BAYESIAN ESTIMATION OF LATENT TRAIT DISTRIBUTIONS CONSIDERING HIERARCHICAL STRUCTURES AND PARTIALLY MISSING COVARIATE DATA

Inaugural-Dissertation

zur Erlangung des akademischen Grades
doctor rerum politicarum
an der Fakultät Sozial- und Wirtschaftswissenschaften der
Otto-Friedrich-Universität Bamberg

vorgelegt von
Diplom-Soziologe Jean-Christoph Gaasch
aus Gerolstein

Bamberg, den 09. Januar 2017

Promotionskommission:

- Prof. Dr. Susanne Rässler (Erstbetreuerin und Sprecherin)
  Otto-Friedrich-Universität Bamberg
  Fakultät Sozial- und Wirtschaftswissenschaften
- Prof. Dr. Claus H. Carstensen (Zweitgutachter)
  Otto-Friedrich-Universität Bamberg
  Fakultät Humanwissenschaften
- Prof. Dr. Thomas Saalfeld
  Otto-Friedrich-Universität Bamberg
  Fakultät Sozial- und Wirtschaftswissenschaften

Tag der mündlichen Prüfung: 12. Oktober 2017

URN: urn:nbn:de:bvb:473-opus4-502904
DOI: http://dx.doi.org/10.20378/irbo-50290

Für meine Eltern

Danksagung

Mein erster Dank gebührt Susanne Rässler, die mir sowohl den fachlichen Impuls als auch die berufliche Möglichkeit für diese Dissertation gab. Insbesondere danke ich ihr für die verständnisvolle Hilfsbereitschaft, mit der sie mich während des gesamten Zeitraumes, der zur Erstellung dieser Arbeit nötig war, betreut und unterstützt hat. Claus Carstensen möchte ich sehr herzlich dafür danken, dass er mir die Welt der empirischen Bildungsforschung nähergebracht und dadurch zahlreiche Anknüpfungspunkte für meine Arbeit geschaffen hat. Ein ganz besonderer Dank gilt Christian Aßmann, den ich jederzeit in Fragen der mathematischen Statistik zu Rate ziehen konnte und der in allen Phasen der Promotion ein offenes Ohr für meine Anliegen hatte.

Erklärung

Hiermit erkläre ich, dass ich die von mir vorgelegte Dissertation mit dem Titel „Bayesian estimation of latent trait distributions considering hierarchical structures and partially missing covariate data“ selbstständig verfasst und keine anderen als die im Quellen- und Literaturverzeichnis genannten Hilfsmittel benutzt habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet. Ich versichere weiterhin, dass ich die beigefügte Dissertation nicht bereits einer anderen Prüfungsbehörde zur Erlangung des Doktorgrades vorgelegt habe und, dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind. Darüber hinaus sei darauf verwiesen, dass sich Inhalte des Kapitels 4 sinngemäß bereits in folgenden Publikationen finden:

Aßmann, C., Gaasch, J.-C., Pohl, S., & Carstensen, C. H. (2015). Bayesian estimation in irt models with missing values in background variables. *Psychological Test and Assessment Modeling, 54* (4), 595-618.

Aßmann, C., Gaasch, J.-C., Pohl, S., & Carstensen, C. H. (2016). Estimation of plausible values considering partially missing background information: A data augmented mcmc approach. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological issues of longitudinal surveys: The example of the national educational panel study* (p. 503-521). Wiesbaden: Springer VS.

(Dipl.-Soz. Jean-Christoph Gaasch)

Zusammenfassung

Groß angelegte sozialwissenschaftliche Studien beinhalten häufig Indikatoren zur Messung eines gemeinsamen Faktors. Fast immer lautet das Ziel anschließender Auswertungen den Zusammenhang zwischen dem gemessenen Konstrukt und weiteren manifesten Variablen zu untersuchen. Daraus ergibt sich eine Messproblematik in dreierlei Hinsicht: Erstens können Fehler durch das jeweilige Erhebungsinstrument für die latente Variable entstehen. Zweitens weisen solche Studien typischerweise hierarchische Datenstrukturen auf, sei es aufgrund des Stichprobendesigns oder des Vorhandenseins korrelierter Beobachtungen im Allgemeinen. Drittens besteht Unsicherheit in Bezug auf fehlende Werte in den Kovariaten, die in Beziehung zu dem Faktor stehen. In dieser Arbeit werden unter Rückgriff auf die Bayesianischen Statistik alle drei Probleme gleichzeitig angegangen. Ich beginne mit der Modellklasse der latenten Regression, welche eine Verbindung von Messmodell und struktureller Analyse herstellt, und entwickle einen neuen Schätzalgorithmus auf Grundlage des Verfahrens der Datenerweiterung. Es können sowohl binäre als auch ordinale Indikatoren in die Schätzung einfließen. Verschiedene Formen der Populationsheterogenität werden berücksichtigt durch die Spezifikation von Multigruppen-, finiten Mischverteilungs- oder zufälligen Interzept-Modellen. Züge aus den Posteriorverteilungen der Modellparameter werden ergänzt durch Züge aus den vollständig bedingten Dichten der fehlenden Werte in Personenkovariaten. Die Verteilungen der fehlenden Werte werden dabei approximiert mittels Klassifikations- und Regressionsbäumen, welche es erlauben sowohl metrische als auch kategoriale Kovariaten zu imputieren und mit nichtlinearen Zusammenhängen umzugehen. Zwei Simulationsstudien mit unterschiedlichem Mechanismus der fehlenden Werte überprüfen die Gültigkeit der vorgeschlagenen Methode. Es zeigt sich, dass der neue Algorithmus fähig ist alle involvierten Parameter in jedem der zwei Szenarien korrekt zu schätzen und besser abschneidet als die stochastische Regressionsimputation und das Eliminierungsverfahren. Die Ergebnisse zweier Beispielanalysen mit Daten des Nationalen Bildungspanels zu Mathematikkompetenz und Essstörungen bei Neuntklässlern demonstrieren die Nützlichkeit der Methode in der empirischen Anwendung. Schließlich werde ich ein R Paket vorstellen, das die im Rahmen dieser Dissertation entwickelten Schätzroutinen bereitstellt.

Schlüsselwörter: Probabilistische Testtheorie, Populationsheterogenität, Markovkette Monte Carlo, multiple Imputationsverfahren, Entscheidungsbäume, statistische Programmierung, R, Nationales Bildungspanel.

Abstract

Large-scale studies in social sciences often involve the measurement of latent constructs and seek to investigate their relationship with additional variables in subsequent analyses. Within this context the analyst has to face three problems: First, there is uncertainty through the particular indicators which measure the trait of interest. Second, large-scale studies typically exhibit hierarchical structures caused by sampling design or a composite population consisting of clustered observations. Third, uncertainty arises due to the presence of missing values in covariates related to the latent construct. This thesis provides a Bayesian estimation strategy that simultaneously addresses all three issues. I start out with the class of latent regression item response models, which combine the fields of measurement models and structural analysis, and develop a novel algorithm based on the device of data augmentation. Binary and ordered polytomous items can both be included in the analysis. Population heterogeneity is taken into account either through multigroup, finite mixture or random intercept specifications. Sampling from the posterior distribution of parameters is enriched by sampling from the full conditional distributions of missing values in person covariates. Approximations for the distributions of missing values are constructed from classification and regression trees, thus allowing for high flexibility in the incorporation of metric as well as categorical variables and nonlinear relationships. The validity of the proposed strategy is evaluated with respect to statistical accuracy by two simulation studies controlling the missing data generating mechanism. I show that the novel algorithm is capable of recovering all involved parameters in each of the two scenarios and clearly outperforms stochastic regression imputation and complete cases analysis. Two illustrations using data from the National Educational Panel Study on mathematical abilities and eating disorders of ninth grade students demonstrate the empirical usefulness of the method. Finally, I introduce an R package which implements the estimation routines presented in the thesis.

Key words: item response theory, population heterogeneity, Markov chain Monte Carlo, multiple imputation, classification and regression trees, statistical computing, R, National Educational Panel Study.

Contents

List of Tables

List of Figures

List of Acronyms

2PNO	Two-Parameter Normal Ogive
BD	Before Deletion
BMI	Body Mass Index
CART	Classification And Regression Trees
CC	Complete Cases analysis
CDF	Cumulative Distribution Function
CRAN	Comprehensive R Archive Network
CTT	Classical Test Theory
DA	Data Augmentation
DAC	Data Augmentation using sequential Cart imputation
DAR	Data Augmentation using sequential stochastic Regression imputation
FMLRM	Finite Mixture Latent Regression item response Model
GRM	Graded Response Model
GYM	GYMnasium
HS	HauptSchule
HDR	Highest Density Region
ICC	Intraclass Correlation Coefficient
iid	independent and identically distributed
IRF	Item Response Function
IRT	Item Response Theory
LSA	Large-Scale Assessment
LRM	Latent Regression item response Model
MAR	Missing At Random
MCAR	Missing Completely At Random
MCMC	Markov Chain Monte Carlo
MGLRM	MultiGroup Latent Regression item response Model

M-H	Metropolis-Hastings
MI	Multiple Imputation
MICE	Multivariate Imputation by Chained Equations
NAEP	National Assessment of Educational Progress
NEPS	National Educational Panel Study
NMAR	Not Missing At Random
PCM	Partial Credit Model
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
PV	Plausible Value
RILRM	Random Intercept Latent Regression item response Model
RMSE	Root Mean Square Error
RS	RealSchule
SC4	neps Starting Cohort 4 of ninth graders
TIMSS	Trends in International Mathematics and Science Study

List of Symbols

$P(A)$	probability of event A
$P(A \cap B)$	joint probability of event A and event B
$g(\cdot)$	arbitrary function
$f(\cdot)$	continuous probability density function
$F(\cdot)$	univariate CDF
$E[\cdot]$	expected value
$\mathbf{1}(A)$	indicator function of event A
$\mathcal{I}_n$	identity matrix of size n
$\vert$	given that
$\in$	is an element of
$=$	equal to
$\neq$	not equal to
$<$	less than
$>$	greater than
$\leq$	less than or equal to
$\propto$	is proportional to
$\forall$	for all
$\sum$	summation
$\prod$	product
$\int$	integral
$\perp$	perpendicular
$-\infty$	negative infinity
$+\infty$	positive infinity
$\lim_{n \to c} g(x)$	the limit of $g(x)$ as n approaches c
$\underset{x}{\text{argmax}}\ g(x)$	the x value that maximizes $g(x)$
$\sim$	is distributed as
$\overset{\text{iid}}{\sim}$	are iid as

$f(\mathbf{D}\|\psi_l, \mathcal{M}_l)$	sampling density for sample data $\mathbf{D}$ given parameter values $\Psi_l = \psi_l$ of statistical model $\mathcal{M}_l$ $(l : 1, \ldots, L)$
$\pi(\Psi_l\|\mathcal{M}_l)$	prior density of Ψ_l
$\pi(\Psi_l\|\mathbf{D}, \mathcal{M}_l)$	posterior density of Ψ_l given $\mathbf{D}$
$f(\mathbf{D}\|\mathcal{M}_l)$	marginal density of $\mathbf{D}$
$\mathbf{D}^*$	auxiliary data
$\exp\{\cdot\}$	natural exponential function
$\ln\{\cdot\}$	natural logarithmic function
$\mathcal{N}(\mu, \sigma^2)$	univariate normal distribution with mean μ and variance σ^2
$\mathcal{N}_d(\underline{\mu}, \Sigma)$	d-dimensional multivariate normal distribution with mean vector $\underline{\mu}$ and covariance matrix Σ
$t_d(\underline{\mu}, \Sigma, \rho)$	d-dimensional multivariate t distribution with mean vector $\underline{\mu}$, covariance matrix Σ and ρ degrees of freedom
$\mathcal{IG}(a, b)$	inverse-gamma distribution with shape parameter a and scale parameter b
$\mathcal{U}(u, v)$	continuous uniform distribution on the interval $[u, v]$
$\Lambda(\cdot)$	standard logistic CDF
$\Phi(\cdot)$	standard normal CDF
R	MCMC chain length
N	number of respondents
G	number of groups within the context of multigroup and finite mixture models
C	number of clusters within the context of random intercept models
S_i	indicator of observed and latent group membership $(i : 1, \ldots, N)$
$\underline{S}$	N-dimensional vector with elements S_i
$\mathbf{Q}$	$N \times C$ design matrix of zeros. Each row has a single entry 1 indicating respondents' cluster membership
N_g	number of respondents in group g $(g : 1, \ldots, G)$
η_g	relative group size
N_c	number of respondents in cluster c $(c : 1, \ldots, C)$
J	number of items

Q_j	number of response categories for item j $(j : 1, \ldots, J)$
$\mathbf{Y}$	$N \times J$ matrix of item responses
$\underline{y}_{i.}$	J-dimensional vector containing ith row of $\mathbf{Y}$
$\underline{y}_{.j}$	N-dimensional vector containing jth column of $\mathbf{Y}$
$y_{i,j}$	element in row i and column j of $\mathbf{Y}$
$\mathbf{Y}^*$	$N \times J$ matrix of continuous variables underlying item responses
$\mathbf{X}$	$N \times (K+1)$ matrix of person covariates explaining the latent trait including a constant
$\mathbf{Z}$	$N \times (M+1)$ matrix of concomitant variables explaining mixture probabilities including a constant
$\mathbf{M}^A$	missing indicator for matrix $\mathbf{A}$
$p_{i,j}$	probability that respondent i correctly answers binary item j
$p_{i,j,q}$	probability that respondent i scores in category q on item j
θ_i	latent trait for respondent i
$\theta_{c,i}$	latent trait for respondent i in cluster c
$\underline{\theta}$	N-dimensional vector of latent trait values
$\mathbf{T}$	$N \times 2$ auxiliary matrix consisting of $\underline{\theta}$ and a negative intercept
$\underline{\gamma}$	$(K+1)$-dimensional vector of regression weights on $\mathbf{X}$
$\underline{\gamma}_g$	$(K+1)$-dimensional vector of group-specific regression weights on $\mathbf{X}$ for group g
$\underline{\zeta}_g$	$(M+1)$-dimensional vector of group-specific multinomial logit regression weights on $\mathbf{Z}$ for group g
$\varepsilon_{i,j}$	residual for respondent i and item j
ϵ_i	residual for respondent i
ω_c	random intercept for cluster c
σ^2_ϵ	residual variance of the latent trait
$\sigma^2_{\epsilon,g}$	group-specific residual variance of the latent trait for group g
Σ_ϵ	$N_g \times N_g$ diagonal matrix with elements $\sigma^2_{\epsilon,g}$
υ^2_ω	residual variance of the random intercept

α_j	discrimination parameter for item j
$\underline{\alpha}$	J-dimensional vector of item discrimination parameters
β_j	difficulty parameter for item j
$\underline{\beta}$	J-dimensional vector of item difficulty parameters
$\underline{\xi}_j$	two-dimensional vector of α_j and β_j
$\underline{\kappa}_j$	$(Q_j + 1)$-dimensional vector of category cutoff parameters for ordinal item j
$\underline{\tau}_j$	$(Q_j - 2)$-dimensional vector of transformed category cutoff parameters for ordinal item j
Ξ	unobserved parameters governing the missing data mechanism

1 Introduction

Latent variables are pervasive. They appear in such different domains as everyday conversations ("I feel good") and political reports ("The country is considered authoritarian"). What characterizes latent variables is that they are not accessible to direct measurement. In this thesis I will follow Skrondal and Rabe-Hesketh (2004, p. 1) and speak of a latent variable as a random variable whose realizations are hidden from us. This definition allows to relate latent variables not only to hypothetical constructs, but also to statistical concepts like measurement error, missing data or counterfactuals. Thus, latent variables are important in two ways: As an object of research in social sciences and, as Cai (2012) phrases it, "perhaps the single most important concept exported from the psychological sciences to the statistical sciences" (p. 118).

Attempts to measure latent traits have a long history. A pioneer in the development of measurement models was Charles Spearman with his contributions to factor analysis and the construction of a general intelligence factor (Spearman, 1904). Over the last decades, item response theory (IRT; e.g., Embretson & Reise, 2000; Lord, 1980; Lord & Novick, 1968; Rasch, 1960) has emerged as the most popular measurement theory. IRT considers multiple observed responses as indicators of the latent trait and aggregates them towards a single score. Most commonly, persons are asked to respond to a given set of items. Beside the obvious testing situations in the fields of psychometrics and educational measurement, indicators also appear in the form of statements in public opinion research or roll call votes in

political sciences (Clinton, Jackman, & Rivers, 2004).

