \times 1)$ vector of residuals, that are assumed to be normally distributed with mean $\mathbf{0}$ and variance $\sigma^2 \mathbf{I}_{n_j}$. All $Q + 1$ regression parameters, $\beta_{0j}, \ldots, \beta_{Qj}$, are treated as varying across Level 2, although it is possible to constrain the variation in one or more parameters to zero. The random regression parameters are treated as outcomes in a Level 2 model

$$\boldsymbol{\beta}_j = \mathbf{W}_j \boldsymbol{\gamma} + \mathbf{u}_j, \tag{6.2}$$

where $\mathbf{u}_j$ is a vector of random effects assumed normally distributed with mean zero and covariance $\mathbf{T}$, $\mathbf{W}_j$ is a matrix consisting of Level 2 characteristics and $\boldsymbol{\gamma}$ is a $(S \times 1)$ vector of fixed effects. This corresponds with the formulation of the multilevel IRT model in Chapter 3, formula (3.4) and (3.5).

Suppose each of $\sum_j n_j$ persons, labeled $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, are exposed to K items, labeled $k = 1, \ldots, K$. A binary response $Y_{ijk} = 1$ or 0 is recorded. Furthermore it is assumed that, conditionally on the item and population parameters, the response Y_{ijk}, with realization y_{ijk}, is an independent Bernoulli random variable, with probability of success $p_{ijk} = P(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k)$. The normal ogive model is used to model the p_{ijk}. This leads to,

$$p_{ijk} = \Phi(a_k \theta_{ij} - b_k), \tag{6.3}$$

where Φ denotes the standard normal cumulative distribution function. Below, the parameters of item k will also be denoted by $\boldsymbol{\xi}_k$, $\boldsymbol{\xi}_k = (a_k, b_k)^t$. Notice, the item difficulty is denoted by the usual choice b while regression coefficients are denoted by β. The two parameter model constitutes a discriminatory parameter a_k for each item $k = 1, \ldots, K$. The restriction $a_k > 0$, $k = 1, \ldots, K$, assure that a student, indexed ij, with a better ability θ_{ij} have a higher probability of getting the k^{th} item correct. To eliminate the effect of guessing in a multiple choice test another set of parameters, the guessing parameters, are introduced in the so called three parameter model. The probability that a student correctly answers an item, indexed k, is represented as the sum of the probabilities that the student guesses and gets the item correct, c_k, plus the probability that the student does not guess, $(1 - c_k)$, and gets the

item correct, $\Phi(a_k\theta_{ij} - b_k)$; that is,

$$P(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k, c_k) = c_k + (1 - c_k)\,\Phi(a_k\theta_{ij} - b_k)\,. \qquad (6.4)$$

An elaborate description of both models can be found in the pioneering work of Birnbaum (1968) and Lord (1980). Discussions and literature reviews are found in Johnson and Albert (1999) and van der Linden and Hambleton (1997).

Formulae (6.1) and (6.2) define the structural model and formula (6.3) or (6.4) the measurement model. Jointly, this defines a multilevel IRT model which will be estimated using SEM.

3. The SEM Algorithm

The EM (expectation-maximization) algorithm is a well-known approach for computing maximum likelihood estimation in a wide variety of situations (see, Dempster, Laird, & Rubin, 1977). Notably, many incomplete data problems can be handled with the EM algorithm. Also, latent variable models and random parameter models turn out to be solvable by EM when they are formulated as missing value problems. In spite of its many appealing features, the EM algorithm has several drawbacks. For example, it can converge to local maxima or saddle points of the log-likelihood function and its limiting position is often sensitive to starting values. In some models, the computation of the E-step involves high dimensional integrations. Therefore, the E-step can be computationally difficult.

SEM (Celeux & Diebolt, 1985) provides an alternative to EM. Particularly, in situations where inference based on complete data is easy, but also in cases where EM is intractable or where the E-step involves high dimensional integrations.

The basic idea underlying SEM is to impute missing data with plausible values and then update parameters on the basis of the complete-data. The SEM algorithm consists of two steps. The S-step generates a complete-data sample by drawing missing data, given the observed data and a current estimate of the parameters. At the M-step, the maximum likelihood estimate of the parameters is computed, based on the complete-data. The entire procedure is iterated a sufficient number of times.

Under specific conditions, the array of estimates corresponding to each draw of pseudo-complete data forms a Markov chain that converges to a stationary distribution (Ip, 1994). The mean of this stationary distribution is close to the maximum likelihood estimate and its variance reflects the information loss due to missing data (Diebolt & Ip, 1996).

4. Maximum Likelihood Estimation

Let $\mathbf{Y}$ be the observed random sample. The values of the Level 1 and Level 2 explanatory variables are known, denoted as, $\mathbf{X}$ and $\mathbf{W}$, respectively. The model has parameters $\boldsymbol{\theta}, \boldsymbol{\xi}$, Level 1 regression coefficients $\boldsymbol{\beta}$, Level 2 regression coefficients $\boldsymbol{\gamma}$ and variance components σ^2 and $\mathbf{T}$. The observed or incomplete-data likelihood of the parameters of interest is given by

$$l\left(\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}; \mathbf{y}\right) = \prod_j \int \left[\prod_{i|j} \int p\left(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right) g\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2\right) d\theta_{ij} \right] h\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right) d\boldsymbol{\beta}_j, \quad (6.5)$$

where $p\left(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right)$ is the IRT model, formula (6.3) , specifying the probability of the observing response pattern $\mathbf{y}_{ij}$ as a function of the ability parameter θ_{ij} and the item parameters $\boldsymbol{\xi}$. Further, $g\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2\right)$ is the density of θ_{ij} and $h\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right)$ is the density of $\boldsymbol{\beta}_j$. The marginal likelihood entails a multiple integral over θ_{ij} and β_j. Computation of two dimensional integrals suffices. An EM algorithm is easily implemented in case all discrimination parameters are equal, that is, in case the measurement error model is the Rasch model (Raudenbush & Sampson, 1999). The probability model is then a member of the regular exponential family of distributions. The lesser restrictive IRT model, where the discrimination parameters may differ per item, is widely applicable but estimating the parameters becomes more difficult. This problem of integration and maximization relates to the estimation of a random-effects model for ordinal data and to the bi-factor full information factor analysis model (Gibbons & Bock, 1987; Gibbons & Hedeker, 1992; Hedeker & Gibbons, 1994). Hedeker and Gibbons (1994) utilized a Gauss-Hermite quadrature to numerically integrate over the distribution of random effects. Fisher's method was used to provide the solution to the likelihood equation. The numerical integration is feasible in these problems. The solution can involve summation over a large number of points when the number of random effects is increased, this could affect the parameter estimates.

An alternative approach is the stochastic EM algorithm which can handle these problems and also further developments of the multilevel model to three or more levels and more complex IRT models, including a guessing parameter. The likelihood should be defined as a function of the complete-data in such a way that a simpler likelihood maximization could be performed if the complete-data were observed. Therefore, assume that there exists a continuous latent variable that underlies each

binary response. The latent variables θ_{ij} are related to the observed responses, Y_{ijk}, of a person, indexed ij, on an item, indexed k. This observation Y_{ijk} can be interpreted as an indicator that a continuous variable with normal density is above or below zero. This variable is denoted as Z_{ijk} with realization z_{ijk}, it follows that

$$Z_{ijk} = a_k\theta_{ij} - b_k + \varepsilon_{ijk}, \tag{6.6}$$

with $\varepsilon_{ijk} \sim N(0,1)$ and $Y_{ijk} = I(Z_{ijk} > 0)$. Here, $I(.)$ is an indicator variable taking the value one if its argument is true and zero otherwise. The latent variable structure yields a model that is equivalent to the normal ogive model. This approach follows the procedure of Albert (1992) and Johnson and Albert (1999). The complete-data likelihood is given by

$$l^c\left(\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}; \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}\right) = \prod_j \left[\prod_{i|j} p\left(\mathbf{z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right) g\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2\right)\right] h\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right), \tag{6.7}$$

where $p\left(\mathbf{z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right)$ represent the IRT model which is normally distributed according to formula (6.6). The maximization of (6.7) becomes easily, which will be shown below, due to the fact that the complete-data likelihood consists of a product of normal densities. In the exponential family case the stochastic EM estimates converge to the maximum likelihood estimates by $O(1/n)$ (Diebolt & Ip, 1995). It must be pointed out that the SEM algorithm provides only convergence in distribution and not a pointwise estimator, like EM. This can be obtained by averaging a sufficient number of successive iterations during the estimation procedure. The values generated by stochastic EM at the M-step, corresponding to each draw of the complete-data, form a Markov chain with a stationary distribution which is approximately centered at the maximum likelihood estimates. The sequence of points represents a set of good guesses, called the plausible region, with respect to various plausible values of the missing data. Usually, the mean of this stationary distribution is considered as an estimate for the parameters. But in the plausible region, the point with the largest observed log-likelihood could also be considered as an estimate for the parameters, this requires the extra effort of evaluating the observed log-likelihood in every iteration (Diebolt & Ip, 1995).

5. Implementation of the SEM Algorithm

The multilevel IRT model can be set up as a missing data problem by defining $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as unobserved variables. The main interest is estimat-

ing the item parameters, $\boldsymbol{\xi}$, the regression coefficients on Level 2, $\boldsymbol{\gamma}$, and the variance on Level 1 and Level 2, σ^2 and $\mathbf{T}$, respectively. The SEM procedure, for current values of the parameters $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, σ^2 and $\mathbf{T}$, completes the observed data by drawing pseudo-complete data, and then computes the maximum likelihood estimates of the parameters based on the completed data. The first step in implementing SEM is creating pseudo-complete data. Hence, samples from the joint distribution of $\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{Y}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}$ are required. Directly drawing a sample from this joint conditional distribution is difficult. It turns out to be easier to use the Gibbs sampler (e.g., see, Gelfand & Smith, 1990; Geman & Geman, 1984) to simulate independent draws from the joint conditional distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Therefore, introduce a continuous latent variable structure that underlies each binary response, formula (6.6). A sample from $\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}$ is obtained by drawing from the distributions $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\xi})$, $p(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2)$ and $p(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T})$. The proposed Gibbs sampler consists of three steps.

First, consider the distribution of $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\xi})$. This conditional distribution of the latent variables $\mathbf{Z}$ given $\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y}$ follows from formula (6.6). For the three parameter normal ogive model, formula (6.4), consider random variables V_{ijk} such that $V_{ijk} = 1$ if a student, indexed ij, knows the correct answer to item k and $V_{ijk} = 0$ if the student does not know the correct answer to item k. The variables Z_{ijk}, formula (6.6), are related to the variables V_{ijk}. That is, several cases arise depending on the value of Y_{ijk}. Suppose that $Y_{ijk} = 0$, then $V_{ijk} = 0$ and $Z_{ijk} < 0$. Next, if $Y_{ijk} = 1$ and $V_{ijk} = 0$, then $Z_{ijk} > 0$. Otherwise if $Y_{ijk} = 1$ and $Z_{ijk} < 0$, then $V_{ijk} = 1$. The Gibbs sampling procedure can be extended to obtain a sample from the distribution of the underlying dichotomous latent variables Z_{ijk} and V_{ijk} (Béguin, 2000; Johnson & Albert, 1999).

Second, the ability parameter $\boldsymbol{\theta}$, given pseudo-complete data $\mathbf{Z}$, and estimates of $(\boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2)$ are independent and distributed as a mixture of normal distributions. From (6.1) and (6.6) it follows that,

$$\begin{aligned} p(\theta_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2) &\propto p(\mathbf{z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi})\, p(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2) \\ &\propto \exp\left[\frac{-1}{2v}\left(\theta_{ij} - \widehat{\theta}_{ij}\right)^2\right] \exp\left[\frac{-1}{2\sigma^2}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2\right] \end{aligned}$$

with

$$\widehat{\theta}_{ij} = \frac{\sum_{k=1}^{K} a_k\,(z_{ijk} + b_k)}{\sum_{k=1}^{K} a_k^2},$$

and $v = \left(\sum_{k=1}^{K} a_k^2\right)^{-1}$. It follows directly from standard Bayesian results for normally distributed observations and a normal prior (e.g., see, Box

& Tiao, 1973; Lindley & Smith, 1972) that

$$\theta_{ij} \mid \mathbf{Z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2 \sim N\left(\frac{\widehat{\theta}_{ij}/v + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/v + 1/\sigma^2}, \frac{1}{1/v + 1/\sigma^2}\right). \tag{6.8}$$

Notice that the posterior mean is a composite estimator; as the sampling variance v of $\widehat{\theta}_{ij}$ increases, the relative amount of weight placed on the prior mean, $\mathbf{X}_{ij}\boldsymbol{\beta}_j$, increases.