A joint model for measurement and structural analysis was developed by Muthén (1979) and has been further examined in Zwinderman (1991) and Adams, Wilson, and Wu (1997). In these publications, a multivariate regression equation is used to model the relationship between the latent trait and additional person covariates. In the following, I refer to models of this type as latent regression item response models (LRMs). LRMs are important in three ways: First, they allow to test hypotheses about structural relationships between the latent trait and covariates. Consider, for example, the case of disparities in students' achievement related to socio-economic status. The results of the Programme for International Student Assessment (PISA; e.g., OECD, 2014) for the year 2012 have shown that on average, socio-economically advantaged students score higher in mathematics than less advantaged students. In Germany, 17% of the variation in mathematical competency is explained by differing socio-economic status levels (OECD, 2013a, p. 36). Mathematical competency and its relation to socio-economic status cannot be directly observed and hence have to be inferred from empirical assessment and covariate data through a suitable model. Second, as noted first by Mislevy (1987), enriching measurement models by a person's background variables can lead to precision gains in parameter estimates. Third, LRMs are employed to generate so-called plausible values (PVs) in large-scale assessments (LSAs). The concept of PVs was introduced by Mislevy (1991) and is based on the work of Rubin (1987) on multiple imputation (MI). In short, PVs are random draws from the domain-specific posterior ability distribution for each respondent and thus explicitly account for uncertainty due to measurement error. These values, considered alone, do not appropriately reflect an individual's proficiency but yield unbiased estimates of performance on the group level. The PVs approach becomes especially relevant when test data arise from

a multiple matrix item sampling design, i.e., each respondent answers a randomly administered subset of the complete test item pool (Mislevy, Beaton, Kaplan, & Sheehan, 1992). Such sampling plans are established, for example, in studies like PISA, the Trends in International Mathematics and Science Study (TIMSS; e.g., Mullis & Martin, 2013), the Programme for the International Assessment of Adult Competencies (PIAAC; e.g., OECD, 2013b) or the National Assessment of Educational Progress (NAEP; e.g., Allen, Carlson, & Zelenak, 1999) in the US.

When PVs shall be provided for secondary data analysis, the person covariates included in the LRM to generate PVs need to match the variables related to latent ability in later analyses. Meng (1994) invented the term (un)congeniality to describe this (mis)match of imputation and analysis model. LSAs are especially affected by this issue. PVs provided through public-use files and scientific-use files should allow the user to answer a preferably wide range of specific research questions. Therefore, hierarchical or clustered data structures inherent to LSAs need to be addressed in the construction of ability scores in order to gain consistent estimates of the respective conditional latent ability distributions. Clustered data can stem from various sources of population heterogeneity. For instance, multistage sampling designs are frequently involved in educational assessments in which schools serve as primary sampling units (see, e.g., Aßmann et al., 2011). Hence, students are nested within schools and cannot be treated as independent observations. Another good example for population heterogeneity is the German school system that groups students quite early into different ability tracks and with it into different learning environments. Students in one track share contextual factors that students in other tracks do not share.

Possibilities to consider clustered data structures involve several alternatives. Regression analysis in general accounts for population heterogeneity through the

covariates contained in the model. According to Gelman and Hill (2006), "Linear regression is a method that summarizes how the average values of a numerical *outcome* vary over subpopulations defined by linear functions of *predictors*" (p. 31). All omitted factors and disturbances affecting the outcome are summarized under a random error term.

LRMs can be extended towards additional levels of analysis by allowing the regression parameters to vary across predefined groups. If an extra model is formulated for the group-specific coefficients (a common distribution or another regression equation), one usually speaks of multilevel analysis or random effects models (e.g., Gelman & Hill, 2006; Hox, 2010; Snijders & Bosker, 2012). Two major features of these models are the introduction of group-level covariates and a decomposition of the total variance into between-group and within-group variability. Continuing with the PISA 2012 example, it has been shown that in Germany the variation in mathematics performance between schools is rather large, 156% of the OECD average (OECD, 2013a, p. 196), and that mainly schools' mean socio-economic status and students' study track account for these differences (OECD, 2013a, pp. 200-201). Such detailed findings highlight important aspects of the education system, which may be very relevant for political decison making.

A further modeling strategy towards hierarchical structures is to specify the latent ability distribution as a mixing distribution (for diverse applications in educational measurement, see von Davier & Carstensen, 2007). The use of mixing distributions resembles the idea of model-based clustering as suggested by Frühwirth-Schnatter and Kaufmann (2008). As opposed to multilevel analysis, model-based clustering assumes a composite population consisting of a finite number of unobserved groups, where each group-specific latent ability distribution is governed by a different set of parameters. Thus, population heterogeneity is not derived from a

deterministic stratification, but is inferred from the data. A review of model-based clustering techniques is given in Fraley and Raftery (2002). Dayton and Macready (1988) were the first to relate the probability of latent group membership to person covariates or concomitant variables as they call them. Introducing mixing probabilities conditional on concomitant variables allows for incorporating additional sources of heterogeneity and hence offers a flexible yet parsimonious way to correctly reproduce clustered data structures. This formulation comprises the cases of equal a priori mixing probabilities for all observations and observable group membership, also known as multigroup modeling (e.g., Bock & Zimowski, 1997; Muthén & Christoffersson, 1981).

As demonstrated by Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003), and described extensively in Wilson and De Boeck (2004), LRMs can be conceptualized within the wider context of nonlinear mixed models. Since the derived likelihood functions involve multivariate integrals and latent variables within these integrals, a Bayesian framework using Markov chain Monte Carlo (MCMC) techniques is eminently suited for parameter estimation in LRMs and their extensions. Bayesian inference has received increased attention in fitting IRT models. The seminal article of Albert (1992) adopts a data augmentation (DA) estimation algorithm (Tanner & Wong, 1987) for measurement models with dichotomous items. Further work adopted Albert's procedure for multidimensional IRT models (Béguin & Glas, 2001) and the confirmatory item factor analysis framework (Ansari & Jedidi, 2000; Edwards, 2010). Patz and Junker (1999a, 1999b) developed the corresponding model counterpart based on the logistic distribution without the use of DA. Bayesian estimation of LRMs incorporating multilevel data structures were adapted by Fox and Glas (2001), Fox (2005) and Johnson and Jenkins (2005), whereas LRM for multiple groups are discussed in Azevedo, Andrade, and Fox (2012). To my knowledge, finite

mixtures of LRMs have not yet been covered in the literature. A closely related model can be found in Lenk and DeSarbo (2000), who apply Bayesian inference for finite mixtures of generalized linear models with random effects.

The main advantages of a Bayesian estimation strategy in LRMs can be summarized as follows:

- A solution for the arising multidimensional integration problems is given,
- compared to existing methods based on maximum likelihood estimation, all involved model parameters are estimated simultaneously and not stepwise (see, e.g., von Davier, Sinharay, Oranje, & Beaton, 2006, for the three-stage estimation process used in NAEP), and
- uncertainty concerning the parameter estimates is addressed.

However, person covariates used to predict the latent trait in LRMs are often seriously afflicted by item nonresponse. Si and Reiter (2013), for example, report less than five percent complete cases on a set of 80 background variables in a TIMSS data file. Such a large amount of data loss poses a great challenge to the estimation of structural parameters. Evidently under such conditions, one has to think of an appropriate strategy for the nonresponse during the stages of data handling and data analysis. While several studies deal with the impact of omitted item responses (e.g., Köhler, Pohl, & Carstensen, 2015; Pohl, Gräfe, & Rose, 2014), there has been little work so far on missing values in background variables.

In order to cope with missing information on individual-level background variables, I propose a fully Bayesian approach in the estimation of LRMs and their extensions. In particular, I employ estimation routines relying on DA methodology and additionaly contain the missing values as a part of the parameter vector. As a

consequence, the algorithm iteratively switches between parameter updates and an imputation step. Among others, this tool has been successfully applied for multivariate panel models by Liu, Taylor, and Belin (2000) and in the field of social network models by Koskinen, Robins, and Pattison (2010). For my research, the compelling argument behind this approach is the joint modeling of the latent trait and partially missing covariate data having regard to correlated observations in homogenous respondent groups. Whilst existing methods yet always lead to a multistage procedure, the new approach simultaneously addresses the uncertainty associated with the estimation of a latent trait variable and the imputation of missing values in manifest predictors. The reciprocal dependence of outcome and predictors is thus reflected to the full extent by the algorithm.

Up to now, the named LSAs treat missing values in context questionnaires as a nominal response category (e.g., OECD, 2014, p. 421-431). Thus, categorical background variables are simply dummy-coded. For the continuous background variables, item nonresponse is completed with the variable mean and a missing indicator is added to the regression as an additional covariate. These methods originate from Cohen, Cohen, West, and Aiken (1975/2002, Chapter 11) and became known as *dummy-coding* and *contrast-coding*. After recoding the context questionnaire, a principal component analysis of all background variables is conducted. Thereby, as many principal components as needed to explain 90% of the variance in the original variables are extracted. The set of principal components then enters the structural model. Aside from the obvious information loss, dummy-variable adjustments for missing values have shown "unacceptably large biases in practical situations and are not advisable in general" (Jones, 1996, p. 222). These results are in line with a recent study by Rutkowski (2011) who found nonnegligible bias and misleading interpretations at the population level when partially missing covariates are dummy

coded.

Some analysts (e.g., Bouhlila & Sellaouti, 2013; Weirich et al., 2014) have presented approaches to MI for context questionnaires in LSAs. Accounting for missing values by MI adds another step to the estimation process. Further, performing MI with a large number of background variables can quickly become a daunting task for the analyst. He or she has to specify a model that specifies how every variable is imputed. In doing so, MI of context questionnaires resembles the "imputation" of PVs in matters of hierarchical data structures. However, MI of multilevel data is still a rather underdeveloped research area. To better understand the mechanisms behind multilevel missing data, Drechsler (2015) theoretically and empirically evaluates the impact of fixed effects imputation models. By drawing on mixed effects logistic regression models, Zinn (2013) introduces a software add-on within the multivariate imputation by chained equations (MICE) framework to impute multilevel binary data.

Due to its high flexibility, I apply nonparametric approximations to the distributions of missing values based on sequential classification and regression trees (CART) as suggested by Burgette and Reiter (2010). Sequential CART belong to the class of hot-deck imputation techniques (see, e.g., Andridge & Little, 2010, for a review). The similiarity of recipients and possible donors is accomplished through predictions based on CART. As CART are intended to operate with metric and categorical variables, they can consequently be utilized for the imputation of both variable types. Burgette and Reiter (2010) and Doove, van Buuren, and Dusseldorp (2014) have further shown that sequential CART are especially well suited when nonlinear dependencies are present in the data.

The thesis proceeds as follows. Chapter 2 gives a brief overview of IRT and outlines different specifications of LRMs. It includes a generalized model for dichoto-

mously and ordered polytomously scored items, as well as the mentioned extensions for hierarchical structures. The philosophy of Bayesian inference and its practical implementation are adapted to the different types of LRMs in Chapter 3. The fourth Chapter proposes the compound sampling algorithm that also imputes missing values in person covariates. Performance and applicability of the estimation routines are demonstrated through a simulation study and two exemplary analyses using the first and second wave of the German National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011) starting cohort in ninth grade. In Chapter 5, I provide a basic guide to the self-created R package *LaRA*, which implements the algorithms described in Chapters 3 and 4. A sketch of ongoing research and forthcoming projects is given in Chapter 6, while Chapter 7 completes the thesis with a few concluding remarks.

2 Latent regression item response models

> The very objective of social science research is not to discover abstract and universal laws but to understand population heterogeneity. (Xie, 2013, p. 6262)
>
> We boil at different degrees.
>
> —Ralph Waldo Emerson, *Society and Solitude*, 1870

2.1 Item response theory

Measurement theories provide the framework for latent trait estimation. They define how the link between observed indicators and a latent variable is established. In the following, I will concentrate on the task of *competency* assessment as an example for latent variable measurement. The terms competency, *ability* and *proficiency* are thereby used interchangeably. Note, however, that measurement theories can be applied to scale any unobservable trait of interest.

Consider J tasks completed by N individuals. This leads to a $N \times J$ test data matrix $\mathbf{Y}$ of the form

$$\begin{array}{r} \\ \text{person } 1 \\ \text{person } 2 \\ \vdots \\ \text{person } N \end{array} \begin{array}{c} \begin{array}{cccc} \text{item } 1 & \text{item } 2 & \cdots & \text{item } J \end{array} \\ \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,J} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,J} \\ \vdots & \vdots & \vdots & \vdots \\ y_{N,1} & y_{N,2} & \cdots & y_{N,J} \end{pmatrix} \end{array},$$

written compactly as

$$\mathbf{Y} = (y_{i,j}), \qquad i : 1, \ldots, N; \quad j : 1, \ldots, J. \tag{2.1}$$

Let vectors $\underline{y}_{i.} = (y_{i,1}, y_{i,2}, \ldots, y_{i,J})$ and $\underline{y}_{.j} = (y_{1,j}, y_{2,j}, \ldots, y_{N,j})'$ denote single rows and columns of $\mathbf{Y}$. Hence, $\mathbf{Y}$ has a two-dimensional structure that consists of items which are nested within persons, the observational units.

Due to its ease of application and computation, classical test theory (CTT) has been the predominant paradigm in psychometrics for a long time. The theory builds on the frequentist reasoning of infinitely replicating the test situation. The basic equation of CTT decomposes the observed test score into two parts, *true score* and *error score*, where the true score is defined as the expected observed score and the error score results from the difference of observed score and true score (Lord & Novick, 1968, Chapter 2). Depending on additional assumptions, the sum score $\widehat{T}_i = \sum_j \underline{y}_{i.}$ serves as an estimator for a person's competency in CTT.

If the total test score is used to estimate a respondent's ability and all items are weighted equally in the construction of the score, the items are interchangeable with each other. Different individual response patterns $\underline{y}_{i.}$ can result in the same value of the latent trait. This neglection of an item-ability relationship precludes the analyst from gaining any insight on how a respondent performs on the item-level. Further, and perhaps the greatest shortcoming associated with it, the results derived from CTT are entirely test- and sample-dependent (for a complete list of CTT properties, see, e.g., Hambleton & Jones, 1993). IRT is an alternative psychometric theory that comprises a wide range of probabilistic measurement models. I now discuss its fundamental assumptions on the basis of models for binary and ordinal response data. Opposed to CTT, these assumptions are testable and can be evaluated for a particular data set.

2.1.1 Binary outcomes

The simplest response format to use in educational measurement is the format of binary test items. It includes the empirical information whether a person correctly responds to an item or not. Let $Y_{i,j}$ be a random variable taking on the value 1 when respondent i is able to solve item j and the value 0 otherwise. In the context of binary response data, IRT aims at modeling the probability that the ith test taker answers item j correctly, i.e.,

$$p_{i,j} = P(y_{i,j} = 1). \tag{2.2}$$

This is done by specifying $p_{i,j}$ as a function of a scalar person parameter θ_i and a set of item parameters $\underset{\sim}{\xi}_j$. The length of vector $\underset{\sim}{\xi}_j$, i.e., the number of item parameters, distinguishes between different IRT models:

1. In a one-parameter model, the only element in $\underset{\sim}{\xi}_j$ is the item difficulty β_j.

2. In a two-parameter model, the test item is characterized by $\underset{\sim}{\xi}_j = (\alpha_j, \beta_j)'$, where α_j is a discrimination parameter.

3. In a three-parameter model, a guessing parameter is added to $\underset{\sim}{\xi}_j$. The model goes back to Birnbaum (1968) and considers the possibility that respondents randomly choose among the given choices. This would enable a respondent to solve an item although he or she lacks in the required level of proficiency. I do not further consider a guessing parameter, because it is not consistent with the intended joint model for binary and ordinal test items.

To link all parameters to the probability of a correct response, a function $F(\cdot)$ from the class of cumulative distribution functions (CDFs) is chosen, which naturally fulfill the constraint $0 \leq p_{i,j} \leq 1$. Usually, $F(\cdot)$ is chosen a standard

logistic or standard normal CDF[1] denoted as $\Lambda(\cdot)$ or $\Phi(\cdot)$. In the following I restrict our analysis to the standard normal link function, because it offers an important computational advantage for MCMC based Bayesian estimation (see Chapter 3). Lord (1952, 1953) already formulated an IRT model which became generally known as the two-parameter normal ogive (2PNO):

$$P(y_{i,j} = 1|\theta_i, \underset{\sim}{\xi}_j) = \Phi(\alpha_j\theta_i - \beta_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha_j\theta_i - \beta_j} \exp\left\{-\frac{s^2}{2}\right\} ds. \tag{2.3}$$

Figure 2.1 illustrates the relationship between $p_{i,j}$ and the corresponding parameters made in equation (2.3) for three exemplary items. The resulting graphs are referred to as item response functions (IRFs). While the probability of a correct response $p_{i,j}$ is graphed on the y-coordinate, latent proficiency θ_i as well as item difficulty β_j are mapped to the x-coordinate.

For all IRFs, we see a monotonically increasing function of θ_i, which means an increased probability of solving the item for higher values of the proficiency variable. A comparison of item 1 and item 2 shows the effect of a change in item difficulty β_j while α_j remains unchanged. The more difficult item 2 with $\beta_2 = 1$ implies a higher ability value to effectuate the same probability of solving the item, which leads to a parallel right shift of the IRF. Regarding the discriminating power of an item, a comparison of item 1 and item 3 reveals the ceteris paribus effect of a change in α_j. Item 3 with $\alpha_3 = 0.5$ is less able to discriminate between persons, because the probabilitiy of solving the item varies less for different values of θ_i. Graphically this is demonstrated through the gentler incline of IRF 3. In terms of regression analysis, β_j and α_j serve as (negative) intercept and slope.

An important implication of the model-based IRF is its generalizability over

1 Both specifications yield very similar results. The standard logistic distribution has marginally fatter tails due to its higher variance.

different samples. This is referred to as *parameter invariance*. IRT models involve further assumptions (e.g., Embretson & Reise, 2000, p. 45-48). *Local independence* assumes that after controlling for the trait level and the specified item parameters, item responses are statistically independent of one another. There are no other relationships among persons or items than those postulated in the model. The probability of solving both items j and j' equals the product of the single probabilities:

$$P(y_{i,j} = 1 \cap y_{i,j'} = 1|\theta_i, \underline{\xi}_j, \underline{\xi}_{j'}) = P(y_{i,j} = 1|\theta_i, \underline{\xi}_j)P(y_{i,j'} = 1|\theta_i, \underline{\xi}_{j'}) \qquad \forall j \neq j'.$$

Moreover, the assumption of *unidimensionality* means that a test measures only one latent trait. Taken local independence and unidimensionality together, the latent competency score in a certain domain is the only thing that governs response behaviour during a test. This enables for computing the joint probability of an individual response pattern as

$$f(\underline{y}_{i\cdot}|\theta_i, \{\underline{\xi}_j\}_{j=1}^J) = \prod_{j=1}^{J} P(y_{i,j} = 1|\theta_i, \underline{\xi}_j). \tag{2.4}$$

An alternative way to represent binary IRT models is in terms of a threshold mechanism which was first formalized in the context of individual level data by McKelvey and Zavoina (1975) and can be found for multivariate binary variables in Maddala (1983, p. 138). Let $y_{i,j}$ take on the value 1 if an underlying continuous variable $y^*_{i,j}$ exceeds the threshold value of 0, otherwise a 0 is observed:

$$y_{i,j} = \begin{cases} 1 & \text{if } \; y^*_{i,j} > 0 \\ 0 & \text{if } \; y^*_{i,j} \leq 0. \end{cases} \tag{2.5}$$

The underlying variable $y^*_{i,j}$ is introduced as an unobservable quantity which is not to be confused with the latent ability θ_i. Rather, the observed test results $y_{i,j}$ have to be seen as imprecise measures of $y^*_{i,j}$. Writing the 2PNO as a linear

regression with $y^*_{i,j}$ as the dependent variable,

$$y^*_{i,j} = \alpha_j \theta_i - \beta_j + \varepsilon_{i,j}, \tag{2.6}$$

where the independent and identically distributed (iid) error term $\varepsilon_{i,j}$ follows a standard normal distribution, implicates $y^*_{i,j}$ to be independent draws from a normal distribution. This further shows the equivalence to equation (2.3):

$$p_{i,j} = P(y^*_{i,j} > 0) = 1 - \Phi(-(\alpha_j \theta_i - \beta_j)) = \Phi(\alpha_j \theta_i - \beta_j).$$

In the following, I use the derivation of binary IRT models outlined in (2.5) and (2.6), because it greatly simplifies the set-up of a DA sampling algorithm. Another reason is that IRT models for ordinal response data follow the underlying variable approach too.

2.1.2 Ordered polytomous outcomes

Ordinal response formats usually consist of more than two categories and there is a distinct order inherent in them. Survey questionnaires commonly hold ordinal variables in the form of Likert items or scales of satisfaction. In educational measurement, test items exist that encompass multiple subtasks in the solution process, so partial credit can be assigned for partial success on an item. Masters (1982) gives an example of such an item from mathematics assessment:

$$\sqrt{7.5/0.3 - 16} \; = \; ?$$

(0) failed $<$
(1) $7.5/0.3 \; = \; 25 \; <$
(2) $25 - 16 \; = \; 9 \; <$
(3) $\sqrt{9} \; = \; 3.$

As another example, persons not only (don't) agree with statements in public

opinion reserach, but choose between graded options like

(0) strongly disagree $<$
(1) disagree $<$
(2) agree $<$
(3) strongly agree.

It now becomes possible to determine the ranking of persons according to their answers. We obtain the response variable $y_{i,j} \in \{0, 1, \ldots, Q_j - 1\}$ having Q_j categories, where the numerical values have no significance except for preserving the ordering of response vector $\underline{y}_{\cdot j}$.

In mixed format tests a score $\tilde{y}_{i,j}$ can be provided to regulate the relative importance of ordinal items to binary items. Thereby the researcher chooses a weighting factor w, so that

$$\tilde{y}_{i,j} = \begin{cases} y_{i,j} & \text{if item is binary} \\ w y_{i,j} & \text{if item is ordinal.} \end{cases} \tag{2.7}$$

If $w = 1$, the score $\tilde{y}_{i,j}$ is equal to the number of completed subtasks.

Models for analyzing ordered polytomous item responses were introduced in the field of psychometrics through the Graded Response Model (GRM) by Samejima (1969). In the GRM, likewise in the case of binary test data, observed item responses can be seen as a ordered polytomous version of an underlying continuous variable $y^*_{i,j}$ as stated in (2.6). From this formulation, one can again link the observed categorical and the underlying continuous variable using a threshold mechanism, namely

$$y_{i,j} = \begin{cases} 0 & \text{if } y^*_{i,j} \leq \kappa_{j,1} \\ 1 & \text{if } \kappa_{j,1} < y^*_{i,j} \leq \kappa_{j,2} \\ \vdots & \vdots \\ Q_j - 1 & \text{if } y^*_{i,j} > \kappa_{j,Q_j-1}, \end{cases} \tag{2.8}$$

where $\underline{\kappa}_j = (\kappa_{j,0}, \kappa_{j,1}, \ldots, \kappa_{j,Q_j})'$ is the $(Q_j + 1)$-dimensional vector of item category cutoff parameters fulfilling the ordering constraint

$$\kappa_{j0} = -\infty < \kappa_{j1} < \ldots < \kappa_{jQ_j-1} < \kappa_{jQ_j} = +\infty. \tag{2.9}$$

When $y^*_{i,j}$ exceeds the threshold $\kappa_{j,q}$ $(q : 0, \ldots, Q_j - 1)$ and remains beneath threshold $\kappa_{j,q+1}$, $y_{i,j}$ is classified as response q. Figure 2.2 depicts mechanism (2.8) for a four-category item. The resulting probability that respondent i achieves grade q on item j, given his latent proficiency and item parameters, is given by

$$\begin{aligned} p_{i,j,q} &= P(y_{i,j} = q|\theta_i, \underline{\xi}_j, \kappa_j) = P(\kappa_{j,q} < y^*_{i,j} \leq \kappa_{j,q+1}) \\ &= \Phi(\kappa_{j,q+1} - (\alpha_j\theta_i - \beta_j)) - \Phi(\kappa_{j,q} - (\alpha_j\theta_i - \beta_j)) \\ &= \Phi(\alpha_j\theta_i - (\beta_j + \kappa_{j,q})) - \Phi(\alpha_j\theta_i - (\beta_j + \kappa_{j,q+1})). \end{aligned} \tag{2.10}$$

Note that $p_{i,j,q}$ consists of the difference of two cumulative probabilities, $P(y_{i,j} \leq q+1)$ and $P(y_{i,j} \leq q)$. Over all ordered polytomous items, the number of categories Q_j does not need to be equal.

Here, I adopt the model specification of Edwards (2010) including an overall item difficulty parameter β_j, which slightly differs from standard GRMs. To ensure identifiability of the model parameters, we additionally impose the restriction $\kappa_{j,1} = 0$. This corresponds to the approach for solving overparameterization in the binary IRT model, where the lower threshold for a correct response is 0. Therefore, (2.10) encompasses binary test items as a special case whenever $Q_j = 2$. Considering the univariate case and the logistic link function, this model type is also known as the proportional odds model originated by McCullagh (1980).

Another possible model for ordered polytomous responses is provided by the Partial Credit Model (PCM; Masters, 1982). It divides $y_{i,j}$ into a sequence of binary items to explicitly model the choice between adjacent categories $y_{i,j,q}$ and $y_{i,j,q+1}$.

In this case, the item category response probabilities $p_{i,j,q}$ are modeled directly as

$$P(y_{i,j} = q|\theta_i, \underline{\xi}_j) = \frac{\exp\left\{\sum_{h=0}^{q}(\theta_i - \beta_{j,h})\right\}}{\sum_{l=0}^{Q_j-1}\exp\left\{\sum_{h=0}^{l}(\theta_i - \beta_{j,h})\right\}}. \tag{2.11}$$

It is apparent from (2.11), that the PCM encompasses the famous one-parameter logistic model for binary responses, also known as the Rasch model (Rasch, 1960). Item parameters $\underline{\xi}_j = (\beta_{j,1}, \ldots, \beta_{j,Q_j-1})'$ can now be interpreted as a vector of item category-specific difficulties. A version introducing the slope parameter α_j to the PCM was developped by Muraki (1992). However, with focussing only on dichotomies, the PCM does not ensure the ordering of $\underline{\kappa}_j$. For this reason, it contradicts the underlying variable formulation outlined in (2.8) and I do not further pursue the PCM.

2.2 Structural component

IRT models are designed to directly compare items and persons on a common scale. To enlarge their scope, the focus of analysis was broadened towards structural analysis in the groundbreaking article of Muthén (1979). Muthén was the first to address the issue that persons may not only differ in terms of their ability, but also in terms of covariates which are correlated with their ability. Many research questions require to perform this type of analysis. For example, factors that influence the acquisition of competencies are of major interest to educational reserach. LSAs ususally survey additional background information on the respondents via context questionnaires. I define a $N \times (K+1)$ matrix

$$\mathbf{X} = (x_{i,k}), \qquad i : 1, \ldots, N; \quad k : 0, \ldots, K, \tag{2.12}$$

containing K person covariates including an intercept. A multivariate regression equation of the form

$$\theta_i = \underline{x}_{i.}\underline{\gamma} + \epsilon_i, \tag{2.13}$$

where the iid error term ϵ_i follows a normal distribution with mean 0 and variance σ^2_ϵ, can be inserted to model the relationship between the latent trait variable θ_i and a set of covariates, where $\underline{x}_{i.} = (1, x_{i,1}, \ldots, x_{i,K})$, $\underline{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_K)'$ is a $(K+1)$-dimensional vector of regression weights and σ^2_ϵ a scalar residual variance parameter. Note that through the distributional assumption made for the residuals ϵ_i, predictions for the person proficiencies are equivalent to normal random variables with mean $\underline{x}_{i.}\underline{\gamma}$ and variance σ^2_ϵ prior to testing. Figure 2.3 visualizes the entire model for fictitious data in form of a path diagram associated with the confirmatory factor analysis framework.[2]

Often, models of this type are labeled LRM in the psychometric literature, because a regression analysis is performed with the dependent variable being a latent construct. The decomposition of a latent trait into fixed effects $\underline{\gamma}$ and random component σ^2_ϵ is demonstrated in Zwinderman (1991). Wilson and De Boeck (2004) published a book in which they show the equality of IRT models and generalized linear and nonlinear mixed models and plead for an integration of both approaches. Moreover, Mislevy (1987) identified improved item parameter estimates through the exploitation of context information, especially for shorter tests with small J.

Another typical application of LRMs is the construction of PVs in LSAs. Here, I briefly describe the procedure currently applied in NAEP. Following von Davier et al. (2006), the analysis is conducted in three stages. In the first stage

2 Although IRT and confirmatory factor analysis stand in seperate traditions, Takane and De Leeuw (1987) demonstrated the analytical similarities of the two methods.

(*scaling*), an IRT model based on the response data $\mathbf{Y}$ is fitted to obtain estimates of item parameters. Stage 2 (*conditioning*) yields maximum likelihood estimates of the LRM structural parameters from a variant of the expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977; Mislevy, 1985; Mislevy, Eugene, & Muraki, 1992) taking into account $\mathbf{Y}$ as well as person covariates $\mathbf{X}$ and fixing item parameters at the results from stage 1. The population model is then used to randomly draw a vector of *plausible values* for each examinee from an approximation to the conditional posterior distribution of individual ability. The concept of PVs was introduced by Mislevy (1991) and is based on the work of Rubin (1987) on MI. These values, considered alone, do not appropriately reflect an individual's proficiency but yield unbiased estimates of performance on the group level. After applying the statistic of interest over all individuals in a group, the results finally need to be averaged over all plausible values according to Rubin's combining rules. The last stage 3 (*variance estimation*) examines variation of the group estimators due to sampling design and measurement error regarding the latent abilities.

2.3 Extensions for clustered observations

LRMs allow for interindividual differences through the inclusion of covariates. When omitted or unobserved clusters are present, these need to be incorporated in the model as well. Consideration of hierarchical data structures is an important prerequisite for valid inference. Cameron and Trivedi (2005, p. 611) demonstrate in the context of survival analysis, how aggregation across heterogeneous groups can lead to confounded results. The multiple forms of population heterogeneity in educational reserach are reviewed in Muthén (1989) and Burstein (1980).

2.3.1 Multigroup

A first method to further consider population heterogeneity is offered by multiple group latent regression item response models (MGLRMs). The MGLRM assumes a composite population consisting of a finite number, say G, of mutually exclusive groups. Within these groups, seperate LRM may hold. Each subpopulation is now characterized by a group-specific $(K+1) \times 1$ vector of regression weights γ_g and a group-specific scalar residual variance $\sigma^2_{\epsilon,g}$. Sample stratification is thereby based on an explicitly observed cluster variable like gender or school type. MGLRMs date back to the early works of Muthén and Christoffersson (1981) and Mislevy (1985), but without providing for covariates except the cluster variable. Often, the specification of a MGLRM is theory driven with the aim to discover substantial differences of covariate effects and variances over groups. These differences are captured through the estimation of group-specific latent trait distributions. Let $\underline{S} = (S_1, \ldots, S_N)'$ denote the vector of individual group membership known prior to analysis, where $S_i \in \{1, \ldots, G\}$. Given S_i, the MGLRM is stated as

$$y^*_{i,j} = \alpha_j \theta_i - \beta_j + \varepsilon_{i,j} \qquad \text{with } \theta_i = \underline{x}_{i.} \underline{\gamma}_{S_i} + \epsilon_i, \tag{2.14}$$

where $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon,Si})$. Whereas group-specific regression weights $\underline{\gamma}_g$ could be treated in basic LRMs through the inclusion of fixed effects and corresponding interactions, group-specific conditional variances $\sigma^2_{\epsilon,g}$ are a distinctive feature of MGLRMs.

Before group means can be compared meaningfully, measurement invariance has to be ensured. Measurement invariance means that items have equal properties across subgroups. Consider an example from educational measurement: The concept is violated, if an item is more difficult for some subjects than others given the same

subject ability. Now it becomes unclear, whether differences in ability are based on group ability differences or differences in item properties. Muthén (1988) considers LRMs to check for the presence of measurement variance through direct effects of person covariates on the test items. This type of analysis is also referred to as differential item functioning (for methodology reviews on the topic, see Millsap and Everson (1993) and Teresi (2006)). This thesis is not concerned with the issue of measurement invariance.

2.3.2 Finite mixture

Just as MGLRMs, finite mixture latent regression item response models (FMLRMs) assume a composite population with group-specific probability distributions as formalized in (2.14). In contrast to MGLRMs, the competency distributions originate from latent groups and individual group membership has to be inferred from the sample data. Prior to analysis, the researcher determines the number of latent groups. The assignment to a group is now characterized as random and based on mixing probabilities $P(S_i = g|\phi)$, where ϕ is a model parameter. This leads us to the unconditional (on S_i) density

$$f(y_{i\cdot}|\{\underline{\xi}_j, \underline{\kappa}_j\}_{j=1}^J, \{\underline{\gamma}_g, \sigma^2_{\epsilon,g}\}_{g=1}^G, \phi) = \sum_{g=1}^{G} P(S_i = g|\phi) f(y_{i\cdot}|\{\underline{\xi}_j, \underline{\kappa}_j\}_{j=1}^J, \underline{\gamma}_g, \sigma^2_{\epsilon,g}) \tag{2.15}$$

which resembles a finite mixture distribution with G components. Frühwirth-Schnatter (2011, p. 261) discusses two alternatives to define ϕ. Mixture component membership depends on

1. unobserved relative group sizes $\{\eta_g\}_{g=1}^G,\ \sum_{g=1}^G \eta_g = 1$,

$$P(S_i = g|\{\eta_g\}_{g=1}^G) = \eta_g, \qquad g : 1, \ldots, G, \tag{2.16}$$

or

2. a vector of (additional) person covariates $\underline{z}_{i.} = (1, z_{i,1}, \ldots, z_{i,m})'$ including a constant via the multinomial logistic regression model

$$P(S_i = g | \{\underline{\zeta}_g\}_{g=1}^{G-1}, \underline{z}_{i.}) = \frac{\exp(\underline{z}_{i.}\underline{\zeta}_g)}{1 + \sum_{g=1}^{G-1} \exp(\underline{z}_{i.}\underline{\zeta}_g)}. \tag{2.17}$$

Standard FMLRMs correspond to alternative 1. and assume constant mixing probabilities over all persons. Using covariates to model the mixing probabilities has been introduced by Dayton and Macready (1988), who coined the term *concomitant variables* for z_i. We will denote by

$$\mathbf{Z} = (z_{i,m}), \qquad i : 1, \ldots, N; \quad m : 0, \ldots, M, \tag{2.18}$$

the matrix of person covariates used to predict latent group membership. In principle, $\underline{z}_{i.}$ may overlap with $\underline{x}_{i.}$ and thus contain covariates that are simultaneously used to predict the latent trait. This provides the possibility to examine complex nonlinear effects on the outcomes $y^*_{i,j}$. A few authors applied the concomitant-variables FMLRM, including examples in educational measurement (Aitkin & Aitkin, 2011, p. 48; Lubke & Muthén, 2005), a study of vision problems (Huang & Bandeen-Roche, 2004) and an economic application of finite mixture binary panel probit models (Aßmann & Boysen-Hogrefe, 2011). Evidence suggests that standard errors and group assignment can benefit from the adding of concomitant variables (Smit, Kelderman, & Van der Flier, 1999, 2000).