Third, the fully conditional distribution of $\boldsymbol{\beta}_j$ entails a normal prior induced by the Level 2 model and normally distributed observations θ_{ij}, that is,

$$\begin{aligned} p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right) &\propto p\left(\boldsymbol{\theta}_j \mid \boldsymbol{\beta}_j, \sigma^2\right) p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right) \\ &\propto \exp\left(\frac{-1}{2\sigma^2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_j\right)^t \mathbf{X}_j^t \mathbf{X}_j \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_j\right)\right) \\ &\quad \exp\left(\frac{-1}{2}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}\right)\right) \end{aligned}$$

with $\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1}\mathbf{X}_j^t\boldsymbol{\theta}_j$. Thus

$$\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T} \sim N\left(\mathbf{D}\mathbf{d}, \mathbf{D}\right), \tag{6.9}$$

where $\Sigma_j = \sigma^2\left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1}$, $\mathbf{d} = \Sigma_j^{-1}\widehat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1}\mathbf{W}_j\boldsymbol{\gamma}$ and variance component $\mathbf{D} = \left(\Sigma_j^{-1} + \mathbf{T}^{-1}\right)^{-1}$. If $\mathbf{X}_j$, $j = 1, \ldots, J$, does not have a full column rank, $\mathbf{X}_j^t\mathbf{X}_j$ has no inverse and the least squares estimator, $\widehat{\boldsymbol{\beta}}_j$, is not the unique solution to the normal equations. Besides, if $\mathbf{X}_j^t\mathbf{X}_j$, when in the form of a correlation matrix, is not nearly a unit matrix, the least squares estimates are sensitive to errors. Estimates depending on a generalized inverse of $\mathbf{X}_j^t\mathbf{X}_j$ are not an estimator for $\boldsymbol{\beta}$ because it depends entirely on what generalized inverse is used in obtaining the estimator (Searle, 1971). Estimation of $\boldsymbol{\beta}_j$ based on the matrix $\left(\mathbf{X}_j^t\mathbf{X}_j + k\mathbf{I}_{Q+1}\right), k \geq 0$ rather than on $\mathbf{X}_j^t\mathbf{X}_j$ has been found to be a procedure that can help to circumvent the difficulties associated with the usual least squares estimates (Hoerl & Kennard, 1970).

At each step, the fully conditional distributions of $\mathbf{Z}$ and $\boldsymbol{\theta}$ are considered at the level of persons, samples are drawn for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. The regression coefficients on Level 1 are sampled for each group j. Eventually, an independent sample $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ is obtained after sufficient draws from the sequentially updated fully conditional distributions.

In case of normal components, a more efficient alternative of updating is a block Gibbs update (Gelman, Carlin, Stern, & Rubin, 1995; Hobert & Geyer, 1998; Roberts & Sahu, 1997). In that case, all of the normal components are updated simultaneously. In order to use this block Gibbs sampler, the density of $\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{Z}, \boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}$ is needed. Treat the regression on the regression parameters, $\boldsymbol{\beta}$, on Level 1 as $J(Q+1)$ prior 'data points'. The joint fully conditional distribution of $\boldsymbol{\theta}_j, \boldsymbol{\beta}_j$ can be deduced from the weighted linear regression of 'observations' $\mathbf{Z}_j^*$ on $\left(\boldsymbol{\theta}_j, \boldsymbol{\beta}_j\right)$, using 'explanatory variables' $\mathbf{X}_j^*$ and 'variance matrix' Σ_j^*, where

$$\begin{aligned} \mathbf{Z}_j^* &= \begin{bmatrix} \mathbf{z}_j + \mathbf{b} \\ \mathbf{0} \\ \mathbf{W}_j \boldsymbol{\gamma} \end{bmatrix}, \mathbf{X}_j^* = \begin{bmatrix} \mathbf{a} \otimes \mathbf{I}_{n_j} & \mathbf{0} \\ \mathbf{I}_{n_j} & -\mathbf{X}_j \\ \mathbf{0} & \mathbf{I}_{Q+1} \end{bmatrix}, \\ \Sigma_j^{*^{-1}} &= \begin{bmatrix} \mathbf{I}_{n_j K} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\mathbf{I}_{n_j} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix}. \end{aligned}$$

It follows that,

$$\left(\boldsymbol{\theta}_j, \boldsymbol{\beta}_j\right)^t \mid \mathbf{Z}_j, \boldsymbol{\xi}, \boldsymbol{\gamma}, \mathbf{T} \sim N\left(\left(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\beta}}_j\right)^t, \left(\mathbf{X}_j^{*^t} \Sigma_j^{*^{-1}} \mathbf{X}_j^*\right)^{-1}\right), \tag{6.10}$$

with

$$\left(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\beta}}_j\right)^t = \left(\mathbf{X}_j^{*^t} \Sigma_j^{*^{-1}} \mathbf{X}_j^*\right)^{-1} \mathbf{X}_j^{*^t} \Sigma_j^{*^{-1}} \mathbf{Z}_j^*.$$

The proposed Gibbs sampler samples successively from (6.6) and (6.10) until an independent sample $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ has been obtained. That is, until convergence of the Gibbs sampler has occurred. This completes the stochastic S-step of the SEM algorithm. The attained pseudo-complete data $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ is then used to estimate $\left(\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right)$. Therefore, the M-step entails computing the estimates of $\left(\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right)$.

Because the item-parameters depend only on the latent data $\mathbf{Z}$ and the ability parameters, $\boldsymbol{\theta}$, according to (6.6) , it follows that

$$\mathbf{Z}_k = \begin{bmatrix} \boldsymbol{\theta} & -\mathbf{1} \end{bmatrix} \boldsymbol{\xi}_k + \boldsymbol{\varepsilon}_k,$$

where $\mathbf{Z}_k = (Z_{11k}, \ldots, Z_{n_1 1k}, \ldots, Z_{n_J Jk})^t$ and $\boldsymbol{\varepsilon}_k = (\varepsilon_{11k}, \ldots, \varepsilon_{n_J Jk})^t$ is a random sample from $N(0,1)$. Therefore,

$$\widetilde{\boldsymbol{\xi}}_k = \left(\mathbf{H}^t \mathbf{H}\right)^{-1} \mathbf{H}^t \mathbf{z}_k, \tag{6.11}$$

with $\mathbf{H} = \begin{bmatrix} \boldsymbol{\theta} & -\mathbf{1} \end{bmatrix}$. The $\widetilde{\boldsymbol{\xi}}$ stands for an estimate of the item parameters based on the pseudo-complete data $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$. The estimate

exclusively based on the observed data are marked with a hat. The same notation will be used for the other parameters.

The estimator of the variance on Level 1, σ^2, follows directly from the regression of $\boldsymbol{\theta}$ on $\mathbf{X}$, with $\boldsymbol{\beta}$ as regression coefficients. Thus,

$$\widetilde{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2, \tag{6.12}$$

which is the maximum likelihood estimator of σ^2 given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

The Level 2 model for school j can be written as

$$\boldsymbol{\beta}_j = \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{u}_j, \tag{6.13}$$

with $E(\mathbf{u}_j) = 0$, $E\left(\mathbf{u}_j\mathbf{u}_j^t\right) = \mathbf{T}$. Because (6.13) is a normal linear model given regression coefficients $\boldsymbol{\beta}_j$ it follows that the generalized least squares estimator of $\boldsymbol{\gamma}$ is

$$\widetilde{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{J}\mathbf{W}_j^t\widetilde{\mathbf{T}}^{-1}\mathbf{W}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{W}_j^t\widetilde{\mathbf{T}}^{-1}\boldsymbol{\beta}_j. \tag{6.14}$$

Likewise it follows that the estimator of $\mathbf{T}$ is

$$\widetilde{\mathbf{T}} = \frac{1}{J}\sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\widetilde{\boldsymbol{\gamma}}\right)\left(\boldsymbol{\beta}_j - \mathbf{W}_j\widetilde{\boldsymbol{\gamma}}\right)^t. \tag{6.15}$$

Notice that an Iterative Generalized Least Squares algorithm (Goldstein, 1995) is needed to compute both estimates in formula (6.14) and (6.15).

In conclusion, the algorithm to estimate all parameters involves iterating two steps. At the S-step, the missing data are sampled, given the observed data and a current estimate of the parameters. Here the S-step is made up of formula (6.6) and (6.10). With use of the Gibbs sampler a pseudo-complete sample is drawn. At the M-step, the missing data are imputed to estimate all parameters, see formula (6.11), (6.12), (6.14) and (6.15).

Eventually, plausible values or estimates from the M-step, based on the augmented data from the S-step, are used in the estimation of the parameters of interest. Therefore, define the parameters of interest $\boldsymbol{\lambda} = \left(\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right)$. The array of points generated by SEM are a Markov chain, denoted by $\left\{\widetilde{\boldsymbol{\xi}}^{(m)}, \widetilde{\sigma}^{2(m)}, \widetilde{\boldsymbol{\gamma}}^{(m)}, \widetilde{\mathbf{T}}^{(m)}, m \in \mathbb{N}\right\} = \left\{\widetilde{\boldsymbol{\lambda}}^{(m)}, m \in \mathbb{N}\right\}$, where m denotes the iteration number. Under some conditions, (Ip, 1994), the sequence $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ is approximately stationary. That is, the

stationary distribution of $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ does not change as m takes on different values. As noted above, usually, the mean of the stationary distribution is considered as an estimate of $\boldsymbol{\lambda}$. That is, after a burn-in period of M_0 iterations,

$$\widehat{\boldsymbol{\lambda}} = \left(\widehat{\boldsymbol{\xi}}, \widehat{\sigma}^2, \widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{T}}\right) = \frac{1}{M - M_0} \sum_{m=M_0+1}^{M} \left(\widetilde{\boldsymbol{\xi}}^{(m)}, \widetilde{\sigma}^{2(m)}, \widetilde{\boldsymbol{\gamma}}^{(m)}, \widetilde{\mathbf{T}}^{(m)}\right). \tag{6.16}$$

Each step of the SEM algorithm incorporates a stochastic step, which prevents the sequence from being immobilized near a saddle point. Accordingly, SEM does not terminate in any stationary point.

As noted above, another estimator for the parameters can also be derived from the values in the plausible region, generated at each M-step. This estimator from the stochastic EM iterates is the point with the largest observed log-likelihood, formula(6.5),

$$\boldsymbol{\lambda}^* = \arg \max_{1 \leq m \leq M} l\left(\boldsymbol{\lambda} \mid \mathbf{y}\right). \tag{6.17}$$

Obtaining this point requires the calculation of this incomplete log-likelihood in every iteration of the stochastic EM algorithm. Gauss-Hermite quadrature can be used to carry out the integration over the parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$. It is also possible to compute the incomplete likelihood via the expected complete likelihood, that is,

$$l\left(\boldsymbol{\lambda} \mid \mathbf{y}\right) = E\left[l^c\left(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{Z}^*\right)\right] = \int_{\mathcal{Z}} l^c\left(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{z}^*\right) k\left(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda}\right) d\mathbf{z}^*, \tag{6.18}$$

where $\mathbf{Z}^*$ represent the augmented data $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ and $k\left(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda}\right)$ is the density of the missing data conditional on the observed data. In this case, computing $\boldsymbol{\lambda}^*$ via (6.18) involves a higher dimensional integration and is consequently computational more demanding. A rough method as Monte Carlo integration of (6.18) is rather difficult because it needs independent samples of the augmented data $\mathbf{Z}^*$ at every iteration. The point in the plausible region which maximizes the observed likelihood is an approximation of the actual maximum likelihood estimator related to the observed likelihood, formula (6.5). For a sufficient number of stochastic EM iterates, that is, for a sufficient number of points in the plausible region gets $\boldsymbol{\lambda}^*$ close to the maximum likelihood estimator. This point can also be used to check whether the stochastic EM estimator, $\widehat{\boldsymbol{\lambda}}$, approximates the maximum likelihood estimator of formula (6.5).

The variances of the estimators are estimated by the inverse of the observed information matrix evaluated at $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$, formula (6.16), or at

the point with the largest observed likelihood $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$, formula (6.17). The observed information matrix is easily computed using Louis identity which relates the observed-data likelihood and the complete-data likelihood (Louis, 1982), that is

$$-\frac{d^2 l\left(\boldsymbol{\lambda};\mathbf{y}\right)}{d\boldsymbol{\lambda} d\boldsymbol{\lambda}^t} = E_{\boldsymbol{\lambda}}\left[-\frac{d^2 l^c\left(\boldsymbol{\lambda};\mathbf{z}^*\right)}{d\boldsymbol{\lambda} d\boldsymbol{\lambda}^t} \mid \mathbf{y}\right] - \mathrm{cov}_{\boldsymbol{\lambda}}\left[-\frac{dl^c\left(\boldsymbol{\lambda};\mathbf{z}^*\right)}{d\boldsymbol{\lambda}} \mid \mathbf{y}\right], \quad (6.19)$$

where the expectation is taken with respect to $k\left(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda}\right)$. The right-hand side of (6.19) is computed with augmented data samples generated independently from $k\left(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda}\right)$ where $\boldsymbol{\lambda}$ is fixed at $\widehat{\boldsymbol{\lambda}}$ or $\boldsymbol{\lambda}^*$.

6. SEM in Comparison with the Gibbs Sampling Approach

It seems worthwhile to compare this implementation of SEM with a fully conditional decomposition of the Gelfand and Smith's (1990) Gibbs sampling, described in Fox and Glas (2001). Define the augmented data $\mathbf{Z}^* = (\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ and the parameters of interest as $\boldsymbol{\lambda}$. This Gibbs sampler generates samples from the following posterior distribution,

$$p\left(\boldsymbol{\lambda} \mid \mathbf{y}\right) = \int\int p\left(\boldsymbol{\lambda} \mid \mathbf{z}^*, \mathbf{y}\right) p\left(\mathbf{z}^* \mid \boldsymbol{\lambda}', \mathbf{y}\right) d\mathbf{z}^* p\left(\boldsymbol{\lambda}' \mid \mathbf{y}\right) d\boldsymbol{\lambda}'. \quad (6.20)$$

In fact, the described Gibbs sampler generates samples from the marginal posterior distributions of parameters $\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}$ and $\mathbf{T}$, including priors for the parameters. There are two natural estimates for $\boldsymbol{\lambda}$ following from formula (6.20) (see, Lehmann & Casella, 1998, pp. 257):

$$\widehat{\boldsymbol{\lambda}}_e = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\lambda}^{(m)} \quad (6.21)$$

$$\widehat{\boldsymbol{\lambda}}_m = \frac{1}{M}\sum_{m=1}^{M} E\left(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{z}^{*(m)}\right). \quad (6.22)$$

Here, $\widehat{\boldsymbol{\lambda}}_e$ is called the empirical estimator (Liu, Wong, & Kong, 1994). The estimator $\widehat{\boldsymbol{\lambda}}_m$ which is often easy to compute assuming that the conditional density $p\left(\boldsymbol{\lambda} \mid \mathbf{z}^*, \mathbf{y}\right)$ is simple, is called the mixture estimator. Finally, the following difference can be notified between these estimates. The SEM estimate, formula (6.16), and the mixture estimate resulting from the Gibbs sampler calculates the means of the expectations of the parameters given the pseudo-complete data, whereas the empirical estimate resulting from the Gibbs sampler calculates the means of

the marginal posterior distributions of the parameters. Liu et al. (1994) showed that the mixture estimator is always better in this situation, i.e., it has a smaller variance than the empirical estimator. That is, the mixture estimator has a smaller variance attributable to the Gibbs sampler in estimating the posterior mean. The posterior variances and credibility intervals are estimated from the sampled values obtained from the Gibbs sampler. Because the posterior density of $\boldsymbol{\lambda}$ given $\mathbf{Z}^*, \mathbf{Y}$ contains a prior for $\boldsymbol{\lambda}$ in formula (6.20), it follows that the mixture estimate, formula (6.22), differs from the SEM estimate, formula (6.16). Moreover, the differences between the sampling schemes will cause different estimates.

7. A Dutch Primary School Language Test

To compare the SEM algorithm with the MCMC algorithm, a dataset from a Dutch primary school language test was analyzed. A multilevel IRT model was estimated with the SEM algorithm and the Gibbs sampler. Furthermore, a comparison was made with a hierarchical model using observed scores.

This research project entailed in investigating whether schools that participate in the central primary school leaving test in the Netherlands on a regular basis perform better than schools that do not participate on a regular basis. The pupils of 97 schools were given a language test for grade 8 students. In this analysis, 24 items designed by the Netherlands National Institute for Educational Measurement (Cito) were used. These items were taken from a standardized Cito test used in most Dutch schools at grade 8, called the primary school leaving test. The total number of pupils for which data were available was 2156. Schools participating in the Cito test (72 schools) on a regular basis are called the Cito schools. The remaining 25 schools will be called the non-Cito schools.

Two students' characteristics were used as a predictor for the students' achievement: socio-economic status (SES) and a non-verbal intelligence test (ISI). The SES is based on four indicators: the education and occupation of the parents. Non-verbal intelligence was measured in grade 7 by using three parts of an intelligence test. The predictors ISI and SES were normally standardized. A predictor labeled End equaled 1 if the school participates in the school leaving test, and equals 0 if this is not the case. A complete description of the data can be found in (Doolaard, 1999, pp. 57). The same predictors were used as in the data-set of Chapter 3. But in this study the dependent variable corresponded to a language test instead of a math test.

The structural model used in the analysis is given by,

$$\begin{aligned}\theta_{ij} &= \beta_{0j} + \beta_1 \text{ISI}_{ij} + \beta_2 \text{SES}_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} \text{End}_j + u_{0j} \\ \beta_1 &= \gamma_{10} \\ \beta_2 &= \gamma_{20},\end{aligned} \tag{6.23}$$

where $e_{ij} \backsim N\left(0, \sigma^2\right)$ and $u_{0j} \backsim N\left(0, \tau^2\right)$. The two-parameter normal ogive model was used as the measurement model.

The following procedure was used to obtain initial estimates. Initial values of the item parameters were computed using Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). A distinct ability distribution was used for every subgroup j. Then the MCMC procedure by Albert (1992) for estimating the normal ogive model was run. As the Gibbs sampler had reached convergence the means of the sampled values of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi})$ were computed. An EM algorithm was used for estimating $\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right)$ with the $\widehat{\boldsymbol{\theta}}$ (see, for instance Bryk & Raudenbush, 1992).

The number of iterations necessary to reach convergence of the SEM algorithm cannot be evaluated simply in a general setting. For the Dutch primary leaving test described above, $5,000$ iterations were "enough" in the sense that after a burn-in period of $1,000$ iterations a substantial increase in the number of iterations did not perturb the values of ergodic averages. Additionally, at every iteration 25 Gibbs sampling steps were taken to generate an independent sample of the pseudo-complete data. The differences in the results were negligible when ranging these Gibbs sampling steps between 20 to 75. The fully conditional decomposition of Gibbs sampling as in Chapter 3 was run for $20,000$ iterations, with a burn-in period of $5,000$ iterations. Non-informative priors were used for the parameters in the Gibbs sampling implementation. A non-informative prior for the difficulty and discrimination parameter, insuring that each item will have a positive discrimination index, and assuming independence between the item difficulty and discrimination parameter leads to the simultaneous noninformative prior $p\left(\xi\right) \propto \prod_{k=1}^{K} I\left(a_k > 0\right)$. A uniform prior was placed on the fixed effects and on the variance components, that is, $p\left(\boldsymbol{\gamma}\right) \propto c$, $p\left(\sigma^2\right) \propto 1/\sigma^2$ and $p\left(\tau^2\right) \propto 1/\tau^2$.

First, the parameter estimates of the measurements model are considered, after that, the parameter estimates of the structural model and further implications of these estimates are considered.

In Table 6.1 and Table 6.2, the estimates of the item parameters resulting from the Gibbs sampler with the mixture estimator and the SEM algorithm are given. The SEM algorithm produces two estimators, the

Table 6.1. Parameter estimates of the discrimination parameter with SEM and the Gibbs sampler.

	SEM				Gibbs Sampler		
	mean		max				
Item	**a**	sd	**a**	sd	**a**	psd	CI
1	.856	.075	.816	.074	.784	.074	[0.646, 0.938]
2	.654	.066	.619	.064	.597	.061	[0.485, 0.724]
3	.928	.086	1.038	.085	.870	.096	[0.698, 1.073]
4	.668	.057	.631	.064	.628	.059	[0.520, 0.751]
5	1.158	.086	1.058	.087	1.089	.099	[0.906, 1.296]
6	1.190	.087	1.165	.085	1.097	.091	[0.927, 1.290]
7	.297	.052	.280	.056	.265	.042	[0.186, 0.351]
8	1.454	.072	1.445	.074	1.407	.122	[1.186, 1.663]
9	.968	.072	.894	.074	.911	.078	[0.767, 1.078]
10	.972	.073	.912	.072	.910	.078	[0.765, 1.073]
11	.927	.083	.845	.082	.845	.084	[0.691, 1.025]
12	1.019	.075	.981	.075	.960	.088	[0.796, 1.143]
13	.738	.060	.652	.061	.696	.064	[0.578, 0.830]
14	1.112	.076	1.047	.075	1.055	.092	[0.888, 1.250]
15	.746	.062	.681	.062	.698	.066	[0.575, 0.833]
16	.562	.055	.571	.053	.525	.053	[0.427, 0.632]
17	.685	.058	.641	.057	.647	.061	[0.533, 0.775]
18	1.042	.062	.964	.062	1.011	.087	[0.850, 1.195]
19	1.174	.083	1.050	.084	1.084	.107	[0.888, 1.304]
20	.977	.071	.884	.072	.914	.082	[0.764, 1.083]
21	.973	.080	.898	.080	.881	.075	[0.743, 1.037]
22	.955	.071	.909	.072	.893	.082	[0.741, 1.062]
23	1.113	.063	.982	.063	1.081	.089	[0.916, 1.265]
24	1	0	1	0	1	0	[1, 1]

mean of the stationary distribution, formula (6.16) , labeled under the column mean, and the point corresponding to the largest observed likelihood, formula (6.17) , labeled under the column max. The multilevel IRT model was identified by fixing two item-parameters, here, $a_{24} = 1$ and $b_{24} = 0$.

The columns labeled sd present the standard deviations of the estimates resulting from the SEM algorithm using Louis identity, formula (6.19). In this application, 100 samples of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ were obtained to compute the observed information matrix. Unlike the SEM estimates are the estimates resulting from the Gibbs sampler calculated in a Bayesian framework. Therefore the posterior standard deviations of the param-

Table 6.2. Parameter estimates of the difficulty parameter with SEM and the Gibbs sampler.

	SEM				Gibbs Sampler		
	mean		max				
Item	**b**	sd	**b**	sd	**b**	psd	CI
1	−.227	.049	−.257	.045	−.259	.044	[−.341, −.168]
2	−.169	.045	−.190	.046	−.197	.038	[−.266, −.119]
3	−.843	.048	−.836	.043	−.870	.051	[−.963, −.766]
4	.332	.042	.315	.042	.313	.040	[.241, .396]
5	−.281	.051	−.284	.052	−.312	.056	[−.414, −.195]
6	.708	.059	.733	.059	.663	.060	[.553, .790]
7	.475	.041	.444	.042	.458	.031	[.400, .521]
8	−.086	.048	−.072	.044	−.109	.069	[−.234, .035]
9	.481	.049	.468	.051	.455	.051	[.362, .560]
10	.100	.047	.080	.045	.073	.049	[−.016, .176]
11	−.451	.050	−.454	.050	−.487	.048	[−.574, −.388]
12	−.222	.048	−.207	.050	−.249	.051	[−.342, −.143]
13	.152	.041	.121	.041	.133	.042	[.056, .218]
14	.052	.049	.031	.049	.026	.055	[−.072, .142]
15	−.045	.043	−.078	.043	−.067	.041	[−.142, .020]
16	.216	.041	.233	.042	.198	.035	[.133, .271]
17	.243	.041	.223	.042	.226	.040	[.152, .309]
18	.160	.043	.126	.044	.147	.054	[.049, .259]
19	−.557	.052	−.591	.050	−.595	.056	[−.698, −.476]
20	−.124	.074	−.132	.068	−.154	.049	[−.244, −.053]
21	.289	.054	.259	.055	.244	.048	[.156, .346]
22	−.177	.046	−.212	.046	−.205	.048	[−.293, −.105]
23	.199	.043	.154	.043	.184	.055	[.083, .299]
24	0	0	0	0	0	0	[0, 0]

eters are denoted by psd. Further, the parameter estimates resulting from the Gibbs sampler are the posterior means. It can be seen that the SEM estimates of the item parameters are close to the mixture estimates resulting from the Gibbs sampler. Confidence intervals are used to compare the uncertainty about the parameter estimates in relation to the different estimators. The Bayesian analogue of a frequentist confidence interval is usually referred to as a credibility interval. In the Bayesian framework the central posterior credibility intervals are calculated as confidence regions for the parameters. The 95% central posterior credibility intervals are given under the column labeled CI. All SEM estimates are well within the computed central posterior credibility intervals. No-

Table 6.3. Parameter estimates of the multilevel model with the Gibbs sampler, stochastic EM, and HLM using sum scores.