In summary, the general intention of IRT mixture modeling is best expressed in the words of Rost (1990):

The primary diagnostic potential of this model lies in its property to ac-

> count for qualitative differences among examinees, and its simultaneous ability to quantify their abilities with respect to the same tasks. ...; yet it is an obvious idea for the psychological practitioner who knows that relevant individual differences are not only differences in how well somebody can do something, but also in how he/she does these things. (p. 281)

2.3.3 Random intercept

A third approach to incorporate hierarchical structures offers the inclusion of further random effects. It is common in multilevel analysis to include effects on three or more levels, see, e.g., Snijders and Bosker (2012, p. 90-92). For example, one seeks to investigate simultaneously the effects of school selectivity and the classroom organization. Early applications in education testing can be found in Raudenbush (1988) and Mislevy and Bock (1989). As in multigroup models there is a composite population and the individual membership is known a-priori, now denoted by C clusters with N_c respondents in cluster c and $\sum_{c=1}^{C} N_c = N$. While fixed group-specific regression parameters are suitable for relative small numbers of groups, drawing on hierrachical structures with regard to schools or classes causes a prohibitively large number of parameters. Difficulties regarding the computation and the statistical properties of the maximum likelihood estimator in this context were studied by Greene (2004). The problem has been discussed extensively under the term *incidental parameter problem* in the statistics literature, see Lancaster (2000) for a survey. Thus, the introduction of normally distributed cluster-specific effects ω_c $(c : 1, \ldots, C)$ with mean 0 and variance υ_w^2 offers an appropriate alternative to the fixed effects approach followed in MGLRMS and FMLRMs. Note that the number of cluster-specific parameters reduces to the variance term υ_w^2 and the variables

ω_c contain the random heterogeneity related to the cth cluster. This corresponds to the most basic multilevel specification, the random intercept latent regression item response model (RILRM). Adopting the notation from Sections 2.3.1 and 2.3.2, the underlying variable can be expressed in the RILRM as

$$y^*_{c,i,j} = \alpha_j \theta_{c,i} - \beta_j + \varepsilon_{c,i,j} \qquad \text{with } \theta_{c,i} = \underline{x}_{c,i.}\underline{\gamma} + \omega_c + \epsilon_{c,i}, \tag{2.19}$$

where $\varepsilon_{c,i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\omega_c \overset{\text{iid}}{\sim} \mathcal{N}(0, \upsilon_w^2)$, $\epsilon_{c,i} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ and $\omega_c \perp \epsilon_{c,i} \forall i, c$ for $c : 1, \ldots, C$; $i : 1, \ldots, N_c$; $j : 1, \ldots J$. Concerning the multilevel structure, one finds items (level-one) nested within persons (level-two) nested within clusters (level-three).

An intraclass correlation coefficient (ICC) can be estimated for the structural component of RILRMs as

$$\frac{\upsilon_\omega^2}{(\upsilon_\omega^2 + \sigma_\epsilon^2)} \tag{2.20}$$

(e.g., Snijders & Bosker, 2012, Section 3.3). This level-two ICC describes the strength of association between two randomly chosen observations from a randomly sampled cluster. Technically, it expresses the proportion of variance that is explained by the clustering. υ_ω^2 and σ_ϵ^2 are often referred to as *between-cluster variance* and *within-cluster variance*. For example, if competencies are assessed on students nested within schools, then the ICC gives the proportion of the variance in the latent competencies that is between the schools' competency means.

In this Chapter, I presented different specifications of LRMs derived from the IRT framework. The choice of a specific model is primarily driven by the focus of analysis and the characteristics of the data. Especially in the case of mixture models, model choice criteria can help to determine the number of mixture components. Chapter 6 identifies a possible approach to Bayesian model selection. After the

statistical modeling framework is set up, one has to face the issue of estimating the unknown parameters. I will move on to this topic in the next Chapter.

3 Bayesian inference

> The practicing Bayesian is well advised to become friends with as many numerical analysts as possible. (Berger, 1985, p. 262)

3.1 The basics

Bayesian inference differs from frequentist and likelihood approaches to statistics in applying Bayes' theorem. The theorem begins with the *prior distribution*, which reflects the prior belief of the researcher concerning a parameter (or a set of parameters). As new information becomes available in form of given sample data, the prior knowledge of the parameter(s) receives an update. With regard to a specific model $\mathcal{M}_l$ $(l : 1, \ldots, L)$ involving parameter(s) Ψ_l, the resulting *posterior distribution* conditional on sample data $\mathbf{D}$ equals

$$\pi(\Psi_l|\mathbf{D}, \mathcal{M}_l) = \frac{f(\mathbf{D}|\psi_l, \mathcal{M}_l)\pi(\Psi_l|\mathcal{M}_l)}{\int f(\mathbf{D}|\psi_l, \mathcal{M}_l)\pi(\Psi_l|\mathcal{M}_l)d\psi_l}. \tag{3.1}$$

The right-hand side of equation (3.1) consists of the sampling density for the sample data, $f(\mathbf{D}|\psi)$[3], which is proportional to the likelihood function $L(\psi|\mathbf{D})$ (i.e., it defines the functional form of $L(\psi|\mathbf{D})$). The second term in the numerator is $\pi(\Psi)$, the prior distribution regarding the parameter(s) in the model. If $\pi(\Psi|\mathbf{D})$ has the same parametric form as $\pi(\Psi)$, one speaks of the conjugacy property of a prior distribution. From the denominator in (3.1), the marginal density $f(\mathbf{D}) = \int f(\mathbf{D}|\psi)\pi(\Psi)d\psi$, it follows that $\pi(\Psi|\mathbf{D})$ is a proper density function. Taken

3 From now on I will drop the statistical model $\mathcal{M}_l$ for notational convenience.

together, Bayesian statistics understands parameter(s) Ψ as unknown quantities and expresses this uncertainty through probability distributions (for a thorough discussion of the principles of Bayesian statistics, see, e.g., the books of Kaplan (2014) and Gelman et al. (2013)).

In Bayesian estimation, point and interval estimates are calculated by suitable summaries of the entire posterior distribution $\pi(\Psi|\mathbf{D})$. Because the marginal density $f(\mathbf{D})$ solely serves as a normalizing constant, (3.1) can be restated as

$$\pi(\Psi|\mathbf{D}) \propto f(\mathbf{D}|\psi)\pi(\Psi). \tag{3.2}$$

From a practical point of view, the main concern for a Bayesian analyst lies in estimating posterior moments and accordingly posterior integrals. In most applied settings, these intervals are not accessible in closed-form. All models described in Chapter 2 serve as typical examples involving multiple integration in high dimensions. Modern sampling approaches come across this task and approximate the demanded integrals by means of simulation. Chib (2008) concludes: "In short, the problem of computing an intractable integral is reduced to the problem of sampling the posterior density." (p. 483).

3.2 Markov chain Monte Carlo

MCMC methods are designed to produce a sample from complex multivariate distributions. MCMC is based on the idea to construct a Markov chain with the property that its invariant distribution is the target distribution of interest, $\pi(\Psi|\mathbf{D})$ in the case of Bayesian inference. The Monte Carlo principle is applied as numerical problems are approximated via random numbers. An in-depth discussion of the mathematical foundations for MCMC can be found in Tierney (1994). Another rich source of background information on the justification of MCMC is the classical book

by Robert and Casella (2004). Here, I will only briefly sketch the main points.

A sequence of random variables $U = \{U_r\}_{r=1}^{R}$ initialized at U_0 is simulated, in which the conditional distribution of each element U_r depends only on the last element U_{r-1}. A transition kernel T describes the probability of moving to the next state $t^{(1)}$ of the Markov chain given the current state $t^{(0)}$ at some arbitrary step r:

$$T(t^{(0)}, t^{(1)}) = P(U_r = t^{(1)} | U_{r-1} = t^{(0)}).$$

The invariant or stationary distribution $\pi(t)$, for the practically more relevant case of a continuous parameter space, is then defined as satisfying

$$\pi_{r+1}(t^{(1)}) = \int \pi_r(t^{(0)}) T(t^{(0)}, t^{(1)}) dt^{(0)}. \tag{3.3}$$

In order that the distribution of U_r converges to an invariant distribution as $R \to \infty$, the Markov chain has to meet the conditions of irreducibility, aperiodicity and positive recurrence (summarized by the term *ergodic*). As a result, even though the random variables U_r are correlated by definition, the averages of the Markov chain give strongly consistent estimates of any function or statistic under the invariant distribution. Generally, the first draws called *burn-in* period are not expected to follow the target distribution and thus are discarded from calculations.

The averages of a Markov chain are used to estimate the posterior expectation of any function $g(U)$ (for instance, means or quantiles). This property relies on a suitable law of large numbers (see, e.g., Chib, 2001, Section 3; Geweke & Keane, 2001, Section 2) which establishes the fact that ergodic Markov chains satisfy

$$\lim_{R \to \infty} \frac{1}{R} \sum_{r=1}^{R} g(U_r) = E[g(U)]. \tag{3.4}$$

In practice, you will apply a large but finite R. A few aspects regarding the

MCMC chain length should be considered. It is required that the simulated values cover the whole range of $\pi(\Psi|D)$ and consecutive draws move through the support of the distribution. High autocorrelation among the samples have an impact on variance estimates. However, the question remains how to find $T(t^{(0)}, t^{(1)})$ that has $\pi(\Psi|D)$ as its stationary distribution.

3.2.1 Gibbs sampling

The most common MCMC technique applied in Bayesian inference is the Gibbs sampling algorithm (Gelfand & Smith, 1990; Geman & Geman, 1984). It serves as a device to generate samples from the joint posterior distribution $\pi(\Psi|\mathbf{D})$ of the parameter vector Ψ. The core idea is the partition of Ψ into P convenient blocks: $\Psi = \{\Psi_p\}_{p=1}^{P}$. While it may be difficult to sample directly from $\pi(\Psi|\mathbf{D})$, simulating from $f(\Psi_p|\cdot)$ is possible. After determining starting values $\{\psi_p^{(0)}\}_{p=1}^{P}$, the Gibbs sampling scheme iteratively simulates for $r : 1, \ldots, R$ from the full conditional distributions

$$\begin{aligned}
\Psi_1^{(r)} &\sim f(\Psi_1|\psi_2^{(r-1)}, \psi_3^{(r-1)}, \ldots, \psi_P^{(r-1)}, \mathbf{D}) \\
\Psi_2^{(r)} &\sim f(\Psi_2|\psi_1^{(r)}, \psi_3^{(r-1)}, \ldots, \psi_P^{(r-1)}, \mathbf{D}) \\
&\vdots \\
\Psi_P^{(r)} &\sim f(\Psi_P|\psi_1^{(r)}, \psi_2^{(r)}, \ldots, \psi_{P-1}^{(r)}, \mathbf{D})
\end{aligned} \tag{3.5}$$

given the sample data D and the current value of all other parameter blocks. Given a sufficiently large number of iterations R, the procedure constitutes an ergodic Markov chain which converges to the posterior distribution. Note that it is necessary to derive the set of full conditional distributions up to a normalizing constant and afterwards to sample from them. This can be achieved for many statistical models, which explains the widespread popularity of the Gibbs sampling algorithm.

3.2.2 Metropolis-Hastings sampling

Often, one of the conditional densities is not of a known form, which makes it difficult to generate samples according to the respective distribution. Complications could also arise concerning convergence behavior or efficiency considerations. A more general solution to sample a parameter block Ψ_p is offered in terms of the Metropolis-Hastings (M-H) algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Peskun, 1973). M-H sampling generates a candidate draw ψ_p^{cand} from a proposal distribution $q_p(\Psi_p^{\text{cand}}|\psi_p^{(r-1)}, \psi_{-p}^{(r-1)})$, where $\psi_{-p}^{(r-1)} = \{\psi_1^{(r-1)}, \ldots, \psi_{p-1}^{(r-1)}, \psi_{p+1}^{(r-1)}, \ldots, \psi_P^{(r-1)}\}$ denotes the parameter blocks excluding $\psi_p^{(r-1)}$ at the current state $r-1$, and accepts this value as the next state $\psi_p^{(r)}$ with probability

$$\min\left\{1,\ \frac{f(\psi_p^{\text{cand}}|\psi_{-p}^{(r-1)}, \mathbf{D}) q_p(\psi_p^{(r-1)}|\psi_p^{(r-1)}, \psi_{-p}^{(r-1)})}{f(\psi_p^{(r-1)}|\psi_{-p}^{(r-1)}, \mathbf{D}) q_p(\psi_p^{\text{cand}}|\psi_p^{(r-1)}, \psi_{-p}^{(r-1)})}\right\}. \tag{3.6}$$

It is important to note that, again, we only need to know the unnormalized target distribution, because we only evaluate the posterior ratio

$$\frac{f(\psi_p^{\text{cand}}|\psi_{-p}^{(r-1)}, \mathbf{D})}{f(\psi_p^{(r-1)}|\psi_{-p}^{(r-1)}, \mathbf{D})}. \tag{3.7}$$

The second ratio

$$\frac{q_p(\psi_p^{(r-1)}|\psi_p^{(r-1)}, \psi_{-p}^{(r-1)})}{q_p(\psi_p^{\text{cand}}|\psi_p^{(r-1)}, \psi_{-p}^{(r-1)})} \tag{3.8}$$

adjusts for a possibly asymmetric proposal distribution which may prefer certain value regions for ψ_p^{cand}.

As opposed to the Gibbs sampling algorithm, the M-H algorithm does not necessarily accept the proposed candidate draws on every iteration. The acceptance

probability of the Gibbs sampler is always 1 because it equals M-H sampling with the full conditional distribution functioning as the proposal density. If a value gets rejected in the M-H algorithm, the Markov chain remains at the current point $\psi_p^{(r-1)}$.

Optionally, M-H steps can be integrated into a Gibbs sampling scheme resulting in a hybrid Metropolis-within-Gibbs algorithm.

3.2.3 Data augmentation

Another solution to posterior sampling from nonstandard distributions involved in Gibbs sampling is DA. It implies the idea of augmenting the parameter space Ψ with auxiliary variables $\mathbf{D}^*$ to simplify computations for sampling from the posterior $\pi(\Psi|\mathbf{D})$.

Whether the introduction of $\mathbf{D}^*$ facilitates sampling from the full conditional distributions crucially depends on the choice of $f(\mathbf{D}^*|\psi)$. Similar to the models presented in Section 2.3, Albert (1992) developed the 2PNO IRT model as a regression on an underlying continuous latent variable employing the normal ogive/probit link. Albert and Chib (1993) generalized the idea to the binary, the proportional odds and the multinomial probit model. These examples show that appropriately selecting a conditional distribution for the auxiliary variable allows to form a Markov chain via iterative simulations

$$\begin{aligned} \Psi^{(r)} &\sim f(\Psi|\mathbf{D}^{*(r-1)}, \mathbf{D}) \\ \mathbf{D}^{*(r)} &\sim f(\mathbf{D}^*|\psi^{(r)}, \mathbf{D}) \end{aligned} \tag{3.9}$$

for $r : 1, \ldots, R$. Note that this principle resembles the logic of Gibbs sampling in simulating draws from the joint posterior $\pi(\Psi, \mathbf{D}^*|\mathbf{D})$. The fundamental concept in DA is that computing the target density relies on the *posterior identity*

$$\pi(\Psi|\mathbf{D}) = \int \pi(\Psi, \mathbf{D}^*|\mathbf{D}) d\mathbf{D}^* = \int f(\Psi|\mathbf{D}, \mathbf{D}^*) f(\mathbf{D}^*|\mathbf{D}) d\mathbf{D}^*. \tag{3.10}$$

Different DA algorithms and their applications are given in van Dyk and Meng (2001) and Tanner and Wong (1987), who originated the method. As will be shown in the next Section, working with the augmented posterior $\pi(\Psi, \mathbf{D}^*|\mathbf{D})$, i.e., treating auxiliary and latent variables along with the key model parameters, can greatly facilitate Gibbs sampling. Furthermore, DA has proved to be very valuable in the context of missing data models, which I will elaborate on in Section 4.1.

3.3 Estimation algorithms

A combination of the MCMC techniques just presented allows for straightforward estimation of the MGLRM, the FMLRM and the RILRM under the underlying variable formulations (2.14) and (2.19). Aßmann, Gaasch, Pohl, and Carstensen (2016) developed a computational convenient Bayesian sampling scheme for a 1PNO LRM with fixed item difficulties for binary outcomes. I supplement this algorithm through the estimation of a 2PNO measurement model, the integration of ordinal outcomes and the incorporation of hierarchical data structures.

The estimability of IRT models in general is hindered by the nonidentification of parameters (e.g., Fox, 2010, Section 4.4). One way for solving it is to put restrictions on $\{\underline{\xi}_j\}_{j=1}^J$: The sum of item difficulties is required to equal zero and the product of item discriminations (with their inherent property $0 < \alpha_j < \infty$) multiplies to one. To take these identifying restrictions into account, the actual item parameter draws denoted as $\tilde{\alpha}_j$ and $\tilde{\beta}_j$ are transformed to

$$\alpha_j = \tilde{\alpha}_j \left(1 / \prod_{j=1}^{J} \tilde{\alpha}_j \right)^{1/J} \quad \text{and} \quad \beta_j = \tilde{\beta}_j - \sum_{j=1}^{J} \tilde{\beta}_j / J \tag{3.11}$$

after each iteration r (Fox, 2010, p. 88-89).

Yet another issue involves missing values occuring in $\mathbf{Y}$. This corresponds to the case of unbalanced panel data structures in the field of longitudinal analysis. Let $\mathbf{M}^Y$ be a $N \times J$ indicator matrix with elements $m_{i,j}^Y = 1$ if item response $y_{i,j}$ is missing and $m_{i,j}^Y = 0$ if not. I define $\mathbf{Y}_{\text{obs}} = (y_{i,j}, (i,j) : m_{i,j}^Y = 0)$ and $\mathbf{Y}_{\text{mis}} = (y_{i,j}, (i,j) : m_{i,j}^Y = 1)$, so one can write $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}\ \mathbf{Y}_{\text{mis}})$. Unobserved test data $\mathbf{Y}_{\text{mis}}$ can either be recoded into wrong responses or be ignored so that the likelihood is provided only for the observed sample data $\mathbf{Y}_{\text{obs}}$. If $\mathbf{Y}_{\text{obs}}$ is ignored, the valid cases and corresponding parameters have to be selected through a missing indicator matrix during estimation. A further alternative is offered through explicitly modeling the missing process, see Pohl et al. (2014) for a discussion of the different approaches.

The treatment of ordinal items requires to implement the mechanism presented in equation 2.8. Sampling of the item category cutoff parameters exactly resembles the estimation of thresholds in ordinal regression models. In the context of univariate outcomes, Albert and Chib first proposed to draw the thresholds seperately from uniform densities (Albert & Chib, 1993, p. 673). This approach showed a bad mixing and convergence behavior of the Markov chains, especially for large sample sizes (Cowles, 1996). To overcome the problems, Albert and Chib (1997) suggested a joint sampling of the thresholds in a single M-H step after performing a reparametrization of the thresholds according to

$$\begin{aligned} \tau_{j,1} &= \ln\{\kappa_{j,1}\} \\ \tau_{j,q} &= \ln\{\kappa_{j,q} - \kappa_{j,q-1}\}, \qquad q : 2, \ldots, Q_j - 1, \end{aligned} \tag{3.12}$$

with inverse map given by

$$\kappa_{j,q} = \sum_{h=1}^{q} \exp\{\tau_{j,h}\}, \qquad q : 1, \ldots, Q_j - 1. \tag{3.13}$$

Note that the reparametrization leaves the transformed thresholds unordered. As described in the respective sampling step of the estimation algorithms 3.3.1 and 3.3.2 below, a multivariate normal prior distribution can be assigned to the transformed thresholds and the parameters of the proposal density can be adjusted by means of optimization.

3.3.1 Multigroup/finite mixture

The likelihood function for the FMLRM marginalized with respect to θ_i is given as

$$f(\mathbf{Y}|\psi,\mathbf{X}) = \prod_{i=1}^{N}\sum_{g=1}^{G} P(S_i = g|\phi)\left[\int_{\theta_i}\left[\prod_{j=1}^{J}\big(\Phi(\alpha_j\theta_i - (\beta_j + \kappa_{j,y_{ij}})) - \Phi(\alpha_j\theta_i - (\beta_j + \kappa_{j,y_{ij}+1}))\big)\right] \times \frac{1}{\sqrt{2\pi\sigma^2_{\epsilon,g}}}\exp\left\{-\frac{1}{2\sigma^2_{\epsilon,g}}(\theta_i - \underline{x}_{i.}\underline{\gamma}_g)^2\right\} d\theta_i\right], \tag{3.14}$$

where $\psi = \{\{\underline{\xi}_j, \underline{\kappa}_j\}_{j=1}^{J}, \{\underline{\gamma}_g, \sigma^2_{\epsilon,g}\}_{g=1}^{G}, \phi\}$ denotes the entire set of parameters.

In MGLRMs, group membership is known a priori and thus a random assignment of observations to groups becomes unnecessary. Accordingly, parameters ϕ which govern group assignment fall away in the equation. Note that compared to (3.14), the indicator function $\mathbf{1}(S_i = g)$ replaces the probability $P(S_i = g|\phi)$ in the likelihood function. Thus, (3.14) encompasses the MGLRM as a special case and the function can be applied to both models. Augmentation of the parameter set with the underlying continuous outcome and the latent trait variable finally yields the augmented data likelihood for the FMLRM,

$$
\begin{aligned}
f(\mathbf{Y}, \mathbf{Y}^*, \underline{\theta}|\psi, \mathbf{X}) = & \prod_{i=1}^{N} \sum_{g=1}^{G} P(S_i = g|\phi) \left[\prod_{j=1}^{J} \exp\left\{ -\frac{1}{2}(y_{i,j}^* - \alpha_j \theta_i + \beta_j)^2 \right\} \right. \\
& \left. \times \left[\sum_{q=0}^{Q_j - 1} \mathbf{1}(y_{i,j} = q) \mathbf{1}(\kappa_{j,q} < y_{i,j}^* \leq \kappa_{j,q+1}) \right] \right] \\
& \times \frac{1}{\sqrt{2\pi\sigma_{\epsilon,g}^2}} \exp\left\{ -\frac{1}{2\sigma_{\epsilon,g}^2}(\theta_i - \underline{x}_{i.}\underline{\gamma}_g)^2 \right\}.
\end{aligned}
\tag{3.15}
$$

Another identifiability problem comes along with the Bayesian estimation of finite mixture models. *Label switching* (e.g., Frühwirth-Schnatter, 2006, p. 15-16) is caused by the invariance of (3.15) to relabeling of the mixture components. I deal with label switching by ordering the vectors of group-specific regression weights $\underline{\gamma}_g$ according to their intercepts at each iteration r,

$$
\gamma_{1,0}^{(r)} < \ldots < \gamma_{G,0}^{(r)}. \tag{3.16}
$$

Hence, the vector $\underline{\gamma}_g$ with the smallest first element is assigned to group one and so forth. Identifiability constraints on the parameter space are an established procedure in order to break the symmetry of the posterior distribution (Stephens, 2000).

After choosing independent conjugate prior distributions and initializing parameters as outlined in Table 3.1, you obtain draws from the joint posterior density through iteratively sampling (with repetitions $r : 1, \ldots, R$) from the following set of full conditional distributions[4]:

ALGORITHM I

I.1 Sampling from $f(Y_{i,j}^*|\underline{\xi}_j, \underline{\kappa}_j, \theta_i, y_{i,j})$

4 Steps **I.4** and **I.5** are omitted when estimating the MGLRM.

The random variables $Y^*_{i,j}$ are independent and produced from a truncated normal distribution with moments

$$\mu_{Y^*_{i,j}} = \alpha_j\theta_i - \beta_j \tag{3.17}$$
$$\sigma^2_{Y^*_{i,j}} = 1, \tag{3.18}$$

where truncation sphere is $(\kappa_{j,q}, \kappa_{j,q+1})$ for $y_{i,j} = q$.

I.2 Sampling from $f(\underline{\xi}_j|\underline{\theta}, \underline{y}^*_{.j})$

Sampling of working item parameters $\underline{\xi}_j = (\tilde{\alpha}_j, \tilde{\beta}_j)'$ for a single item j is based on the linear regression equation

$$\underline{y}^*_{.j} = \mathbf{T}\underline{\xi}_j + \underline{e}_j, \tag{3.19}$$

where $\mathbf{T}$ is a $N \times 2$ auxiliary matrix consisting of $(\underline{\theta} \ -1)$. If we assume $\underline{e}_j$ normally distributed, it follows a bivariate normal distribution for the item parameters with covariance matrix and mean vector

$$\Sigma_{\xi_j} = \left(\mathbf{T}'\mathbf{T} + \Omega^{-1}_{\xi_j}\right)^{-1} \tag{3.20}$$
$$\underline{\mu}_{\xi_j} = \Sigma_{\xi_j}\left(\mathbf{T}'\underline{y}^*_{.j} + \Omega^{-1}_{\xi_j}\underline{\nu}_{\xi_j}\right). \tag{3.21}$$

I.3 Sampling from $f(\underline{\tau}_j|\underline{\xi}_j, \underline{\theta}, \underline{y}_{.j})$

Draws for the transformed item category cutoff parameters $\underline{\tau}_j$ are retained via a M-H step following Albert and Chib (1997). Given that the chain is currently in state $\underline{\tau}_j$, candidate values $\underline{\tau}^{\text{cand}}_j$ are sampled from a multivariate t_{Q_j-2} proposal density with mean vector $\underline{\mu}_{\underline{\tau}_j} = \widehat{\underline{\tau}}_j$, covariance matrix $\Sigma_{\tau_j} = \widehat{V}$ and ρ degrees of freedom, where $\widehat{\underline{\tau}}_j = \underset{\tau_j}{\text{argmax}} \ln\{f(\underline{y}_{.j}|\underline{\xi}_j, \underline{\kappa}_j, \underline{\theta})\pi(\underline{\tau}_j)\}$ and $\widehat{\mathbf{V}}$

is the inverse of the Hessian of $\ln\{f(\underline{y}_{\cdot j}|\underline{\xi}_j, \underline{\kappa}_j, \underline{\theta})\pi(\underline{\tau}_j)\}$ evaluated at $\widehat{\underline{\tau}}_j$. The probability of accepting candidate values $\underline{\tau}_j^{\text{cand}}$ is computed through the ratio

$$acc_{\tau_j} = \frac{f(\underline{y}_{\cdot j}|\underline{\xi}_j, \underline{\tau}_j^{\text{cand}}, \underline{\theta})\pi(\underline{\tau}_j^{\text{cand}})}{f(\underline{y}_{\cdot j}|\underline{\xi}_j, \underline{\tau}_j, \underline{\theta})\pi(\underline{\tau}_j)} \frac{f_t(\underline{\tau}_j|\widehat{\underline{\tau}}_j, \widehat{\mathbf{V}}, \rho)}{f_t(\underline{\tau}_j^{\text{cand}}|\widehat{\underline{\tau}}_j, \widehat{\mathbf{V}}, \rho)} \tag{3.22}$$

and is given by $\min(1, acc_{\tau_j})$. Conversely, the probability of remaining at $\underline{\tau}_j^{(r-1)}$ equals $1 - \min(1, acc_{\tau_j})$.

I.4 Sampling S_i

The individual latent group indicator is drawn from a multinomial distribution, whose group probabilities correspond to the full conditional probability

$$\begin{aligned} P(S_i = g|\{\xi_j\}_{j=1}^J, \underline{\gamma}_g, \sigma^2_{\epsilon,g}, \{\underline{\zeta}_g\}_{g=1}^{G-1}, \theta_i, \underline{y}^*_{i\cdot}, \underline{x}_{i\cdot}, \underline{z}_{i\cdot}) \propto \\ \left[\prod_{j=1}^J \exp\left\{-\frac{1}{2}(y^*_{i,j} - \alpha_j\theta_i + \beta_j)^2\right\}\right] \\ \times \frac{1}{\sqrt{2\pi\sigma^2_{\epsilon,g}}} \exp\left\{-\frac{1}{2\sigma^2_{\epsilon,g}}(\theta_i - \underline{x}_{i\cdot}\underline{\gamma}_g)^2\right\} \\ \times \frac{\exp(\underline{z}_{i\cdot}\underline{\zeta}_g)}{1 + \sum_{g=1}^{G-1}\exp(\underline{z}_{i\cdot}\underline{\zeta}_g)}. \end{aligned} \tag{3.23}$$

I.5 Sampling from $f(\underline{\boldsymbol{\zeta}}|\underline{S})$

Draws for all parameters governing the cluster probabilities $\underline{\boldsymbol{\zeta}} = \{\underline{\zeta}_g\}_{g=1}^{G-1}$ are retained via a M-H step. Candidate values $\underline{\boldsymbol{\zeta}}^{\text{cand}}$ are sampled from a $(G-1)(M+1)$-dimensional multivariate normal proposal density with mean vector $\underline{\mu}_\zeta = \widehat{\underline{\boldsymbol{\zeta}}}$ and covariance matrix $\Sigma_\zeta = \widehat{\mathbf{W}}$, where $\widehat{\underline{\boldsymbol{\zeta}}} = \underset{\zeta}{\text{argmax}}\ \ln\{f(\underline{S}|\underline{\boldsymbol{\zeta}})\pi(\underline{\boldsymbol{\zeta}})\}$ and $\widehat{\mathbf{W}}$ is the inverse of the Hessian of $\ln\{f(\underline{S}|\{\underline{\zeta}_g\}_{g=1}^G)\pi(\{\underline{\zeta}_g\}_{g=1}^G)\}$ evaluated at $\widehat{\underline{\boldsymbol{\zeta}}}$. The probability of accepting candidate values $\underline{\boldsymbol{\zeta}}^{\text{cand}}$ is computed through the ratio

$$acc_{\zeta} = \frac{f(\{S_i\}_{i=1}^{N}|\underline{\zeta}^{\text{cand}})\pi(\underline{\zeta}^{\text{cand}})}{f(\{S_i\}_{i=1}^{N}|\{\underline{\zeta}_g\}_{g=1}^{G})\pi(\zeta)} \frac{f_{\mathcal{N}}(\{\underline{\zeta}_g\}_{g=1}^{G}|\widehat{\underline{\zeta}}, \widehat{\mathbf{W}})}{f_{\mathcal{N}}(\underline{\zeta}^{\text{cand}}|\widehat{\underline{\zeta}}, \widehat{\mathbf{W}})} \tag{3.24}$$

and is given by $\min(1, acc_{\zeta})$.

I.6 Sampling from $f(\theta_i|\{\underline{\xi}_j\}_{j=1}^{J}, \{\gamma_g, \sigma^2_{\epsilon,g}\}_{g=1}^{G}, \underline{y}^*_{i.}, S_i)$

Let $\underline{B}_{i.} = \underline{y}^*_{i.} + \underline{\beta}$, where $\underline{\beta}$ is a vector including all item difficulties. This allows for stating the conditional distribution of the individual abilities as normal with moments

$$\sigma^2_{\theta_i} = \left(\underline{\alpha}'\underline{\alpha} + \sigma^{-2}_{\epsilon,S_i}\right)^{-1} \tag{3.25}$$

$$\mu_{\theta_i} = \sigma^2_{\theta_i}\left(\underline{\alpha}'\underline{B}_{i.} + \sigma^{-2}_{\epsilon,S_i}\underline{x}'_{i.}\underline{\gamma}_{S_i}\right), \tag{3.26}$$

where $\underline{\alpha}$ is a vector which contains all item discrimination parameters.