	SEM				Gibbs Sampler			HLM	
	mean		max						
Fixed Effects	Par.	sd	Par.	sd	Par.	psd	CI	Par.	sd
γ_{00}	.334	.204	.349	.197	.327	.206	$[-.074, .729]$	.361	.044
γ_{01}	.262	.237	.273	.225	.277	.236	$[-.183, .740]$	.223	.051
γ_{10}	.184	.014	.196	.013	.194	.018	$[.160, .231]$	.156	.010
γ_{20}	.158	.014	.168	.014	.168	.017	$[.136, .204]$	.127	.011
Random Effects	Par.	sd	Par.	sd	Par.	psd	CI	Par.	
σ	.423	.020	.439	.021	.445	.027	$[.387, .506]$	.443	
τ	.223	.010	.216	.009	.294	.027	$[.222, .390]$	.191	

table, the posterior standard deviations are, in almost all cases, larger than the standard deviations related to the SEM estimates. More detailed information concerning this point will be provided later.

Table 6.3 presents the results of estimating the fixed effects and random components of the model with the Gibbs sampler and stochastic EM. The main result of the analysis is that conditionally on SES and ISI, the Cito schools perform better than the non-Cito schools. The fixed effect, γ_{01}, models the contribution of participating in the school leaving exam to the ability level of the students via its influence on the intercept β_{0j}. This intercept β_{0j} is defined as the expected achievement of a student in school j when controlling for SES and ISI. Thus a positive value of γ_{01} indicates a positive effect of participating in the school leaving exam to the students' abilities. Further, there is a highly significant association between the Level 1 predictors ISI and SES and the ability of the students. Obviously, students with high ISI and SES scores perform better than students with lower scores. The residual variance for the school-level, τ_0, is the variance of the achievement of students in school j, β_{0j}, around the grand mean, γ_{00}, when controlling for SES and ISI. Obviously, the use of a multilevel model is justified, because a substantial proportion of the variation in the outcome at the student level was between the schools.

The fixed and random effects are generally quite the same for the SEM and the Gibbs sampling estimates, except for the Level 2 variance, τ. The significant difference between the Level 2 variance-estimates results in different intraclass correlation coefficients. The proportion of variance in ability accounted for by group-membership, after controlling for the Level 1 predictor variables is .345 according to the SEM variance-estimates and .330 according to the SEM variance-estimates which maximizes the observed likelihood. This coefficient is .398 in case of the variance-estimates resulting from the Gibbs sampler. As an additional check the fixed effects and variance components are also estimated from the observed scores using HLM for windows (Bryk, Raudenbush, & Congdon, 1996). For comparative purposes, the unweighted sums of the item responses were rescaled such that their mean and variance were equal to the mean and variance of the posterior estimates of the ability parameters, respectively. The standard deviations of the HLM estimates are given under the column labeled sd. The estimate of the Level 2 variance component is smaller in the HLM analysis whereas the estimate of σ is almost similar in comparison to the other estimates. The intraclass correlation coefficient consisting of these variance components, is .301, which is smaller than the estimates of the intraclass correlation coefficient from the SEM approach. Furthermore, the estimates of the fixed effects are smaller except for the main effect, γ_{00}. In conclusion, the multilevel IRT analysis, estimated with the Gibbs sampler and SEM, measures a greater variance between students' abilities which results in a larger school-level effect. Further, a sharper distinction in students' achievements is attained.

The standard deviations of the SEM estimates are larger than the standard deviations of the estimates resulting from the analysis in HLM using observed scores. Obviously, the estimates resulting from HLM are based on the observed scores, which results in more accurate estimates, that is, the HLM analysis does not take the uncertainty of the ability parameter into account. It can be seen from Tables 6.1 to 6.3 that the standard deviations related to the stochastic EM estimates are smaller, in most cases, than the posterior standard deviations. This observation was also made in Chapter 3 and Glas, Wainer and Bradlow (2000). It seems that the smaller size of the standard deviations in the frequentist framework is related to the fact that they are based on an asymptotic approximation that does not take the skewness into account.

Finally, Figure 6.1 shows the plausible region of the variance components. The region of interest contains the parameter estimates of (σ, τ), obtained at every iteration of the stochastic EM algorithm. The most central point, that is the mean of (σ, τ), correspond to the stochastic

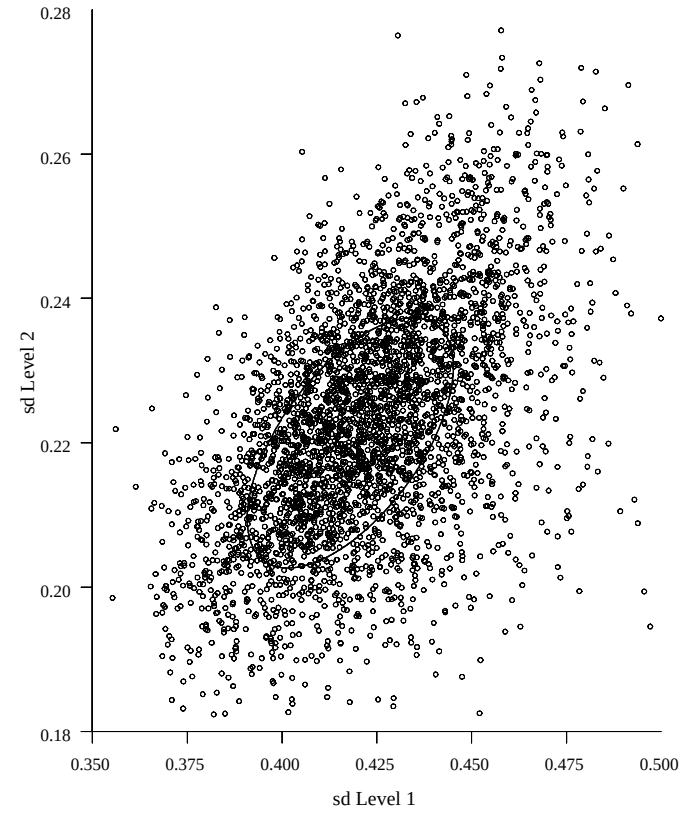

Figure 6.1. Plausible region for (σ, τ), generated by the stochastic EM algorithm.

EM estimate of (σ, τ), formula (6.16). The point with the largest observed log-likelihood, formula (6.17), lies in the circle close to the mean. The points within the circle represent estimates of (σ, τ) with high observed log-likelihood values, that is, the corresponding log-likelihood values are close to each other and therefore close to the highest observed log-likelihood. This illustrates the general idea behind stochastic EM. The parameters of interest are estimated by taking the mean over all points within the plausible region, where all points correspond to high observed log-likelihood values. As a result, this estimate lies close to the maximum likelihood estimate, which is checked by computing the observed log-likelihood at every iteration.

8. Discussion

In this chapter, a stochastic EM algorithm is used to estimate the parameters of a multilevel IRT model. As mentioned, the multilevel IRT model has several advantages, by treating the dependent ability parameters as latent variables in a multilevel model and using an IRT model to model these variables. Although direct parametric inference is hard because the likelihood function is very complex, maximum likelihood estimates can be obtained with the stochastic EM algorithm.

The use of a SEM algorithm for estimating the parameters of a multilevel IRT model has several appealing features. First, the algorithm is easy to implement. Second, although the amount of computation involved can be prohibitive, the SEM algorithm can handle the numerical integrations needed also in cases with more than two levels. Moreover, there are no limitations to the number of parameters or the number of explanatory variables. It must be remarked that MML or Bayes model estimation procedures are possible but require the calculation of two-dimensional integrals in the case of two levels. The implementation of the Gibbs sampler also has no limitations to the number of levels (Fox & Glas, 2001). Moreover, other measurement error models can be used to model the latent ability parameters.

The comparison with the Gibbs sampler showed that both methods estimate the parameters by sampling the missing data. SEM performs direct inference based on the pseudo-complete data whereas the Gibbs sampler samples the entire posterior distributions of the parameters. Both methods gave almost similar results. It must be pointed out that the differences between the standard deviations and the posterior standard deviations needs further research.

The convergence of this implementation of the algorithm is held up through the Gibbs sampling procedure for sampling the pseudo-complete data. The convergence is speeded up by the block Gibbs sampler, but a further improvement could be the use of another samplings-technique to sample all pseudo-complete data at once. General techniques for simulating draws directly from the target density as rejection sampling or importance sampling (Gelman et al., 1995) could improve the rate of convergence. Furthermore, the number of iterations needed to get a stable estimate could be reduced.

Appendix 6.A: Geometric Convergence of SEM

In this section, it will be shown that the convergence of the SEM algorithm, described above, depends on the convergence of two Markov chains. Because of the nested structure of the algorithm, it is shown that

the convergence of one Markov chain can be deduced from the convergence of another Markov chain, which results in the convergence of the SEM algorithm. First, the Markov chain formed by the imputed pseudo-complete data at each iteration is considered. Second, the Markov chain formed by the estimates of the parameters at each iteration will be observed. A consequence of this approach is that it can be shown that SEM is similar to Data Augmentation (Tanner & Wong, 1987).

As above, define $\mathbf{Z}^* = (\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$, $\left\{\mathbf{Z}^{*(m)}\right\} = \left\{\mathbf{Z}^{*(0)}, \mathbf{Z}^{*(1)}, \dots\right\}$ is a sequence of random variables with the Markov property, that is, the future of the process is independent of the past given only its present value. Therefore, $\left\{\mathbf{Z}^{*(m)}\right\}$ is a Markov chain containing the random variables $\left\{\mathbf{Z}^{*(m)}, m \in \mathbb{N}\right\}$ where $\mathbf{Z}^*$ takes on values $\mathbf{z}^*$, in $\mathbb{R}^n$. The probabilistic motion of the chain $\left\{\mathbf{Z}^{*(m)}\right\}$ is defined by a transition (probability) kernel. The Gibbs sampler, as described above, defines a stochastic process with transition kernel (Gibbs kernel), as defined in formulae (6.6), (6.8) and (6.9),

$$\begin{aligned} K\left(\mathbf{z}^{*(m)}, \mathbf{z}^{*(m+1)}\right) = & f\left(\mathbf{z}^{(m+1)} \mid \boldsymbol{\theta}^{(m)}, \widetilde{\boldsymbol{\xi}}\left(\mathbf{z}^{(m)}, \boldsymbol{\theta}^{(m)}\right), \mathbf{y}\right) \\ & f\left(\boldsymbol{\theta}^{(m+1)} \mid \mathbf{z}^{(m+1)}, \boldsymbol{\beta}^{(m)}, \widetilde{\sigma}^2\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\beta}^{(m)}\right)\right) \\ & f\left(\boldsymbol{\beta}^{(m+1)} \mid \boldsymbol{\theta}^{(m+1)}, \widetilde{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}^{(m)}\right), \widetilde{\mathbf{T}}\left(\boldsymbol{\beta}^{(m)}\right)\right), \quad (6.A.1) \end{aligned}$$

where $\widetilde{\boldsymbol{\xi}}\left(\mathbf{z}^{(m)}, \boldsymbol{\theta}^{(m)}\right)$ is the maximum likelihood estimate of $\boldsymbol{\xi}$ based on $\left(\mathbf{z}^{(m)}, \boldsymbol{\theta}^{(m)}\right)$, for the other parameters the same notation applies. In the sequel, the parameters $\left(\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right)$ are written as $\boldsymbol{\lambda}$.