I.7 Sampling from $f(\underline{\gamma}_g|\underline{\theta}_{[g]}, \sigma^2_{\epsilon,g}, \mathbf{X}_{[g.]})$

Let the indices $[g]$ and $[g.]$ select the elements of $\underline{\theta}$, respectively the rows of $\mathbf{X}$ for which the condition $S_i = g$ holds. Further, let Σ_{ϵ} be a $N_g \times N_g$ diagonal matrix with elements $\sigma^2_{\epsilon,g}$. Draws from the conditional distribution of $\underline{\gamma}_g$ are obtained from a multivariate normal with covariance matrix and mean vector

$$\Sigma_{\gamma_g} = \left(\mathbf{X}'_{[g.]}\Sigma^{-1}_{\epsilon}\mathbf{X}_{[g.]} + \Omega^{-1}_{\gamma_g}\right)^{-1} \tag{3.27}$$

$$\underline{\mu}_{\gamma_g} = \Sigma_{\gamma_g}\left(\mathbf{X}'_{[g.]}\Sigma^{-1}_{\epsilon}\underline{\theta}_{[g]} + \Omega^{-1}_{\gamma_g}\nu_{\gamma_g}\right). \tag{3.28}$$

I.8 Sampling from $f(\sigma^2_{\epsilon,g}|\underline{\theta}_{[g]}, \underline{\gamma}_g, \mathbf{X}_{[g.]})$

Choosing the conjugate prior, $\sigma^2_{\epsilon,g}$ is distributed inverse gamma with shape and scale parameter

$$a_{\sigma^2_{\epsilon,g}} = a^0_{\sigma^2_{\epsilon,g}} + N_g/2 \tag{3.29}$$

$$b_{\sigma^2_{\epsilon,g}} = \Big(b^0_{\sigma^2_{\epsilon,g}} + 0.5(\underline{\theta}_{[g]} - \mathbf{X}_{[g.]}\underline{\gamma}_g)' \times (\underline{\theta}_{[g]} - \mathbf{X}_{[g.]}\underline{\gamma}_g)\Big)^{-1}. \tag{3.30}$$

3.3.2 Random intercept

Summarizing all parameters as $\psi = \{\{\underline{\xi}_j, \underline{\kappa}_j\}_{j=1}^J, \underline{\gamma}, \sigma^2_\epsilon, \upsilon^2_\omega\}$ and integrating out the random effects θ_i and ω_c yields the likelihood function

$$\begin{aligned} f(\mathbf{Y}|\psi, \mathbf{X}) = \prod_{c=1}^{C} \int_{\omega_c} \Bigg[\prod_{i=1}^{N_c} \int_{\theta_{c,i}} \Bigg[\prod_{j=1}^{J} \Big(& \Phi(\alpha_j\theta_{c,i} - (\beta_j + \kappa_{j,y_{c,i,j}})) \\ & - \Phi(\alpha_j\theta_{c,i} - (\beta_j + \kappa_{j,y_{c,i,j}+1})) \Big) \Bigg] \\ & \times \frac{1}{\sqrt{2\pi\sigma^2_\epsilon}} \exp\left\{ -\frac{1}{2\sigma^2_\epsilon}(\theta_{c,i} - \omega_c - \underline{x}_{c,i}\underline{\gamma})^2 \right\} d\theta_{c,i} \Bigg] \\ & \times \frac{1}{\sqrt{2\pi\upsilon^2_\omega}} \exp\left\{ -\frac{\omega_c^2}{2\upsilon^2_\omega} \right\} d\omega_c. \end{aligned} \tag{3.31}$$

Augmentation of the parameter set with $\mathbf{Y}^*, \underline{\theta}$ and $\underline{\omega}$ results in the augmented data likelihood

$$\begin{aligned} f(\mathbf{Y}, \mathbf{Y}^*, \underline{\theta}, \underline{\omega}|\psi, \mathbf{X}) = \prod_{c=1}^{C} \prod_{i=1}^{N_c} \Bigg[\Bigg[& \prod_{j=1}^{J} \exp\left\{ -\frac{1}{2}(y^*_{c,i,j} - \alpha_j\theta_{c,i} + \beta_j)^2 \right\} \\ & \times \left[\sum_{q=0}^{Q_j-1} \mathbf{1}(y_{c,i,j} = q)\mathbf{1}(\kappa_{j,q} < y^*_{c,i,j} \leq \kappa_{j,q+1}) \right] \Bigg] \\ & \times \frac{1}{\sqrt{2\pi\sigma^2_\epsilon}} \exp\left\{ -\frac{1}{2\sigma^2_\epsilon}(\theta_{c,i} - \omega_c - \underline{x}_{c,i.}\underline{\gamma})^2 \right\} \Bigg] \\ & \times \frac{1}{\sqrt{2\pi\upsilon^2_\omega}} \exp\left\{ -\frac{\omega_c^2}{2\upsilon^2_\omega} \right\}. \end{aligned} \tag{3.32}$$

Note that the changes in (3.32) compared to (3.15) only concern the structural

component. Sampling of item parameters remains unchanged and is identical to ALGORITHM I. After choosing independent conjugate prior distributions as outlined in Table 3.1 and initializing parameters, you obtain draws from the joint posterior density through iteratively sampling (with repetitions $r : 1, \ldots, R$) from the following set of full conditional distributions:

ALGORITHM II

II.1 according to **I.1**

II.2 according to **I.2**

II.3 according to **I.3**

II.4 Sampling from $f(\theta_{c,i}|\{\xi_j\}_{j=1}^J, \underline{\gamma}, \sigma_\epsilon^2, \omega_c, \underline{y}^*_{c,i})$

Let $\underline{B}_{c,i.} = \underline{y}^*_{c,i.} - \underline{\beta}$, where $\underline{\beta}$ is a vector including all item difficulties. This allows for stating the conditional distribution of the individual abilities as normal with moments

$$\sigma^2_{\theta_{c,i}} = \left(\underline{\alpha}'\underline{\alpha} + \sigma_\epsilon^{-2}\right)^{-1} \tag{3.33}$$

$$\mu_{\theta_{c,i}} = \sigma^2_{\theta_{c,i}} \left(\underline{\alpha}'\underline{B}_{c,i.} + \sigma_\epsilon^{-2}(\underline{x}'_{c,i.}\underline{\gamma} + \omega_c)\right), \tag{3.34}$$

where $\underline{\alpha}$ is a vector which contains all item discrimination parameters.

II.5 Sampling from $f(\underline{\gamma}|\underline{\theta}, \sigma_\epsilon^2, \underline{\omega}, \mathbf{X}, \mathbf{Q})$

Let Σ_ϵ be a $N \times N$ diagonal matrix with elements $\sigma^2_{\epsilon,g}$ and $\mathbf{Q}$ be a $N \times C$ design matrix of zeros. Each row of $\mathbf{Q}$ has a single entry 1 indicating the respondents' cluster membership. Draws from the conditional distribution of $\underline{\gamma}$ are obtained from a multivariate normal with covariance matrix and mean vector

$$\Sigma_\gamma = \left(\mathbf{X}'\Sigma_\epsilon^{-1}\mathbf{X} + \Omega_\gamma^{-1}\right)^{-1} \tag{3.35}$$
$$\mu_\gamma = \Sigma_\gamma\left(\mathbf{X}'\Sigma_\epsilon^{-1}(\underline{\theta} - \mathbf{Q}\omega) + \Omega_\gamma^{-1}\nu_\gamma\right). \tag{3.36}$$

II.6 Sampling from $f(\sigma_\epsilon^2|\underline{\theta}, \underline{\gamma}, \underline{\omega}, \mathbf{X}, \mathbf{Q})$

Choosing the conjugate prior, σ_ϵ^2 is distributed inverse gamma with shape and scale parameter

$$a_{\sigma_\epsilon^2} = a^0_{\sigma_\epsilon^2} + N/2 \tag{3.37}$$
$$b_{\sigma_\epsilon^2} = \left(b^0_{\sigma_\epsilon^2} + 0.5(\underline{\theta} - \mathbf{X}\gamma - \mathbf{Q}\underline{\omega})'(\underline{\theta} - \mathbf{X}\gamma - \mathbf{Q}\underline{\omega})\right)^{-1}. \tag{3.38}$$

II.7 Sampling from $f(\omega_c|\underline{\theta}, \underline{\gamma}, \sigma_\epsilon^2, \upsilon_\omega^2, \mathbf{X})$

Let the indices $[c]$ and $[c.]$ select the elements of $\underline{\theta}$, respectively the rows of $\mathbf{X}$ belonging to cluster c. The cluster-specific random intercepts follow a normal distribution with moments

$$\sigma^2_{\omega_c} = \left(\upsilon_\omega^{-2} + N_c/\sigma_\epsilon^2\right)^{-1} \tag{3.39}$$
$$\mu_{\omega_c} = \sigma^2_{\omega_c}\left(\sigma_\epsilon^{-2}(\underline{\theta}_{[c]} - \mathbf{X}_{[c.]}\underline{\gamma})'(\underline{\theta}_{[c]} - \mathbf{X}_{[c.]}\underline{\gamma})\right). \tag{3.40}$$

II.8 Sampling from $f(\upsilon_\omega^2|\underline{\omega})$

Choosing the conjugate prior, υ_ω^2 is distributed inverse gamma with shape and scale parameter

$$a_{\upsilon_\omega^2} = a^0_{\upsilon_\omega^2} + C/2 \tag{3.41}$$
$$b_{\upsilon_\omega^2} = \left(b^0_{\upsilon_\omega^2} + 0.5\underline{\omega}'\underline{\omega}\right)^{-1}. \tag{3.42}$$

The purpose of this Chapter is to set up two Metropolis-within-Gibbs sampling algorithms to estimate the MGLRM/FMLRM and the RILRM. Having completely

observed the matrices of person covariates $\mathbf{X}$ and $\mathbf{Z}$, the estimation schemes allow effectively for Bayesian inference on the corresponding parameter vector. If missing values are present in $\mathbf{X}$ or $\mathbf{Z}$, the analyst has to think of an appropriate missing data technique to complete the matrices prior to estimation. In the next Chapter I will show how the imputation of missing values can be incorporated into both algorithms, again using the device of DA.

4 The case of missing values in person covariates

> There is an old saying "If all a man has is a hammer, then every problem looks like a nail." The trouble for statisticians is that recently some of the problems have stopped looking like nails. (Breiman, 2001, p. 204)

4.1 Data augmentation continued

Context questionnaires in LSAs are almost always affected by item nonresponse. For instance, respondents may refuse to answer questions due to privacy concerns or fatigue effects. The impact of partially missing covariates on student ability estimates has been studied by Rutkowski (2011) using simulated data that mimic a multiple matrix sampling assessment design. She found a shift in the ability distribution for subgroups defined by a background variable with values missing at random (see explanation below) when dummy-coding is applied.

How can the nonresponse be treated instead? The default option in many statistical software packages is *complete cases analysis* (CC; also known as listwise deletion) which excludes all observations having a missing value on any covariate from estimation. Beside the inefficient use of the sample in situations with high rates of missingness, the method may give biased estimates, especially when observations are missing at random (Little & Rubin, 2002, p. 41-44).

MI has evolved into the contemporary solution for treating incomplete data in a variety of research fields like, e.g., epidemiology and medical statistics (Rässler, Rubin, & Zell, 2008) or data fusion problems (Rässler, 2002). Reiter and Raghu-

nathan (2007) give a general overview of the main adaptations of MI. The framework is introduced by (Rubin, 1976, 1978) and its theoretical foundations are explained thoroughly in Rubin (1987). Before missing data can be (multiply) imputed, a model for the *missing data mechanism* needs to be set up. Rubin distinguishes three different distributions:

- Missing data are missing completely at random (MCAR) if the probability that a missing value occurs is equal for every observation. Thus, the missingness happens completely random and does not depend on any observed data.
- Missing data are missing at random (MAR) if the missingness depends on other observed variables. For example, students in the lower educational track may refuse more often to respond to a school satisfaction scale than students in the intermediate and upper educational tracks.
- Missing data are not missing at random (NMAR) if the probability that a missing value occurs is related to the variable itself. Students who stay away from school often could rather avoid to answer questions about their absenteeism. This mechanism is also referred to as *nonignorable*.

The MAR mechanism, which plays an essential role in MI, is formally expressed as

$$f(\mathbf{M}^X|\mathbf{X}, \Xi) = f(\mathbf{M}^X|\mathbf{X}_{\text{obs}}, \Xi) \; \forall \mathbf{X}_{\text{mis}}, \Xi, \tag{4.1}$$

where $\mathbf{M}^X$ is defined a $N \times K$ matrix indicating the missing ($m^X_{i,k} = 1$) and observed ($m^X_{i,k} = 0$) parts of covariate matrix $\mathbf{X} = (\mathbf{X}_{\text{obs}}\ \mathbf{X}_{\text{mis}})$ corresponding to the definitions for $\mathbf{M}^Y$ in Section 3.3 and Ξ are the unobserved parameters governing the missing data mechanism.[5]

5 In the following, the notation used refers to the MGLRM.

Additionally, if the assumption

$$f(\Psi, \Xi) = f(\Psi) f(\Xi) \tag{4.2}$$

holds (i.e., prior independence) then one speak of the *distinctness* between Ξ and the MGLRM parameters. Under the MAR and the distinctness assumptions Bayesian inference for the MGLRM can be made based on the *observed data posterior density* $\pi(\Psi, \mathbf{Y}^*, \underline{\theta}|\mathbf{Y}, \mathbf{X}_{\text{obs}})$ *ignoring* the missing data mechanism (Little & Rubin, 2002, p. 120), because

$$\begin{aligned} \pi(\Psi, \mathbf{Y}^*, \underline{\theta}, \Xi|\mathbf{Y}, \mathbf{X}_{\text{obs}}, \mathbf{M}) &\propto [f(\mathbf{Y}, \mathbf{Y}^*, \underline{\theta}|\psi, \mathbf{X}_{\text{obs}})\pi(\Psi)][f(\mathbf{M}|\mathbf{X}_{\text{obs}}, \Xi)\pi(\Xi)] \\ &\propto \pi(\Psi, \mathbf{Y}^*, \underline{\theta}|\mathbf{Y}, \mathbf{X}_{\text{obs}})\pi(\Xi|\mathbf{X}_{\text{obs}}, \mathbf{M}). \end{aligned} \tag{4.3}$$

Moreover, the *missing data pattern* critically affects the way MI can be conducted. There are two main patterns concerning item nonresponse in multivariate settings. Monotone patterns describe the case where columns of $\mathbf{X}$ can be ordered by occurence of the nonresponse. They require the positions of missing values for column $\underline{x}_{\cdot k}$ to be a subset of the positions of missing values for all succeeding columns $\underline{x}_{\cdot k'}$ $(k' > k)$ if covariates are sorted by the amount of missing data in ascending order. Among other desirable properties of monotone patterns, the full conditional distribution of missing values in variable $\underline{x}_{\cdot k}$ is defined given all completely observed columns and variables in columns 1 to $k-1$ (Little & Rubin, 2002, p. 144). On the contrary, nonmonotone patterns reveal an arbitrary or generic drop out schema. Data structures like latent trait variables or latent group membership introduced in Chapter 2 may also be regarded as an own pattern of missingness. However, opposed to item nonresponse, there is no chance at all of observing them.

In the context of LRMs with partially missing covariate data it becomes necessary to integrate the latent trait vector $\underline{\theta}$ into the imputation model for $\mathbf{X}_{\text{mis}}$ and

thereby avoid a missmatch of the imputation and the analysis model. Instead of estimating an IRT model in a preceding step and using the resultant scores $\widehat{\underline{\theta}}$ together with $\mathbf{X}$ for MI, I choose to make inferences about $\underline{\theta}$ and at the same time impute missing covariates $\mathbf{X}_{\text{mis}}$. This further brings the advantage that hierarchical model parameters reflecting the stratification of the data may also be used for imputation purposes. Presuming ignorability as stated by (4.3), the DA algorithm facilitates the set up of such an estimation scheme. Writing the observed data posterior as

$$\begin{aligned}
\pi(\Psi, \mathbf{Y}^*, \underline{\theta}|\mathbf{Y}, \mathbf{X}_{\text{obs}}) &= \int \pi(\Psi, \mathbf{Y}^*, \underline{\theta}, \mathbf{X}_{\text{mis}}|\mathbf{Y}, \mathbf{X}_{\text{obs}}) d\mathbf{X}_{\text{mis}} \\
&= \int \pi(\Psi, \mathbf{Y}^*, \underline{\theta}|\mathbf{Y}, \mathbf{X}) f(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}) d\mathbf{X}_{\text{mis}} \\
&\propto \int f(\mathbf{Y}, \mathbf{Y}^*, \underline{\theta}|\psi, \mathbf{X}) \pi(\Psi) f(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}) d\mathbf{X}_{\text{mis}}, \qquad (4.4)
\end{aligned}$$

where $\Psi = \{\{\underline{\xi}_j, \underline{\kappa}_j\}_{j=1}^J, \{\underline{\gamma}_g, \sigma^2_{\epsilon,g}\}_{g=1}^G, \phi\}$, samples from the joint posterior

$$\pi(\Psi, \mathbf{Y}^*, \underline{\theta}, \mathbf{X}_{\text{mis}}|\mathbf{Y}, \mathbf{X}_{\text{obs}}) \qquad (4.5)$$

are obtained using the Gibbs sampler (3.9) with $\mathbf{D}^* = \mathbf{X}_{\text{mis}}$. Thus, in every iteration r, all model parameters involved in (3.15) and (3.32) are sampled according to the existing MCMC schemes ALGORITHM I and ALGORITHM II given a filled-in matrix $\mathbf{X}$. Then missing values in $\mathbf{X}$ are completed with draws from their full conditional distributions. This updated version of $\mathbf{X}$ is used in the succeeding iteration $r+1$ to start the sequence again and re-estimate model parameters. Reaching convergence, one obtains approximate draws from the respective marginal posterior densities. In the upcoming Section, I will describe the procedure utilized during the imputation step for $\mathbf{X}_{\text{mis}}$ in more detail.