Here, convergence of a chain is considered in terms of its transition probabilities. Convergence has occurred if the chain $\left\{\mathbf{Z}^{*(m)}\right\}$ has reached a stable or stationary state. That is, the strongest form of stability, the density of $\mathbf{Z}^{*(m)}$ does not change as m takes on different values. By definition there exists a target or stationary density of $\mathbf{Z}^*$ given the data $\mathbf{Y}$ and the parameters $\boldsymbol{\lambda}$. Clearly,

$$\pi\left(\mathbf{z}^* \mid \boldsymbol{\lambda}, \mathbf{y}\right) = f\left(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \mathbf{y}\right),$$

is the stationary density of the chain $\left\{\mathbf{Z}^{*(m)}\right\}$. This density is invariant because it satisfies

$$\pi\left(\mathbf{z}^* \mid \boldsymbol{\lambda}, \mathbf{y}\right) = \int K\left(\mathbf{x}, \mathbf{z}^*\right) \pi\left(\mathbf{x} \mid \boldsymbol{\lambda}, \mathbf{y}\right) d\mathbf{x}, \quad (6.A.2)$$

which stems from formula (6.*A*.1). This invariant density is important because it defines the stationary process, but it also defines the long

term or ergodic behavior of the chain. Furthermore, it turns out that the iterated kernels, formula (6.*A*.1) , converge to this invariant density. This can be seen as follows. The probability that the chain at $\mathbf{x}$ will be in the set $\mathbf{A}$ after m transitions is defined as,

$$P\left(\mathbf{Z}^{*(m)} \in \mathbf{A} \mid \mathbf{Z}^{*(0)} = \mathbf{x}\right) = \int_{\mathbf{A}} K^m(\mathbf{x}, \mathbf{z}^*)\, d\mathbf{z}^*. \tag{6.A.3}$$

Thus, after infinite many iterations,

$$\begin{aligned} P\left(\mathbf{Z}^* \in \mathbf{A} \mid \mathbf{Z}^{*(0)} = \mathbf{x}\right) &= \lim_{m\to\infty} \int_{\mathbf{A}} K^m(\mathbf{x}, \mathbf{z}^*)\, d\mathbf{z}^* \\ &= \lim_{m\to\infty} \int_{\mathbf{A}} \int K^{m-1}(\mathbf{x}, \mathbf{w})\, K(\mathbf{w}, \mathbf{z}^*)\, d\mathbf{w} d\mathbf{z}^* \\ &= \int_{\mathbf{A}} \int \pi(\mathbf{z}^* \mid \boldsymbol{\lambda}, \mathbf{y})\, K(\mathbf{w}, \mathbf{z}^*)\, d\mathbf{w} d\mathbf{z}^* \\ &= \int_{\mathbf{A}} \pi(\mathbf{z}^* \mid \boldsymbol{\lambda}, \mathbf{y})\, d\mathbf{z}^*. \end{aligned} \tag{6.A.4}$$

This follows from formula (6.*A*.2) and (6.*A*.3) . Here, it will be shown that not only the convergence of iterated kernels to the invariant distribution is guaranteed, but that also information can be given on the rate of convergence of the considered Markov chains. Therefore, two definitions will be given that provides information on the rate of convergence (see, e.g., Meyn & Tweedie, 1993; Tierney, 1994).

Definition 6-6.A.1 (*Geometric Ergodicity*) *A Markov chain, with a transition kernel* $K^m(\mathbf{x}, .)$ *and invariant distribution* π, *is geometrically ergodic if there exists a nonnegative real-valued function* C *and constant* $\rho > 1$ *such that*

$$\|K^m(\mathbf{x}, .) - \pi\|_{tv} \leq C(\mathbf{x})\, \rho^{-m}$$

for all $\mathbf{x}$.

Here, $\|.\|_{tv}$ is the total variation norm,

$$\|\mu\|_{tv} = \sup_{\mathbf{A}\subset\mathbb{R}^n} \mu(\mathbf{A}) - \inf_{\mathbf{A}\subset\mathbb{R}^n} \mu(\mathbf{A}),$$

which measures the difference between two probability distributions. The next definition gives a stronger form of ergodicity that comprehends geometric ergodicity.

Definition 6-6.A.2 (*Uniform Ergodicity*) *A Markov chain, with a transition kernel* $K^m(\mathbf{x}, .)$ *and invariant distribution* $\boldsymbol{\pi}$, *is uniformly*

ergodic if there exists a constant $C < \infty$ *and* $\rho > 1$ *such that*

$$\|K^m(\mathbf{x}, .) - \pi\|_1 \leq C\rho^{-m},$$

where $\|.\|_1$ *is defined as the* L^1*-norm,*$\|f\|_1 = \int |f(\theta)|\, d\theta$.

If the Markov chain is uniformly ergodic, the m-step transition probabilities converges, with a uniformly geometric rate to the stationary distribution of the Markov chain. Meyn and Tweedie (1993, pp. 395) gave conditions under which a Markov chain is uniformly ergodic. These conditions will be discussed here to prove that the Markov chain $\left\{\mathbf{Z}^{*(m)}\right\}$ is uniformly ergodic. All parts of the space can be reached by the Markov chain, no matter what starting point, as a result of the strict positivity of the Gaussian distributions given by formula $(6.A.1)$. Therefore, $\left\{\mathbf{Z}^{*(m)}\right\}$ with invariant distribution π is *irreducible*. There are no specific portions of the state space which can only be visited at certain regularly spaced times, thus the chain is *aperiodic*. Since $K\left(\mathbf{z}^{*(m)}, \mathbf{z}^{*(m+1)}\right)$ is also *continuous* and the $\mathbf{Z}^{*(m)}$'s take values in a *compact* space, $\left\{\mathbf{Z}^{*(m)}\right\}$ is uniformly ergodic. As a result, after a burn-in period, $\mathbf{Z}^{*(m)}$ is distributed according to the stationary density π.

The convergence of the SEM algorithm can be described as the iterative action of two dual Markov kernels (Diebolt & Robert, 1994). The first Markov chain is described above. The second Markov chain is formed by iterations of parameter estimates using the imputed data at each iteration. It will be shown that the second Markov chain converges because of the convergence of the first chain $\left\{\mathbf{Z}^{*(m)}\right\}$. That is, it will be shown that this second Markov chain is geometrically ergodic by means of the uniform ergodicity of the first Markov chain.

Firstly, the kernel and stationary density of this second chain are obtained. This is done using the principles of Data Augmentation. As a result, it is shown that this implementation of the SEM algorithm is a special case of the Data Augmentation algorithm. Secondly, the convergence of the second Markov chain is proven with use of the achieved formulae of the kernel and stationary distribution.

The SEM algorithm consists of sampling pseudo-complete data $\mathbf{Z}^* = (\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ given the data and an estimate of the parameters, say, $\widetilde{\boldsymbol{\lambda}}' = \left(\widetilde{\boldsymbol{\xi}}', \widetilde{\sigma}'^2, \widetilde{\boldsymbol{\gamma}}', \widetilde{\mathbf{T}}'\right)$, and estimating the parameters $\widetilde{\boldsymbol{\lambda}} = \left(\widetilde{\boldsymbol{\xi}}, \widetilde{\sigma}^2, \widetilde{\boldsymbol{\gamma}}, \widetilde{\mathbf{T}}\right)$. It is shown that given $\mathbf{Z}^* = (\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ it is easy to calculate $\widetilde{\boldsymbol{\lambda}}$, using formulae $(6.11), (6.12), (6.14)$ and (6.15). From this it can be seen that the iterative SEM algorithm induces samples of maximum likelihood estimates

with the following marginal distribution:

$$\Psi\left(\widetilde{\boldsymbol{\lambda}} \in \mathbf{A} \mid \mathbf{y}\right) = \int\int I\left(\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^*\right) \in \mathbf{A}\right) \pi\left(\mathbf{z}^* \mid \widetilde{\boldsymbol{\lambda}}', \mathbf{y}\right) d\mathbf{z}^* \Psi\left(\widetilde{\boldsymbol{\lambda}}' \mid \mathbf{y}\right) d\widetilde{\boldsymbol{\lambda}}' \tag{6.A.5}$$

where $I(.)$ is defined as the indicator function and $\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^*\right)$ represents the functions for estimating the parameters $\boldsymbol{\lambda}$ given the data and the imputed data. Specifically, from formulae (6.11), (6.12), (6.14) and (6.15), it can be seen that

$$\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^*\right) = E_{\Psi}\left[\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{z}^*\right], \tag{6.A.6}$$

where the expectation is taken with respect to the stationary distribution Ψ.

A Monte Carlo implementation to calculate the multiple integrals in formula (6.*A*.5) is provided by the Data Augmentation algorithm. To calculate $\Psi\left(\widetilde{\boldsymbol{\lambda}} \mid \mathbf{y}\right)$ using Monte Carlo integration:

1. Generate $\widetilde{\boldsymbol{\lambda}}_l' \backsim \Psi\left(\widetilde{\boldsymbol{\lambda}}' \mid \mathbf{y}\right)$, $l = 1, \ldots, L$.

2. Generate, for each $\widetilde{\boldsymbol{\lambda}}_l'$, $\mathbf{Z}_{lk}^* \backsim \pi\left(\mathbf{z}^* \mid \widetilde{\boldsymbol{\lambda}}_l', \mathbf{y}\right)$.

These two steps are alternated repeatedly, say K times. Eventually, calculate $\widehat{\Psi}\left(\widetilde{\boldsymbol{\lambda}} \in \mathbf{A} \mid \mathbf{y}\right) = \frac{1}{L}\sum_l \frac{1}{K}\sum_k I\left(\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^*\right) \mid \mathbf{y}, \mathbf{z}_{lk}^*\right)$. Step 1 is done by calculating $\widetilde{\boldsymbol{\lambda}}'$ given the imputed $\mathbf{z}^*$ and the data $\mathbf{y}$ using $\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^*\right)$ (see, Lehmann & Casella, 1998, pp. 291). Step 2 is done via the Gibbs sampler defined in the so-called S-step of the SEM algorithm. Notice, this entire procedure is the SEM algorithm described in the former section if $L = 1$. Furthermore, Tanner and Wong (1987) note that this algorithm will work even if $L = 1$. Therefore, this implementation of the SEM algorithm is proven to be an implementation of the Data Augmentation algorithm of Tanner and Wong (1987).

Formula (6.*A*.5) and (6.*A*.6) shows that $\Psi\left(\widetilde{\boldsymbol{\lambda}} \mid \mathbf{y}\right)$ is the invariant stationary density of the Markov chain $\left\{E_{\Psi}\left[\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{z}^{*(m)}\right]\right\} = \left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$,

$$\Psi\left(\widetilde{\boldsymbol{\lambda}} \mid \mathbf{y}\right) = \int K^*\left(\widetilde{\boldsymbol{\lambda}}', \widetilde{\boldsymbol{\lambda}}\right) \Psi\left(\widetilde{\boldsymbol{\lambda}}' \mid \mathbf{y}\right) d\widetilde{\boldsymbol{\lambda}}' \tag{6.A.7}$$

with kernel

$$K^*\left(\widetilde{\boldsymbol{\lambda}}', \widetilde{\boldsymbol{\lambda}}\right) = \int I\left(\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^*\right)\right) \pi\left(\mathbf{z}^* \mid \mathbf{y}, \widetilde{\boldsymbol{\lambda}}'\right) d\mathbf{z}^*. \tag{6.A.8}$$

This defines the motion of the chain $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$, that is, the probabilistic motion from $\widetilde{\boldsymbol{\lambda}}^{(m)} = \widetilde{\boldsymbol{\lambda}}'$ to state $\widetilde{\boldsymbol{\lambda}}^{(m+1)} = \widetilde{\boldsymbol{\lambda}}$. The estimator for $\boldsymbol{\lambda}$ is

$$\frac{1}{M}\sum_{m=1}^{M} E_{\Psi}\left[\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{z}^{*(m)}\right] = \frac{1}{M}\sum_{m=1}^{M} \widetilde{\boldsymbol{\lambda}}^{(m)} \underset{(M\to\infty)}{\longrightarrow} E_{\Psi}\left[\boldsymbol{\lambda} \mid \mathbf{y}\right] \text{ a.s.}, \tag{6.A.9}$$

if the Markov chain converges. Notice that the expectation is taken with respect to the stationary distribution Ψ. The described iteration scheme is needed to calculate the mean of this conditional expectation of $\boldsymbol{\lambda}$ given the pseudo-complete data. Clearly, the proposed estimator is equivalent to the SEM estimator, formula (6.16).

Under regularity conditions on $K^*\left(\widetilde{\boldsymbol{\lambda}}', \widetilde{\boldsymbol{\lambda}}\right)$ the Data Augmentation algorithm converges, (Tanner & Wong, 1987). In this section, a different approach is used to establish the convergence of the algorithm. As a result, a geometric convergence is derived, which guarantees fast convergence to the marginal distributions of the conditional expectation of the parameters given the pseudo-complete data.

The main properties of the chain $\left\{\mathbf{Z}^{*(m)}\right\}$ can be transferred to the chain $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ because of a duality principle (Diebolt & Robert, 1994). Instead of deducing properties of the Markov kernel formula $(6.A.8)$, to establish the convergence of $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$, as above, it appears to be more convenient to show the convergence of $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ through the convergence of $\left\{\mathbf{Z}^{*(m)}\right\}$. A duality principle relates the distributions of the two chains $\left\{\mathbf{Z}^{*(m)}\right\}$ and $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$;

$$\Psi\left(\widetilde{\boldsymbol{\lambda}} \in \mathbf{A} \mid \mathbf{y}\right) = \int I\left(\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^*\right) \in \mathbf{A}\right) \pi\left(\mathbf{z}^* \mid \widetilde{\boldsymbol{\lambda}}, \mathbf{y}\right) d\mathbf{z}^*. \tag{6.A.10}$$

The geometric convergence of $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ can be expressed in the following manner:

$$\begin{aligned}\left\|K^{*(m)}\left(\widetilde{\boldsymbol{\lambda}}', .\right) - \Psi\right\|_{tv} &= \sup_{\mathbf{A}}\left|\int I\left(\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y}, \mathbf{z}^* \in \mathbf{A}\right)\right)\right. \\ &\qquad \left.\left[K^{(m)}\left(\mathbf{x}, \mathbf{z}^*\right) - \pi\left(\mathbf{z}^* \mid \widetilde{\boldsymbol{\lambda}}', \mathbf{y}\right)\right]\right| d\mathbf{z}^*\end{aligned}$$

$$\begin{aligned}\left\|K^{*(m)}\left(\widetilde{\boldsymbol{\lambda}}',.\right)-\Psi\right\|_{tv} &\leq \sup_{\mathbf{A}}\int I\left(\widetilde{\boldsymbol{\lambda}}\left(\mathbf{y},\mathbf{z}^{*}\in\mathbf{A}\right)\right)\\ &\quad\left|K^{(m)}\left(\mathbf{x},\mathbf{z}^{*}\right)-\pi\left(\mathbf{z}^{*}\mid\widetilde{\boldsymbol{\lambda}}',\mathbf{y}\right)\right|d\mathbf{z}^{*}\\ &\leq\int\left|K^{(m)}\left(\mathbf{x},\mathbf{z}^{*}\right)-\pi\left(\mathbf{z}^{*}\mid\widetilde{\boldsymbol{\lambda}}',\mathbf{y}\right)\right|d\mathbf{z}^{*}\\ &=\left\|K^{m}\left(\mathbf{x},.\right)-\pi\right\|_{1}. \qquad (6.\text{A}.11)\end{aligned}$$

From formula (6.*A*.11) follows the geometric convergence of $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ because of the uniform convergence of $\left\{\mathbf{Z}^{*(m)}\right\}$. Also, Liu (1991) proved the geometric rate of convergence of the Data Augmentation algorithm under mild conditions. Finally, Diebolt and Robert (1994) showed the geometric rate of convergence of the Data Augmentation algorithm for estimation of finite mixture distributions with use of a duality principle.

Epilogue

Multilevel models are often used in the analysis of hierarchical structured data because dependencies between different levels are properly described without wasting any information. A proper specification of a model should also include the measurement error of the variables. For example, in school effectiveness research, students' abilities or degree of skills are analyzed in relation to school characteristics. Students' abilities and certain school characteristics cannot be observed directly and are measured using tests or questionnaires. These measurements cannot be made without an error. In this thesis, a model is introduced for dealing with measurement error in both the dependent and independent variables of a structural multilevel model.

A classical true score model and an item response theory model are proposed to model measurement error. The combination of a multilevel model with one or more latent variables modeled by a classical true score model or an item response theory model is called a multilevel true score model or a multilevel IRT model, respectively. In Chapter 2, the effects of measurement error on the estimation of the parameters of a multilevel model are analyzed using the multilevel true score model. It is shown that attenuated parameter estimates are obtained if the measurement error in the manifest variables is ignored. Modeling the measurement error by a classical true score model or an item response theory model leads to disattenuated parameter estimates. The effects of measurement error are also shown with a simulation study.

Shrinkage estimators are used to estimate the random regression coefficients. The shrinkage estimators are biased but have a mean squared error that is less than the mean squared error of the least squares estimator. Further research is needed to investigate the relationship between the amount of disattenuation and bias of the parameter estimates. Also, alternative ways of shrinkage could lead to estimators with a lower

mean squared error. For example, estimating a random regression coefficient of group j with shrinkage towards observed variables related to the characteristics of school j could be extended by shrinkage towards the observed variables related to the characteristics of all J schools. This double shrinkage estimator could lead to a lower mean squared error.

In Chapter 3, a Markov chain Monte Carlo estimation procedure is described to estimate the parameters of a structural multilevel model where a latent dependent variable is measured by the normal ogive model. It is shown that the evaluation of the multiple integrals needed to solve the estimation equations in an MML framework, are avoided by using the Gibbs sampler. Measurement errors are taken into account and prior knowledge could be incorporated to restrict parameters or to incorporate knowledge from previous research. The Gibbs sampler is easily implemented but requires a lot of iterations. Future research should focus on other sampling techniques or a more efficient implementation of the Gibbs sampler to reduce the amount of iterations. It should be investigated whether combining an efficient implementation of the block Gibbs sampler with the Metropolis-Hastings algorithm would lead to a reduction in the required number of iterations. That is, instead of sampling parameters from their conditional distributions, sample parameters from the simultaneous distribution using an efficient approximation, from which sampling is possible.

It is shown in Chapter 4 that the Gibbs sampler can also be used to estimate the structural multilevel model where some of the explanatory variables are modeled by an item response theory model or a classical true score model. The Bayesian formulation of the multilevel IRT model results in a straightforward model identification. The model is identified by fixing each latent ability scale. The multilevel true score model needs prior knowledge about the variance components for identification of the model. With a simulation study and a real data example it is shown that the required prior knowledge highly affects the parameter estimates. Further, the multilevel IRT model has a lower mean squared error than the multilevel true score model. Correcting for measurement error with the normal ogive model results in more variance of the parameter estimates but less bias and a better fit of the model. Using observed scores, instead of modeling the latent variables by an IRT model or a classical true score model, always lead to a lower mean squared error. More research is needed to explain the differences between the IRT model and the classical true score model as measurement models. This includes, specifying guidelines for choosing the classical true score model or an IRT model as a measurement model, and specifying conditions under which the IRT model results in a better fit of the model.

In the first chapters of the thesis it is assumed that the multilevel IRT model is correct. In Chapter 5, attention is focused on analyzing residuals, detecting outliers, testing heteroscedasticity at Level 1, and checking the sensitivity of the prior distributions. It is shown that Bayesian latent residuals have standard normal marginal distributions which can be used to assess the extremeness of the realized marginal distributions. Rao-Blackwellised estimators are derived for the Bayesian latent residuals at the item level. Also, tests for detecting outliers and for heteroscedasticity at Level 1 are easily computed as a by-product of the Gibbs sampler.

The multilevel IRT model is specified with proper priors to insure that the posterior distributions are proper. The comparison with the use of improper priors shows no differences. Further, small changes in the prior specifications do not result in major differences in the parameter estimates. As an advantage, the Bayes factor is well-defined using proper priors. In some cases, sampling from the conditional distributions is difficult when proper priors are used. Then, the Metropolis-Hastings algorithm is used to obtain samples from the conditional distributions. The convergence of the algorithm is highly depended on the proposal distributions. More research needs to be done about the specification of the parameters of the proposal distribution to establish an acceptable convergence rate.

Besides the Bayesian MCMC estimation procedure for estimating the parameters of a multilevel IRT model other procedures may result in parameter estimates with a smaller mean squared error. In Chapter 6, a stochastic EM algorithm is implemented resulting in estimates close to the maximum likelihood estimates. Further research could focus on the differences between the parameter estimates resulting from SEM, the Gibbs sampler, and estimates from other procedures like marginal maximum likelihood (MML) or Bayes modal estimation using Gauss-Hermite quadrature. In this comparison, the mean squared error of the different estimates, and the amount of computer time and prior knowledge should be taken into account to specify the advantages of the different estimation methods.

More research is needed to develop a statistical computer package for handling measurement errors. The Bayesian approach is computer intensive and requires an efficient implementation. But the statistical inference can be very misleading when the measurement error is ignored and the lack of programs impedes the use of modeling measurement error within a structural multilevel model.

Samenvatting

Data bezitten vaak een hiërarchische of geneste structuur. Voorbeelden hiervan zijn data afkomstig uit een enquete waarbij respondenten gekoppeld zijn aan een interviewer, longitudinale data waarbij meerdere observaties per individu beschikbaar zijn, en toets resultaten van studenten binnen klassen en scholen. Vanaf begin jaren tachtig is er een klasse van modellen ontwikkeld, de multiniveau modellen, die rekening houdt met de geneste structuur van de data en die de mogelijkheid biedt om verklarende variabelen te incorporeren op verschillende niveaus. Tevens is er specialiseerde software ontwikkeld voor het analyseren van multiniveau data met behulp van een multiniveau model.

Vooral in onderwijseffectiviteitsonderzoek wordt het multiniveau model veel gebruikt. Hiermee worden de effecten van het onderwijs op individuele leerprestaties bepaald, terwijl gecontroleerd wordt voor relevante achtergrondkenmerken van leerlingen, klassen en scholen. De verzamelde multiniveau data bestaan uit, onder andere, toets resultaten, achtergrondkenmerken, en school- en klaskenmerken. Bepaalde kenmerken zoals, socio-economische kenmerken en klasse-grootte zijn direct observeerbaar. Eigenschappen als leerprestaties en sociale vaardigheden zijn niet direct observeerbaar en worden aangeduid als latente variabelen. In praktijk, worden de niet direct observeerbare kenmerken geschat op basis van een aantal vragen, ook wel items genoemd. Hierbij worden de gemaakte meetfouten overigens vaak genegeerd.

In dit proefschrift wordt een nieuw model geïntroduceerd voor het analyseren van multiniveau data waarbij rekening wordt gehouden met meetfouten in geobserveerde afhankelijke en verklarende variabelen. Onderzocht wordt of de meetfouten invloed hebben op verdere analyses en daaruit voortvloeiende conclusies. Tevens wordt onderzocht of het gebruik van een meetmodel (een item response model c.q. een IRT model) er toe leidt dat de geschatte parameters gecorrigeerd worden voor meet-

fouten in geobserveerde variabelen. Er wordt een schattings-procedure ontwikkeld voor het simultaan schatten van alle parameters in het model. Verder wordt het IRT model vergeleken met het klassieke test theorie model als meetmodel. Hieronder volgt een samenvatting van de genoemde aspecten en gerapporteerde conclusies.

Een meetmodel is een model voor de relatie tussen de geobserveerde variabelen en de meting van een (latente) construct. Meetmodellen worden vooral gebruikt om een schatting te krijgen van de betrouwbaarheid van de meting. In de klassieke testtheorie onderscheidt men de geobserveerde en ware, niet direct observeerbare totaal score op een toets. Een nadeel van de klassieke testtheorie is dat de definitie en schatting van betrouwbaarheid meestal populatieafhankelijk is. Item response theorie stelt niet de toetsscore centraal maar de items en de antwoorden op de items. Een item response theorie model beschrijft de samenhang tussen de latente vaardigheid en het antwoordgedrag op een verzameling items. IRT komt tegemoet aan de eerder genoemde nadelen van klassieke testtheorie omdat betrouwbaarheid hier conditioneel op de waarde van de latente variabele gedefinieerd is, waardoor de schatting van de betrouwbaarheid van een individuele meting niet meer van de verdeling van de latente variabele hoeft af te hangen. Daarnaast heeft het als voordelen, onder andere, de scheiding van de item moeilijkheid en de latente variabele, en de toepasbaarheid in onvolledige designs. De combinatie van een multiniveau model met latente variabelen gemodelleerd met een item response theorie model wordt het multilevel IRT model genoemd. Het klassieke test theorie model als meetmodel resulteert in het multilevel ware score model.

In hoofdstuk twee wordt aangetoond dat het negeren van meetfouten in de geobserveerde afhankelijke en/of verklarende variabelen van een multiniveau model kan leiden tot een onzuiverheid in de parameter schattingen. Daarnaast wordt aangetoond dat de variantie in de afhankelijke variabele aanzienlijk toeneemt. Met behulp van een simulatie studie worden de effecten van meetfouten geïllustreerd. De geschatte parameters gecorrigeerd met behulp van de beide meetmodellen worden vergeleken met geschatte parameters op basis van som scores. Met name de schattingen van de variantie termen wijken van elkaar af. De geschatte parameters, gecorrigeerd met een twee parameter IRT model, liggen het dichtst bij de werkelijke parameters.

De intraklasse correlatie coëfficient is een maat voor de proportie variantie van de afhankelijke variabele verklaard door variabelen op groepsniveau. De verschillen in geschatte varianties leiden tot grote verschillen in schattingen van de intraklasse correlatie coëfficient. In hoofdstuk drie is een multilevel IRT model geanalyseerd waarbij de afhankelijke

latente variabele van een multiniveau model gemodelleerd is met een IRT model. Een simulatie studie laat onder meer zien dat de intraklasse correlatie coëfficient afwijkt van de gesimuleerde waarde wanneer geobseerveerde scores gebruikt worden. Met dit multilevel IRT model zijn tevens de uitkomsten van een CITO rekentoets geanalyseerd. De verklarende variabelen hebben een groter effect op de rekenvaardigheid wanneer de meetfouten gemodelleerd worden met een IRT model. De vaardigheden geschat met het multilevel IRT model discrimineren de leerlingen beter, hetgeen tevens leidt tot een groter groepseffect.