4.2 Sequential classification and regression trees as an imputation tool

In order to establish highly flexible approximations to the distributions of missing values in covariates, Burgette and Reiter (2010) propose to adopt CART for the construction of conditional imputation models. Schenker and Taylor (1996) were actually the first to suggest that data mining methods could be used for imputation purposes. See also the general remark on CART as an imputation engine by Hastie, Tibshirani, and Friedman (2009, p. 333) and Reiter (2005) for an application in the context of partially synthetic data generation. The flexibility of CART to incorporate nonlinear dependencies among the variables with missing data has been further highlighted by Doove et al. (2014). Recently, a related approach has been successfully applied to take individual skip patterns into account during income imputation in the adult cohort of the NEPS by Aßmann, Würbach, Goßmann, Geisser, and Bela (2015).

CART is a popular and widely used *data mining* tool which goes back to Breiman, Friedman, Olshen, and Stone (1984). Typically, it aims at the classification of observations and making predictions based on the independent variables involved in building the tree. Instead of specifying a parametric form which connects the dependent variable with a set of covariates (e.g., through a regression function), CART rather behaves as an adaptive heuristic. Its functioning is now illustrated using a short example.

The `kyphosis` data frame available via the R package *rpart* (Therneau, Atkinson, & Ripley, 2015) holds data on 81 children who have had corrective spinal surgery. A tree is fitted to all patients where the binary variable *Kyphosis*, indicating if a kyphosis was present after the operation (absent versus present), serves as the outcome. The predictors are age in month (*Age*), number of vertebrae involved

(*Number*) and number of the topmost vertebra operated on (*Start*). Classification of children is achieved through recursively partitioning the data into mutually exclusive groups (called *nodes* in the CART terminology). The assignment to a group fulfills the condition that within a node the best achievable homogeneity with regard to the outcome is revealed. At the same time intergroup heterogeneity is maximized. The partitioning is based on binary splits in one selected predictor variable. Opposed to other machine learning algorithms like the C4.5 (Quinlan, 1993), CART only conducts binary splits in continuous as well as in categorical covariates. Returning to the example, the right and left son/daughter nodes are defined by variable *Start* (Figure 4.1). If $Start_i < 8.5$ for obervation i, this child is classified a present kyphosis. Children with a value greater than 8.5 enter the left son/daughter node. Further splitting of the resulting partitions is continued similiarly until perfect node purity is achieved or an abort criterion is fulfilled. The choice of predictor variables considered for splitting and the strategy for finding optimal cutpoints are based on a suitable measure of node impurity. CART utilizes the empirical Gini index which reduces to

$$2\frac{N_{\text{kyphosis}}}{N}\left(1-\frac{N_{\text{kyphosis}}}{N}\right), \tag{4.6}$$

where N_{kyphosis} is the number of children with a kyphosis within a node (this always holds for binary outcomes). Note that differences in node impurity must be calculated before and after a possible split. Continuing with the classification tree for the kyphosis data, the best split in the left son/daughter node is again determined through the predictor *Start*. Children with large numbers of the topmost verbetra operated on, i.e., $Start_i > 14$, finally end up in an abscent kyphosis. CART carries on allocating the remaining observations to a node according to the logic just explained.

Burgette and Reiter (2010) use CART for specifying the imputation model within the MICE framework by van Buuren and Groothuis-Oudshoorn (2011), which is, in a slightly different formulation, also known as the sequential regressions approach (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). The basic idea of MICE is to define a full conditional model for each variable $\underline{x}_{\cdot k}$ plagued by missing values seperately without presupposing a joint distribution. After initializing the missing values, an appropriate model (according to the level of measurement) is fitted to all respondents originally having an observed value for $\underline{x}_{\cdot k}$. Imputations are then generated from the corresponding posterior predictive distribution which leads to a Gibbs-like algorithm. A complete cycle comprising all variables having missing values is repeated for several iterations to stabilize the results. Note that previous imputations do not directly enter the imputation model but updated versions of each variable with missing values are added sequentially to the full conditional distributions of the subsequent columns of $\mathbf{X}$. One problem of the MICE approach is that it is not known if the joint distribution even exists. If the rearranging of $\mathbf{X}$ by the number of missing values yields a monotone missing data pattern, the joint distribution exists given that the sequential imputations are performed from left to right.

Let the partially observed columns of $\mathbf{X}$ be ordered so that the missing rate in each column is nondecreasing from left to right. Given starting values as unconditional draws from the observed values, replacements for missing values are created subsequently for each column in two steps:

a) A tree is built based on all remaining variables in $\mathbf{X}$ plus the updated draws from posterior sampling taking into account only the complete observations. The resulting binary partition of the data along the set of predictors provides the outcome pools of possible donors. In this way the nonparametric

characterization of the full conditional distribution is defined.

b) Every respondents with a missing value is assigned to one of these identified donor groups. Draws from the empirical distribution within the end nodes are finally obtained using the Bayesian bootstrap (Rubin, 1981).

The application of CART to approximate the full conditional distributions of missing values is particularly useful because the analyst does not need to specify the imputation models. Further, the hot-deck imputation technique ensures valid substitutes taken from the empirical distribution, so that the regression coefficients in the structural model remain interpretable in contrast to the dummy-coding method.

The following sampling steps are added to ALGORITHM I and ALGORITHM II resulting in data augmentation using sequential CART imputation (DAC) as described above:

I.9 Sampling from $f(\mathbf{X}_{\text{mis}}|\underline{\theta}, \underline{S}, \mathbf{X}_{\text{obs}})$

II.9 Sampling from $f(\mathbf{X}_{\text{mis}}|\underline{\theta}, \mathbf{Q}\underline{\omega}, \mathbf{X}_{\text{obs}})$

4.3 Simulation studies

I set up two simulation studies to assess the statistical accuracy of the proposed strategy and compare it with two alternative methods for the treatment of missing covariate data (see Appendix C for the program code to run the simulation studies and Appendix D for detailed information on the computer software and hardware used). For each scenario, a single data generating process and missing data mechanism are generated. The different estimation procedures being checked against each other are then conducted for 200 replications of these. In the first scenario, only continuous person covariates are included. DA is employed either via DAC or

using sequential stochastics regression imputations (DAR) to approximate the full conditional distributions of missing values. Scenario 2 considers continuous as well as categorical person covariates and compares DAC with a CC. The two scenarios additionally differ in the severity of missingness.

Both data generating processes satisfy the following conditions: A response matrix $\mathbf{Y}$ is simulated assuming a MGLRM according to equations (2.14) with a sample setup of $N = 2{,}000$ respondents allocated equally to $G = 2$ groups. The respondents face a test of altogether $J = 20$ items of which the first 18 are binary and the last two are ordinal with $Q_j = 4$ categories. Item parameters are fixed across replications and were generated once via $\tilde{\alpha}_j = 1 + \mathcal{U}(-0.3, 0.3)$ and $\tilde{\beta}_j = 0 + \mathcal{U}(-0.7, 0.7)$, where $\mathcal{U}(u, v)$ is a continuous uniformly distributed random number in the interval $[u, v]$. To fulfill the identifying restrictions $\sum_{j=1}^{J} \beta_j = 0$ and $\prod_{j=1}^{J} \alpha_j = 1$, item difficulty and discrimination parameters are finally derived according to (3.11). The item category offset parameters for the two ordinal items are set to $\kappa_{19} = (0, 0.5, 1.6)'$ and $\kappa_{20} = (0, 0.7, 1.2)'$. So far, simulations are identical for both scenarios. The varying specifications of the latent trait distribution are given in Sections 4.3.1 and 4.3.2.

Tables 4.1 and 4.3 provide the true parameter values and mean posterior moments over the 200 replications obtained from the full sample estimates before deletion (BD) and the different estimation procedures. Beside the averaged estimates, simulation results are evaluated in terms of the root mean square error (RMSE) and the proportion of 95% highest posterior density regions (HDRs) that contain the true parameter values[6] (coverage). RMSEs and coverages for the different estima-

6 For example, the approximate 99% confidence interval for a binomial proportion 0.95, i.e., $0.95 \pm q_{0.995}\sqrt{0.95(1-0.95)/200}$ ($200 \times 0.95(1-0.95) = 9.5 > 9$), allows the coverages to lie between 0.91 and 0.99 for all parameters, where q is the quantile function of the standard normal distribution.

tion procedures are presented in Tables 4.2 and 4.4.

4.3.1 Comparison with stochastic regression imputation

In scenario 1, two person covariates X_k $(k : 1, 2)$ explaining differences in latent trait θ_i are generated from a multivariate normal distribution, where the variables each have a mean of 1, a variance of 4 and a correlation equal to 0.5. The corresponding parameters of the population model including two intercepts are set to $\underline{\gamma}_1 = (-0.5, 0.2, 0.2)'$, $\underline{\gamma}_2 = (1, 0.4, -0.2)'$, $\sigma^2_{\epsilon,1} = 0.7^2$ and $\sigma^2_{\epsilon,2} = 0.5^2$.

Then, observations in X_1 and X_2 are deleted via MCAR according to

$$P(X_{i,1} = \text{“missing”}) = P(X_{i,2} = \text{“missing”}) = 0.1, \tag{4.7}$$

which results, on average, in 10% missing values for both variables.

Whereas the algorithm DAC was already presented in Section 4.2, DAR proceeds as follows. Instead of the sequential CART imputation step, a univariate normal full conditional distribution is specified for each variable X_k. After the missing values in $\mathbf{X}$ are completed by imputations, the following two regression equations can be calculated:

$$\underline{x}_{.k} = \mathbf{W}_k \underline{\varphi}_k + \underline{e}_k \qquad \text{with } \underline{e}_k \sim \mathcal{N}(0, \sigma^2_{e,k} \mathcal{I}_{N_{\text{mis}}}), \qquad \text{for } k : 1, 2, \tag{4.8}$$

where $\mathbf{W}_k = (1 \ \mathbf{X}_{-k} \ \underline{\theta} \ \underline{S})$ and $\{\underline{\varphi}_k, \sigma^2_{e,k}\}^2_{k=1}$ are the usual regression parameters. The least squares estimators $\widehat{\underline{\varphi}}_k$ and $\widehat{\sigma}^2_e$ are then used to generate new imputations. Each originally missing value in X_k is replaced by a random normal sample with mean $\mathbf{W}_{k[\text{mis}]} \widehat{\varphi}_k$ and variance $\widehat{\sigma}^2_e$. This imputation procedure is referred to as *stochastic regression imputation* (see, e.g., van Buuren, 2012, p. 13) and belongs to the traditional methods to handle missing data.

While the individual latent traits are initialized to random draws from a stan-

dard normal distribution, I adopt the following starting values and vague prior specifications about parameter blocks $\{\underline{\gamma}_g, \sigma^2_{\epsilon,g}\}^2_{g=1}$, $\{\underline{\xi}_j\}^{20}_{j=1}$ and $\{\underline{\tau}_j\}^{20}_{j=19}$ (see Table 3.1 for the prior distributions):

- $\underline{\gamma}_g^{(0)} = (0,0,0)'$, $\underline{\nu}_{\gamma_g} = (0,0,0)'$ and $\Omega_{\gamma_g} = 100\mathcal{I}_3$.
- $\sigma^{2(0)}_{\epsilon,g} = 1$, $a^0_{\sigma^2_{\epsilon,g}} = 1$ and $b^0_{\sigma^2_{\epsilon,g}} = 1$.
- $\underline{\xi}_j^{(0)} = (1,0)'$, $\underline{\nu}_{\xi_j} = (0,0)'$ and $\Omega_{\xi_j} = 100\mathcal{I}_2$.
- $\underline{\tau}_j^{(0)} = (0,0)'$, $\underline{\nu}_{\tau_j} = (0,0)'$ and $\Omega_{\tau_j} = 100\mathcal{I}_2$.

Each of the repeated estimations is based on MCMC chains of length $R =$ 12,000. After discarding the first 2,000 iterations as burn-in and retaining every second iteration, inference is finally made on the remaining 5,000 simulated draws from the joint posterior distribution. This requires a total run time of under eight hours.

Table 4.1 shows the true parameter values, mean posterior means and standard deviations over 200 replications obtained from BD, DAC and DAR. For the BD estimates you find overall unbiased results for all parameters. The results indicate a correct implementation of the algorithm and further serve as a benchmark to assess the relative performance of the different imputation methods in the case of partially missing covariate data. Similar results are revealed for my suggested approach. There is no notable difference between columns BD and DAC of the table. In contrast, the stochastic regression imputation method reported in columns DAR leads to biased estimates of the structural parameters. For example, the bias for coefficient $\gamma_{2,1}$ is -0.038 and for variance $\sigma^2_{\epsilon,2}$ adds up to 0.1. From the mean posterior standard deviations in the second block of Table 4.1 one can see that the structural parameter estimates obtained from DAR have slightly higher uncertainty.

Of course, the item parameter estimates do not differ across BA, DAC and DAR because the corresponding full conditional distributions do not involve $\mathbf{X}$.

Turning now to Table 4.2, the findings further demonstrate the advantage of DAC compared to DAR. We can see the RMSEs in accordance with the averaged posterior standard deviations for BD and DAC, which highlights the statistical accuracy of my approach. The significantly higher values concerning DAR result from the biases and larger variances of the estimates. For DAC, the observed number of HDRs covering the particular true parameter approximately conform with BD and the expected theoretical values. In column DAR, the coverages drop to 0.04 for variance $\sigma^2_{\epsilon,2}$.

It should be noted that stochastic regression imputation could be an equally efficient imputation method because the use of full conditional normal distributions exactly reproduces the chosen simulation setup. If the interactions of group membership and covariate effect were considered in the imputation model as well, the results of DAR would match the estimates for BD and DAC. Note that in real world applications the true imputation model is never known to the researcher. However, sequential CART is capable to gauge the impact of heterogeneous data structures and reliably recovers the true parameter values.

4.3.2 Comparison with complete cases analysis

For the second scenario, I simulate three background variables X_k $(k : 1, 2, 3)$ from a multivariate normal distribution, where the variables have means $\mu_{X_1} = \mu_{X_2} = 1$ and $\mu_{X_3} = 0$, variances of $\sigma^2_{X_1} = \sigma^2_{X_2} = 4$ and $\sigma^2_{X_3} = 1$ and pairwise correlations equal to 0.5. X_3 is transformed into a binary variable with a split value of 0, i.e, $\mathbf{X} = (X_1 \; X_2 \; \mathbf{1}(X_3 > 0))$. The population model parameters are now changed to $\underline{\gamma}_1 = (-0.5, 0.2, 0.2, 0.3)'$, $\underline{\gamma}_2 = (1, 0.4, -0.2, -0.5)'$, $\sigma^2_{\epsilon,1} = 0.7^2$ and $\sigma^2_{\epsilon,2} = 0.5^2$.

Concerning the missing data mechanism, the rates of missingness for X_1, X_2 and X_3 lie around 15%, 7% and 7% and depend on the latent trait variable according to

$$P(X_{i,1} = \text{“missing”}) = \Phi(-1 - \theta_i/2) \tag{4.9}$$

and

$$P(X_{i,2} = \text{“missing”}) = P(X_{i,3} = \text{“missing”}) = \Phi(-1.7 - \theta_i). \tag{4.10}$$

Thus, scenario 2 poses more challenges on the missing data techniques than scenario 1, because it includes a categorical person covariate and it increases the number of incomplete observations.

As expected, the CC results in Tables 4.3 and 4.4 display biased estimates of the structural parameters. The bias rises to 0.27 for coefficient $\gamma_{1,0}$ and none of the resulting 200 HDRs covered the true parameter value. Consistent with the results of Section 4.3.1, my data augmented sampling algorithm revealed unbiased estimation of all parameters. Further, inspection of RMSEs and coverage rates suggest no severe loss of statistical accuracy compared to BD. The mean posterior standard deviations of the regression parameters differ only little from the associated RMSEs. These results are supported by coverages which meet the 95% confidence level for all parameters. It is evident that the combination of DA and sequential CART imputations offers a suitable solution for the treatment of missing covariates in the context of LRMs.

4.4 Examples using the German National Educational Panel Study

In order to illustrate the empirical usefulness of the suggested algorithms, I provide two exemplary applications using data from the NEPS. The NEPS is a large scale longitudinal study in Germany that aims to picture the acquisition of competencies over the entire life course (see Blossfeld et al., 2011, for the conception of the study and its embeddedness in the research landscape). The educational biographies of the respondents are therefore split into eight stages. Along with the stages, Figure 4.2 depicts the five theoretical dimensions which are central to the NEPS. To gain a deeper insight into educational attainment over time, a multicohort sequence design was implemented in the NEPS (Figure 4.3). The sampling plan covers six starting cohorts ranging from early childhood to advanced adulthood with a total number of over $60,000$ target persons who were selected between 2009 and 2012.