In hoofdstuk vier zijn een of meerdere verklarende variabelen in een multiniveau model gemodelleerd met een twee parameter IRT model en het klassieke test theorie model. De modellen zijn met behulp van een simulatie studie en een rekentoets met elkaar vergeleken. Daarbij is de gemiddelde kwadratische fout (MSE) tussen de geobserveerde en voorspelde scores geëvalueerd. Het multilevel IRT model met alle latente verklarende variabelen gemodelleerd met het twee parameter IRT model resulteert in de kleinste MSE waarde.

Het multilevel IRT model is eenvoudig te identificeren door de schaal van de latente variabelen vast te leggen. De identificatie van het multilevel ware score model vereist a priori informatie over de variantie van de meetfout. Deze variantie is moeilijk te schatten, en deze schatting heeft invloed op de schattingen van de overige parameters.

De standaard schattingsmethode voor item response theorie modellen is "Marginal Maximum Likelihood". Het simultaan schatten van alle parameters van een multilevel IRT model met deze methode is problematisch aangezien er veel meervoudige integralen uitgerekend moeten worden. Standaard methoden schieten hierin te kort, of kunnen slechts gedeeltelijke de klasse van multilevel IRT modellen schatten. Met behulp van recent ontwikkelde technieken, Markov chain Monte Carlo (MCMC), kunnen de parameters wel simultaan worden geschat. In hoofdstuk drie en vier worden implementaties gegeven waarbij latente variabelen in een multiniveau model gemodelleerd worden met het IRT model en met het klassieke test theorie model. Simulatie studies laten zien dat de parameters met beide meetmodellen behoorlijk nauwkeurig geschat kunnen worden. De schattingsmethode is flexibel en biedt mogelijkheden om andere meetmethoden te gebruiken waardoor een realistische manier van modelleren mogelijk is. Aan de andere kant is de methode tijdrovend, maar met de toenemende computer snelheid lijkt dit in de toekomst geen obstakel meer te zijn.

Met deze Bayesiaanse schattingsmethode worden de gemiddelden van de a posteriori verdelingen van de parameters gebruikt als schatters. In hoofdstuk zes wordt een andere schattingsprocedure gebruikt om de

meest aannemelijke schatter te bepalen. Deze methode levert vergelijkbare schattingen voor de parameters op.

Het multilevel IRT model bevat een aantal aannames die gecontroleerd moeten worden. Het toetsen van deze verschillende aannames van het multilevel IRT model vereist nog veel onderzoek. Hoofdstuk vijf laat een aantal aspecten zien maar is geenszins volledig. Er wordt aandacht besteed aan het analyseren van de residuen en de gebruikte priors, het bepalen van uitschieters, en het toetsen op heteroscedasticiteit. Bayesiaanse residuen hebben verschillende marginale verdelingen en zijn hierdoor moeilijk te vergelijken. Dit in tegenstelling tot de Bayesiaanse latente residuen die een standaard normale verdeling hebben en direct vergelijkbaar zijn. In hoofdstuk vijf zijn schatters voor deze latente residuen afgeleid. Tevens zijn formules afgeleid om uitschieters te kunnen identificeren. De residuen en uitschieters zijn eenvoudig te berekenen tijdens de schattingsprocedure. In hoofdstuk drie en vier zijn niet-informatieve priors gebruikt voor het analyseren van de data sets. Deze priors zijn vaak oneigenlijk, omdat de integraal over hun domein niet gelijk aan één is. Dit heeft als nadeel dat de gerelateerde a posteriori verdelingen ook oneigenlijk kunnen zijn. In hoofdstuk vijf is aangetoond dat het gebruik van niet-informatieve eigenlijke priors leidt tot nagenoeg dezelfde uitkomsten.

References

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

Aitkin, M., & Longford, N.T. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*, 1-43.

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

Albert, J.H., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika, 82*, 747-759.

Baltagi, B.H. (1995). *Econometric analysis of panel data.* Chichester: Wiley.

Béguin, A.A. (2000). *Robustness of equating high-stakes tests.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation of multidimensional IRT models. To appear in *Psychometrika.*

Bernardinelli, L., Pascutto, C., Best, N.G., & Gilks, W.R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine, 16*, 741-752.

Bernardo, J.M., & Smith, A.F.M. (1994). *Bayesian theory.* New York, NY: John Wiley & Sons, Inc.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord, & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading: Addison-Wesley.

Bock, R.D. (Ed.) (1989). *Multilevel analysis of educational data.* San Diego, CA: Academic Press, Inc.

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

Bollen, K.A. (1989). *Structural equations with latent variables.* New York, NY: John Wiley.

Bosker, R.J., Blatchford, P., & Meijnen, G.W. (1999). Enhancing educational excellence, equity and efficiency. In R.J. Bosker, B.P.M. Creemers & S. Stringfield (Eds.). *Evidence from evaluations of systems and schools in change* (pp. 89-112). Dordrecht/Boston/London: Kluwer Academic Publishers.

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

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

Brooks, S.P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434-455.

Bryk, A.S., & Raudenbush, S.W. (1987). Applying the hierarchical linear model to measurement of change problems. *Psychological Bulletin, 101*, 147-158.

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

Bryk, A.S., Raudenbush, S.W., & Congdon, R.T. (1996). *Hlm for Windows.* Chicago: Scientific Software International, Inc.

Buonaccorsi, J.P. (1991). Measurement errors, linear calibration and inferences for means. *Computational Statistics & Data Analysis, 11*, 239-257.

Buonaccorsi, J.P., & Tosteson, D. (1993). Correcting for nonlinear measurement errors in the dependent variable in the general linear model. *Communications in Statistics, Theory & Methods, 22*, 2687-2702.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8,* 158-233.

Carlin, B.P., & Louis, T.A. (1996). *Bayes and empirical Bayes methods for data analysis.* London: Chapman & Hall, Inc.

Carroll, R.J., Ruppert, D., & Stefanski, L.A. (1995). *Measurement error in nonlinear models.* London: Chapman & Hall.

Celeux, G., Chauveau, D., & Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation, 55*, 287-314.

Celeux, G., & Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly, 2*, 73-82.

Chaloner, K., & Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika, 75*, 651-659.

Chen, M.-H., & Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics, 8*, 69-92.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician, 49,* 327-335.

Cochran, W.G. (1968). Errors of measurement in statistics. *Technometrics, 10,* 637-666.

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

Cowles, M.K., & Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 833-904.

de Leeuw, J., & Kreft, I.G.G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics, 11*, 57-86.

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

Diebolt, J., & Ip, E.H.S. (1996). Stochastic EM: method and application. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 259-273). London: Chapman & Hall.

Diebolt, J., & Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B, 56*, 363-375.

Doolaard, S. (1999). *Schools in change or schools in chains.* 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.

Freedman, L.S., Carroll, R.J., & Wax, Y. (1991). Estimating the relation between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology, 134*, 310-320.

Fuller, W.A. (1987). *Measurement error models.* New York, NY: John Wiley & Sons, Inc.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., & Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*, 972-985.

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. (1995). Inference and monitoring convergence. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 131-143). London: Chapman & Hall.

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

Gelman, A., Meng, X.-L., & Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica 6*, 733-807.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

Gibbons, R.D., & Bock, R.D. (1987). Trend in correlated proportions. *Psychometrika, 52*, 113-124.

Gibbons, R.D., & Hedeker, D.R. (1992). Full-information bi-factor analysis. *Psychometrika, 57*, 423-463.

Gilks, W.R. (1996). Full conditional distributions. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 75-88). London: Chapman & Hall.

Glas, C.A.W., Wainer, H., & Bradlow E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (271-287). Boston MA: Kluwer Academic Publishers.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

Goldstein, H. (1989). Models for multilevel response variables with an application to growth curves. In R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 107-125). San Diego, CA: Academic Press, Inc.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8,* 369-395.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN.* London, Multilevel Models Project, Institute of Education, University of London.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff Publishing.

Hedeker, D.R., & Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50*, 933-944.

Hobert, J.P., & Geyer, C.J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis, 67*, 414-430.

Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55-67.

Hofman, R.H., & Bosker, R.J. (1999). *De schakels in weer samen naar school: een onderzoek naar de veronderstelde ketenstructuur van het WSNS-beleid.* De Lier: Academisch Boeken Centrum.

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

Hoijtink, H., & Molenaar, I.W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171-189.

Hox, J.J. (1995). *Applied multilevel analysis* (2nd ed.). Amsterdam: TT-Publikaties.

Hüttner, H.J.M., & van den Eeden, P. (1995). *The multilevel design: A guide with an annotated bibliography 1980-1993.* Westport: Greenwood Press.

Ip, E.H.S. (1994). A stochastic EM algorithm in the presence of missing data - theory and applications. Technical Report DMS 93-01366, Department of Statistics, Stanford University.

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

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Kass, R.E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*, 1343-1370.

Kreft, I.G.G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage Publications.

Lavine, M., & Schervish, M.J. (1999). Bayes factors: What they are and what they are not. *The American Statistician, 53*, 19-122.

Lehmann, E.L. (1986). *Testing statistical hypotheses (2nd ed.).* New York, NY: Springer-Verlag New York, Inc.

Lehmann, E.L., & Casella, G. (1998). *The theory of point estimation* (2nd ed.). New York, NY: Springer-Verlag New York, Inc.

Lindley, D.V., & Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34*, 1-41.

Liu, J. (1991). *Correlation structure and convergence rate of the Gibbs sampler.* Unpublished doctoral dissertation, University of Chicago, Chicago, IL.

Liu, J.S., Wong, H.W., & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika, 81*, 27-40.

Longford, N.T. (1990). *VARCL. Software for variance component analysis of data with nested random effects (maximum likelihood).* Princeton, NJ: Educational Testing Service.

Longford, N.T. (1993). *Random coefficient models.* New York, NY: Oxford University Press Inc.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

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

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.

MacEachern, S.N., & Berliner, L.M. (1994). Subsampling the Gibbs sampler. *The American Statistician, 48*, 188-190.

Mallick, B.K., & Gelfand, A.E. (1996). Semiparametric errors-in-vari ables models: A Bayesian approach. *Journal of Statistical Planning and Inference, 52*, 307-321.

Mason, W.M., Wong, G.Y., & Entwisle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 72-103). San Francisco: Jossey-Bass.

Mathsoft, Inc. (1999). *S-Plus 2000 Programmer's guide and Computer Program.* Seattle, WA: Data analysis products division, Mathsoft.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric monographs, 15.*

McDonald, R.P. (1982). Linear versus nonlinear models in latent trait theory. *Applied Psychological Measurement, 6*, 379-396.

McDonald, R.P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York, NY: Springer.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22,* 1142-1160.

Meyn, S.P., & Tweedie, R.L. (1993). *Markov chains and stochastic stability.* London: Springer-Verlag.

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

Mislevy, R.J., & Bock, R.D. (1989). A hierarchical item-response model for educational testing. In R.D. Bock (Eds.), *Multilevel analysis of educational data* (pp. 57-74). San Diego: Academic Press.

Morris, C.N. (1983). Parameteric empirical Bayes inference: Theory and applications (with discussion). *Journal of the American Statistical Association, 78*, 47-65.

Müller, P., & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika, 84*, pp 523-537.

Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Muthén, K.L., & Muthén, B.O. (1998). *Mplus. The comprehensive modeling program for applied researchers.* Los Angeles, CA: Muthén & Muthén.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B, 57*, 99-138.

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.

Pauler, D.K., Wakefield, J.C., & Kass, R.E. (1999). Theory and methods - Bayes factors and approximations for variance component models. *Journal of the American Statistical Association, 94*, 1242-1253.

Rao, C.R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13*, 85-116.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon, R.T., Jr. (2000). *HLM 5. Hierarchical linear and nonlinear modeling.* Lincolnwood, IL; Scientific Software International, Inc.

Raudenbush, S.W., & Sampson, R.J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Richardson, S. (1996). Measurement error. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 401-417). London: Chapman & Hall.

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

Roberts, G.O., & Sahu, S.K. (1997). Updating schemes, correlation structure, blocking and parametrization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B, 59*, 291-317.

Rosner, B., Willett, W.C., & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine, 8*, 1051-1069.

Rubin, D.B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6*, 377-400.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12,* 1151-1172.

Schaalje, G.B., & Butts, R.A. (1993). Some effects of ignoring correlated measurement errors in straight line regression and prediction. *Biometrics, 49,* 1262-1267.

Scheerens, J., & Bosker, R.J. (1997). *The foundations of educational effectiveness.* Oxford: Pergamon.

Searle, S.R. (1971). *Linear models.* New York, NY: John Wiley & Sons, Inc.

Seltzer, M.H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics, 18*, 207-235.

Seltzer, M.H., Wong, W.H., & Bryk, A.S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics, 21*, 131-167.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis.* London: Sage Publications Ltd.

Tanner, M.A., & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528-550.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics, 22*, 1701-1762.

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

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.

Wakefield, J., & Morris, S. (1999). Spatial dependence and errors-in-variables in environmentol epidemiology. In J.M. Bernardo, J.O.

Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 6* (pp. 657-684). New York, NY: Oxford University Press, Inc.

Wei, G.C.G., & Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's Data Augmentation algorithms. *Journal of the American Statistical Association, 85*, 699-704.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics.* New York, NY: John Wiley & Sons, Inc.

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.

Author Index

A

Adams, R.J., 30–31, 110
Aitkin, M., 3, 7, 31, 35, 40
Albert, J.H., 19, 31, 33–35, 40–41, 62, 64–65, 82, 86, 88, 91, 113, 115–116, 123

B

Baltagi, B.H., 2
Béguin, A.A., 31, 62, 82, 116
Berliner, L.M., 68
Bernardinelli, L., 4
Bernardo, J.M., 62, 75, 96
Best, N.G., 4
Birnbaum, A., 113
Blatchford, P., 5, 76
Bock, R.D., 2, 30–31, 35, 40, 67, 85, 98, 110, 114, 123
Bollen, K.A., 83
Boomsma, A., 30, 56
Bosker, R.J., 1, 5, 7, 57, 76, 91, 93, 104
Box, G.E.P., 15, 36, 64, 86–87, 94–95, 117
Bradlow, E.T., 31, 43, 127
Brant, R., 86, 91
Brooks, S.P., 13
Brown, W., 2
Bryk, A.S., 2, 30–33, 40, 42, 51, 57, 65, 67, 110, 123, 127
Buonaccorsi, J.P., 3–4
Burstein, L., 2
Butts, R.A., 19

C

Campbell, D.T., 56
Carlin, B.P., 14, 20, 74, 97
Carlin, J.B., 13, 31, 33, 51, 57, 74, 96–97, 108, 118, 129
Carroll, R.J., 3–4, 8–9, 14, 24, 56, 58–59, 82
Casella, G., 62–63, 67, 111, 121, 133
Celeux, G., 111, 113
Chaloner, K., 86, 91
Chauveau, D., 111
Chen, M.-H., 21, 68
Cheong, Y.F., 2, 67
Chib, S., 82, 86, 88, 91
Cochran, W.G., 16
Congdon, R.T., 2, 42, 67, 127
Cook, T.D., 56
Cowles, M.K., 14, 20

D

de Leeuw, J., 2, 30, 57
Dempster, A.P., 2, 113
Diebolt, J., 111, 113, 115, 132, 134–135
Doolaard, S., 46, 122
Draper, D., 2
Du, Z., 31

E

Entwisle, B., 2

F

Fox, J.-P., 62, 64–65, 121, 129
Freedman, L.S., 9
Fuller, W.A., 3, 14, 16, 56, 60

G

Gelfand, A.E., 4, 12, 33, 57, 62, 86, 89, 97, 111, 116, 121
Gelman, A., 13–14, 31, 33, 51–52, 57, 74, 96–97, 108, 118, 129
Geman, D., 4, 12, 33, 62, 116
Geman, S., 4, 12, 33, 62, 116
Geyer, C.J., 118
Gibbons, R.D., 53, 114
Gilks, W.R., 4, 99
Glas, C.A.W., 31, 43, 62, 64–65, 121, 127, 129

Goldstein, H., 2–3, 7, 30, 46, 56–57, 60, 85, 93, 110, 119
Greenberg, E., 82

H

Hambleton, R.K., 3, 10, 113
Healy, M., 2
Hedeker, D.R., 53, 114
Hills, S.E., 33, 57, 97
Hobert, J.P., 118
Hoerl, A.E., 117
Hofman, R.H., 5, 76
Hoijtink, H., 30–31, 56
Hox, J.J., 2
Hüttner, H.J.M., 2

I

Ip, E.H.S., 111, 113, 115, 119

J

Johnson, V.E., 19, 62, 82, 88, 91, 113, 115–116
Junker, B.W., 31, 41, 82, 99, 110

K

Kass, R.E., 87, 97, 99, 106, 108
Kennard, R.W., 117
Kong, A., 68, 121–122
Kreft, I.G.G., 2, 30, 57

L

Laird, N.M., 2, 113
Lavine, M., 87
Lehmann, E.L., 97, 111, 121, 133
Lindley, D.V., 15, 36, 64, 117
Liu, J., 135
Liu, J.S., 68, 121–122
Longford, N.T., 1–3, 7, 30
Lord, F.M., 3, 9–10, 21, 56, 59–60, 72, 113
Louis, T.A., 74, 97, 121

M

MacEachern, S.N., 68
Mallick, B.K., 4
Mason, W.M., 2
McDonald, R.P., 69
Meijnen, G.W., 5, 76
Meng, X.-L., 52, 74, 108
Meyn, S.P., 131–132
Mislevy, R.J., 30–31, 40, 51, 67, 98–99, 110, 123
Molenaar, I.W., 31
Morris, C.N., 51
Morris, S., 3–4
Müller, P., 3–4
Muraki, E., 40, 67, 98, 123
Muthén, B.O., 2, 83
Muthén, K.L., 2

N

Novick, M.R., 9–10, 21, 56, 59–60, 72

O

O'Hagan, A., 51, 87

P

Pascutto, C., 4
Patz, R.J., 31, 41, 82, 99, 110
Pauler, D.K., 87, 108
Plewis, I., 2

R

Racine-Poon, A., 33, 57, 97
Raftery, A.E., 106, 108
Rao, C.R., 111
Rasbash, J., 2
Raudenbush, S.W., 2, 30–33, 40, 42, 57, 67, 110, 114, 123, 127
Richardson, S., 4, 57
Roberts, G.O., 13, 39, 67, 118
Robert, C.P., 62–63, 67, 132, 134–135
Roeder, K., 3–4
Rosner, B., 9
Rubin, D.B., 2, 13, 31, 33, 51, 57, 74, 96–97, 108, 113, 118, 129
Ruppert, D., 3–4, 8, 14, 24, 56, 58–59, 82

S

Sahu, S.K., 13, 39, 67, 118
Sampson, R.J., 110, 114
Schaalje, G.B., 19
Scheerens, J., 7
Schervish, M.J., 87
Searle, S.R., 117
Seltzer, M.H., 31, 51–52, 65, 67, 97
Shao, Q.-M., 21, 68
Smith, A.F.M., 4, 12, 15, 33, 36, 57, 62, 64, 75, 86, 89, 96–97, 111, 116–117, 121
Snijders, T.A.B., 1, 57, 91, 93, 104
Spiegelman, D., 9
Stefanski, L.A., 3–4, 8, 14, 24, 56, 58–59, 82
Stern, H.S., 13, 31, 33, 51–52, 57, 74, 96–97, 108, 118, 129
Swaminathan, H., 10

T

Tanner, M.A., 51, 62, 130, 133–134
Tiao, G.C., 15, 36, 64, 86–87, 94–95, 117
Tierney, L., 62, 99, 131
Tosteson, D., 3–4
Tweedie, R.L., 131–132

V

van den Eeden, P., 2
van der Linden, W.J., 3, 10, 113

W

Wainer, H., 31, 43, 127
Wakefield, J., 3–4
Wakefield, J.C., 87, 108
Wang, H., 31
Wasserman, L., 97, 99
Wax, Y., 9
Wei, G.C.G., 51
Willett, W.C., 9
Wilson, M., 30–31, 110
Wong, G.Y., 2
Wong, H.W., 68, 121–122
Wong, W.H., 31, 51, 62, 65, 67, 130, 133–134
Woodhous, G., 2
Wu, M., 30–31, 110

Y

Yang, M., 2

Z

Zellner, A., 4, 57, 86–87, 91
Zimowski, M.F., 40, 67, 98, 123˙

Subject Index

A

Attenuation, **16**
Augmenting data, 34, **88**

B

Bartlett's approximation, 95
Bayes factors, 51, 87, 107–108
Bayes modal, 31, 52
Bayesian latent residuals, 5, 86, **88**–91, 101
Bayesian residual analysis, **87**–88

C

Classical residuals, 86–87
Classical test theory, 9, 59
  classical additive measurement error model, 3
  classical true score model, 4–5, 14–15, **59**–61
    prior information, 5, 66, 70, **80**
  group specific error variance, 5, 21, 60
  observed score, 9, 59
  propensity distribution, 9
  response distribution, 60
  true score, **9**–10, 59–61
Composite estimator, 17, 64, 117
Confidence regions, 41
Credibility intervals, 41

D

Data Augmentation algorithm, 62, 130, 132–134
Deviance adjusted residuals, 87
Deviance residuals, 87
Direct product, 37
  kronecker product, 37
  tensor product, 37

E

EM algorithm, 2, 40, 113

F

Fisher scoring algorithm, 2

G

Gauss-Hermite quadrature, 53, 114, 120
Generalized least squares
  algorithm, 119
  estimator, 38, 119

H

Heteroscedasticity, 30, 86, **93**, 104
Highest posterior density interval, 21, 73, 93
  HPD, 21, 73
  ratio of variances, 94, 105
  various scale parameters, 95
HLM for Windows, 2, 44, 127
Homoscedasticity, 3, 5, 30, 86

I

Intraclass correlation coefficient, **44**, 127
Inverse-chi-square distribution, 38, 98
Inverse-gamma distribution, 98
Inverse-Wishart distribution, 39, 98
Item response theory, 10, 61
  item parameters, 10
  latent ability, 10
  normal ogive model, **60**
  three parameter model, 112
Iteratively reweighted least squares, 2

L

Latent variable, 8, 58
Least squares estimator, 16, 37, 66

M

Manifest variable, 9, 14
Marginal likelihood, 107
Marginal maximum likelihood, 31, 52, 129
Marginal posterior distribution, 91
Markov chain Monte Carlo, **12**, 31, **62**, 88
  block Gibbs sampler, 13, 118, 129
  convergence, 13, 67
  Gibbs sampler, 5, **12**, 34, 62, 90, 116, 121
    burn-in period, 46
    confidence interval, **125**
    convergence, 20, **39**, 41, 75
    empirical estimator, **121**
    Gibbs kernel, 130
    initial values, 67
    mixture estimator, **121**
    starting values, **40**, 123
  MCMC, 4–5, 12, 31
Maximum likelihood, 6, 113–**114**, 120, 130
Measurement error, **2**–4
  correlated predictor variables, 68
  ignorable, 14
  measurement error model, 3–5, **9**, **59**, 123
    polytomously scored items, 82
    three-parameter IRT model, 82
  non-ignorable, 14
  nondifferential, 8, 59
  predictor variables, 55
  response error, 2, 5, 12, 14
  response variance, 2, 9, 56
Metropolis-Hasting, 51, 82
  algorithm, **99**
  within Gibbs sampling, 93
MLwiN, 2
Mplus, 2
Multilevel IRT model, 5, **10**, 20, **33**, 61, 82, 86, 110–111, 129
  analysis, 48
  convergence, 20, 41, 75, 108, 123
  full conditional distributions, 35
  full posterior distribution, 35
  identification, **12**, **35**, 72, 124
  initial values, 71
Multilevel model, **1–2**, 7, 32
  attenuation, 17–18, 23
Multilevel true score model, 14, 61, 82
  identification, 72
  initial values, 72

N

Normal ogive model, 10, 33, 112

O

Observed variables, 59
  manifest, 59
  proxies, 59
  surrogate, 59
Outlier, 5, 86
  detection, 86, **91**
  outlying probability, 92, 108

P

Pearson residuals, 87
Posterior distribution
  improper, 97
  normal approximation, 96
Posterior predictive checks, 82, 108
Posterior predictive data, 51, 74
  distributions, 75
Posterior variance, 41
Prior distribution, 5
  conjugate, 97
  improper, 87, 97
  informative prior, 97
  noninformative prior, 35, 97
    fixed effects, 123
    improper, 97
    item parameters, 35
    Jeffreys' prior, 51
    vague prior, 97
    variance components, 123
  proper, 97, 108
  sensitivity, 86, 106

R

Rao-Blackwellised estimator, 89
Realized residuals, 87, 90

S

School effectiveness, 7
Shrinkage estimator, 17, 64
Stochastic EM, 5–6, 111, **113**, 116
  algorithm, 120
  burn-in period, 123
  complete-data, 113, 116
  convergence, 123, 129
  dual Markov kernels, 132
  geometric convergence, 134
  Markov chain, 119, 130
    aperiodic, 132
    ergodic behavior, 131
    geometric ergodic, **131**
    invariant, 130
    irreducible, 132
    rate of convergence, 131
    stationary density, 130
    stationary distribution, 113, 120

transition kernel, 130
uniform ergodic, **132**
maximum likelihood, 120
pseudo-complete data, 134
uniform convergence, 135
Student's t-distribution, 97
Surrogate, 8

T

Total variation norm, 131

V

VARCL, 2˙