For the two analyses I employ data from the first two waves of the NEPS cohort sample of students in ninth grade[7] (SC4; see Skopek, Pink, & Bela, 2013, for the documentation of the SC4 scientific use files). Identical to the NEPS cohort sample of students in fifth grade, access to the children of this cohort is gained via an institutional context. The data were collected in schools in Germany between fall 2010 and winter 2010/2011. For the purpose of sampling, the population of schools is partitioned into six school type strata (Aßmann et al., 2011). Both factors, the institutional context of schools and the stratified sampling approach, give reason to assume a differentiated hierarchical data structure.

7 This thesis uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:6.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

4.4.1 Mathematical competencies at grade 9

Following Weinert et al. (2011), "four areas of individual abilities and competencies are differentiated and assessed in the NEPS: (A) domain-general cognitive abilities/capacities, (B) domain-specific cognitive competencies, (C) metacompetencies and social competencies, and (D) stage-specific (curriculum- or job related) attainments, skills, and outcome measures" (p. 71). With respect to (B), three competence domains are measured: *reading literacy and oral language comprehension*, *mathematical literacy* and *scientific literacy*. I chose the second domain as an example for latent variable modeling with person covariates. The relationship of mathematical competency with secondary school type, gender, repeating a grade and parents' socio-economic status will be analyzed. Mathematical competency was assessed in the first wave of SC4. The corresponding test comprises four content areas: *quantity*, *change and relationships*, *space and shape* and *data and chance* (Neumann et al., 2013).

From the 15,629 ninth graders participating in the first wave, students in special needs schools did not attend any competency assessment. After considering only the regular schools, I follow standard recommendations and restrict my sample to students with a valid response to at least three mathematics test items (588 cases are omitted before merging data files). Table 4.5 lists variable information, response format and frequency distribution of all $J = 22$ tasks that had to be solved in the test. 20 items have a binary format (*simple multiple-choice* and *short constructed response*). Items *item03* and *item16* are *complex multiple-choice* items consisting of three subtasks. Consequently, I treat these as ordinal items with $Q_j = 4$ categories. From the set of binary items, the easiest item is *item15* which was solved by 86% of the students. The missing rates are quite low and reach 5% for the items positioned

later in the test. The short constructed response formatted item *item17* demands from the students to write down a number into an empty field. It has the highest percentage of missing values with 21% (for an overview and further results from the mathematics test data in SC4, see Duchhardt & Gerdes, 2013). Overall, the histogram of grouped test scores in Figure 4.4 shows a normal distribution which is slightly more spread out on the right. Similar to other LSAs, the data of NEPS competence tests are scaled using IRT (Pohl & Carstensen, 2012).

In addition to the test results, I consider two clustering variables (*schooltype* and *school*) and three student variables to gauge their effect on mathematical proficiency (*female*, *repeat* and *hisei*). Table 4.6 gives a detailed overview of the background information used. In my analysis the available school type variable (Bayer, Goßmann, & Bela, 2014) was transformed to cover only the three traditional tracks of the German secondary education system: Hauptschule (*HS*; lower track), Realschule (*RS*; intermediate track) and Gymnasium (*GYM*; upper track). For observations where an assignment to these tracks was not possible or unclear, e.g., students in comprehensive schools with no separation into school branches, I declare the variable missing and exlude these observations from analysis (1,139 cases are omitted before merging files). The school identifier *school* assigns a unique number to each school and serves as the second clustering variable. *female* and *repeat* are binary variables indicating whether the student is female and the student ever repeated a school year respectively. Regarding socio-economic status, there are many operationalizations implemented in the NEPS. In line with recent analyses of the PISA data (OECD, 2013a, p. 132), I took the highest occupational level of parents measured by the index ISEI-08 (Ganzeboom, 2010) and calculated a variable *hisei* as the higher ISEI-08 score of either the students' mother or the students' father or the only available score. To change the scale of the regression coefficient associated with

hisei, the original values are divided by 100. Merging mathematics test data and all student information together results in a final data set with 13,075 observations.

Descriptive statistics for the person covariates considered in the application are displayed in Table 4.7. School type is represented by two dummy variables, where *HS* serves as reference category. With 40% of students, *GYM* is the most frequently occuring educational track. There are 511 schools in total. Half of the students are girls and one fifth of them ever repeated a grade. The only quantitative predictor *hisei* ranges from 1.16 to 8.90, with higher values indicating a higher level of occupational status. The total amount of missing data is to be considered as moderate-to-medium. At most, about 20% of the values are missing univariately for *hisei*. The ratio of students having complete background information is 79%.

To examine different specifications of the population model, I analyze the five models listed below. Each MCMC run consists of $R = 60{,}000$ iterations, where the last 50,000 are used for inference. Memory burden is reduced through storing only every 10th iteration which results in a sample of size 5,000 from the joint posterior distribution. I use the functions contained in the R package *LaRA* (Chapter 5) for performing estimation. The program took about 40 hours to execute as a shared memory parallel job (see Appendix D).

$\mathcal{M}_1$ empty LRM

$$y^*_{i,j} = \alpha_j[\gamma_0 + \epsilon_i] - \beta_j + \varepsilon_{i,j},$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ for $i : 1, \ldots, 13075$; $j : 1, \ldots, 22$.

$\mathcal{M}_2$ empty 3MGLRM by school type

$$y^*_{i,j} = \alpha_j[\gamma_{S_i,0} + \epsilon_i] - \beta_j + \varepsilon_{i,j},$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon,S_i})$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i : 1, \ldots, 13075$;

$S_i \in \{\text{“HS”}, \text{“RS”}, \text{“GYM”}\}$; $j : 1, \ldots, 22$.

$\mathcal{M}_3$ 3MGLRM by school type

$$\begin{aligned} y^*_{i,j} = \alpha_j[&\gamma_{S_i,0} + \gamma_{S_i,1} female_i + \gamma_{S_i,2} repeat_i + \gamma_{S_i,3} hisei_i + \epsilon_i] \\ &- \beta_j + \varepsilon_{i,j}, \end{aligned}$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon,S_i})$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i : 1, \ldots, 13075$;

$S_i \in \{\text{“HS”}, \text{“RS”}, \text{“GYM”}\}$; $j : 1, \ldots, 22$.

$\mathcal{M}_4$ empty school-level RILRM

$$y^*_{c,i,j} = \alpha_j[\omega_c + \gamma_0 + \epsilon_{c,i}] - \beta_j + \varepsilon_{c,i,j},$$

where $\omega_c \overset{\text{iid}}{\sim} \mathcal{N}(0, v^2_\omega)$, $\epsilon_{c,i} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$ and $\varepsilon_{c,i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $c : 1, \ldots, 511$;

$i : 1, \ldots, N_c$; $j : 1, \ldots, 22$.

$\mathcal{M}_5$ school-level RILRM

$$\begin{aligned} y^*_{c,i,j} = \alpha_j[&\omega_c + \gamma_0 + \gamma_1 female_{c,i} + \gamma_2 repeat_{c,i} + \gamma_3 schooltype{:}RS_{c,i} \\ &+ \gamma_4 schooltype{:}GYM_{c,i} + \gamma_5 hisei_{c,i} + \epsilon_{c,i}] - \beta_j + \varepsilon_{c,i,j}, \end{aligned}$$

where $\omega_c \overset{\text{iid}}{\sim} \mathcal{N}(0, v^2_\omega)$, $\epsilon_{c,i} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$ and $\varepsilon_{c,i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $c : 1, \ldots, 511$;

$i : 1, \ldots, N_c$; $j : 1, \ldots, 22$.

The trace plots and cumulative means indicate good convergence behavior of the algorithms (Figure 4.5; convergence diagnostics are only provided for the most complex model $\mathcal{M}_5$). Also, succeeding posterior samples of the single parameters show a nonsignificant autocorrelation (Figure 4.6).

Table 4.8 compares the estimates of models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. While the results that emerge from the empty LRM ($\mathcal{M}_1$) show an ability distribution for all students with a mean of 0.181, the empty MGLRM ($\mathcal{M}_2$) reveals school type-specific competency distributions with the highest mean being associated with *GYM* followed by *RS* and *HS*. There seems to be a clear performance gap between the different educational tracks. In the same way, the conditional variances $\sigma^2_{\epsilon,g}$ increase over the higher educational tracks *RS* and *GYM*. Adding the person covariates *female*, *repeat* and *hisei* to the empty MGLRM ($\mathcal{M}_3$) shows interactions with the clustering variable: first, the negative effect of gender on mathematical competency is significantly stronger for *RS* and *GYM* relative to *HS*. Second, grade repetition is much more relevant for gaining mathematical knowledge at the *GYM* than at the other two educational tracks. Likewise, third, socio-economic status is positively related to student achievement for all school types but plays a greater role at the *GYM*. Figure 4.7 graphs the distribution of posterior mean ability scores, also referred to as expected a posteriori estimates, for students by school type. We can clearly see the shifts in means and residual variances of the group-specific density curves.

There are 511 different schools in the considered sample. To test the dependency of students nested within a school, Table 4.9 summarizes the estimation results from an empty RILRM ($\mathcal{M}_4$) and a RILRM with additional background variables ($\mathcal{M}_5$). Note that the variable *schooltype* now no longer serves as a clustering variable but enters the model as a covariate. Both models confirm a significant difference between the schools. The ten smallest and largest random intercepts obtained from model $\mathcal{M}_5$ can be compared in Figure 4.8. In model $\mathcal{M}_4$, the level-two ICC is $0.223/(0.223+0.180) = 0.55$. This value gives quite a strong positive correlation between the proficiency levels of two randomly selected students from the same school.

School belonging contributes more to the variability in mathematical competency than student's interindividual differences. After controlling for person covariates in model $\mathcal{M}_5$, the between-school variance decreases to 0.054 and the student-level accounts for three quarters of the variability, $0.054/(0.054+0.159) = 0.25$. In line with the previous results, *female* and *repeat* are negatively associated with mathematical literacy and the higher educational tracks, as well as a higher socio-economic status, cause an increase in abilities.

4.4.2 Eating disorders at grade 9

Eating disorders are a public health issue. In their cost-of-illness study, Krauth, Buser, and Vogel (2002) calculate the health care cost of anorexia nervosa to be 65 million euro and 10 million euro for bulimia nervosa in Germany during 1998. Because these numbers are based on health and insurance data for in-patient care and rehabilitation only, Simon, Schmidt, and Pilling (2005) conclude that they grossly underestimate the real economic burden of eating disorders. Seen from a medical point of view, early diagnosis and intervention for these illnesses is critical and significantly improves the prognosis (Steinhausen & Seidel, 1991). Hence, Morgan, Reid, and Lacey (1999) developed a clinical screening tool called the SCOFF questionnaire. The authors designed $J = 5$ questions with either positive or negative answers. Risk persons, that might have an eating disorder, are defined through a SCOFF score of two or higher, i.e., $\sum_{j=1}^{J} y_{i,\cdot} \geq 2$. Morgan et al. (1999) found "100% sensitivity for anorexia and bulimia, separately and combined, with a specificity of 87.5% for controls." (p. 1467). Due to its briefness and good psychometric properties, the SCOFF scale is often implemented in social science surveys.

There is an ongoing debate about the dimensionality of the SCOFF scale. In a spanish adolescent sample, Muro-Sans, Amador-Campos, and Morgan (2008), using

exploratory factor analysis, found a bidimensional structure for the total sample and females. They distinguish between cognitive and behavior-related aspects of eating disorders resulting in the two factors *loss of control over food* and *purging behaviors*. For males, only a single-factor solution is reported. Beside exploratory factor analysis, Hansson, Daukantaité, and Johnsson (2015) also applied confirmatory models and found similar results in a swedish population. An alternative approach to explore different types of eating disorders was followed by McBride, McManus, Thompson, Palmer, and Brugha (2013). They identified latent groups with specific eating patterns based on the SCOFF questionnaire and the respondents' body mass index[8] (BMI). In a second step, multinomial logistic regression analysis was utilized to examine the influence of socio-demographic variables on each subgroup. Starting out with the results of McBride et al., I use different specifications of the FMLRM to combine both modeling steps. The SCOFF instrument plus students' height and weight were surveyed in the second wave of NEPS SC4.

Table 4.11 gives variable information, item wording and the frequency distribution of the SCOFF instrument. I recoded the items so that 1 indicates the possible existence of an eating disorder and 0 reflects normal eating behavior. As the SCOFF questionnaire comprises only five items, I reduced the sample of ninth graders to cases with complete test data (2,129 from originally 15,133 observations in wave 2 are deleted). After merging the screening and covariate data files, $N = 12{,}460$ students finally enter the models. Regarding the SCOFF questionnaire, students considerably more often negate the single items: the frequencies of a positive response lie between six and 32 percent. About one quarter of the ninth graders answer two or more items with yes and thus belong to the risk group (Figures 4.9 and 4.10). Whereas approximately half of the students are female, the average BMI is 20.81.

8 the BMI is calculated by weight in kilograms divided by the square of height in meters.

The missing rate is quasi zero for variable *female*, but rises to 11% for variable *bmi* (178 observations lie outside the range $[10, 50]$ and thus are considered implausible and recoded to missing values).

As item difficulty and discrimination parameters are neither interpretable nor meaningful for this example, I set $\beta_j = 0$ and $\alpha_J = 1$ and skip the corresponding parameter blocks in the DAC sampling algorithm. Hence, this model resembles the finite mixture binary panel probit regression model with concomitant variables (Aßmann & Boysen-Hogrefe, 2011). In detail, I choose the following five models for analysis. The numbers of MCMC iterations and burn-in samples are set equal to those from example 1 in Section 4.4.1. Due to the comparatively low number of items and covariates, the reported overall execution time of the *LaRA* functions was less than 24 hours.

$\mathcal{M}_6$ empty LRM

$$y^*_{i,j} = [\gamma_0 + \epsilon_i] + \varepsilon_{i,j},$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ for $i : 1, \ldots, 12460$; $j : 1, \ldots, 5$.

$\mathcal{M}_7$ LRM

$$y^*_{i,j} = [\gamma_0 + \gamma_1 female_i + \gamma_2 bmi_i + \epsilon_i] + \varepsilon_{i,j},$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ for $i : 1, \ldots, 12460$; $j : 1, \ldots, 5$.

$\mathcal{M}_8$ empty 2FMLRM

$$y^*_{i,j} = [\gamma_{S_i,0} + \epsilon_i] + \varepsilon_{i,j}, \qquad \text{with } P(S_i = g|\{\eta_g\}^2_{g=1}) = \eta_g,$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon,S_i})$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i : 1, \ldots, 12460$; $S_i \in \{1,2\}$; $g : 1,2$; $j : 1, \ldots, 5$.

$\mathcal{M}_9$ **2FMLRM**

$$y^*_{i,j} = [\gamma_{S_i,0} + \gamma_{S_i,1} female_i + \gamma_{S_i,2} bmi_i + \epsilon_i] + \varepsilon_{i,j} \qquad \text{with}$$

$$P(S_i = g | \{\eta_g\}^2_{g=1}) = \eta_g,$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon,S_i})$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i : 1, \ldots, 12460$; $S_i \in \{1,2\}$; $g : 1,2$; $j : 1, \ldots, 5$.

$\mathcal{M}_{10}$ **2FMLRM**

$$y^*_{i,j} = [\gamma_{S_i,0} + \gamma_{S_i,1} female_i + \gamma_{S_i,2} bmi_i + \epsilon_i] + \varepsilon_{i,j}, \qquad \text{with}$$

$$P(S_i = 1 | \underline{\zeta}_1, bmi_i) = \frac{\exp(\zeta_{1,0} + \zeta_{1,1} bmi_i)}{1 + \exp(\zeta_{1,0} + \zeta_{1,1} bmi_i)},$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon,S_i})$ and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i : 1, \ldots, 12460$; $S_i \in \{1,2\}$; $g : 1,2$; $j : 1, \ldots, 5$.

As can be seen from Figures 4.11 and 4.12, no convergence problems occur in the MCMC run of the most complex model $\mathcal{M}_{10}$. Starting with the simpler model specifications $\mathcal{M}_6$ and $\mathcal{M}_7$, a positve coefficient is found for *female* as well as for *bmi* (Table 4.14). Accordingly, girls and students with a higher BMI are rather affected by eating disorders. The three finite mixture models $\mathcal{M}_8$ to $\mathcal{M}_{10}$ reveal the same effects on latent eating disorder scores (Table 4.15). What is interesting about these models is that none of the multinomial logit intercepts were significant, i.e., the latent clusters occur with equal probabilities. Overall, the data do not seem

to support a discrimination between students in terms of varying cluster-specific parameters. The population models are indistinguishable from each other. In $\mathcal{M}_{10}$, concomitant variable *bmi* does not control the mixture probabilities either. These findings, while preliminary, suggest that a two latent cluster solution is not suitable for the data.

In this Chapter the data augmented MCMC aproach towards the LRM and its extensions is supplemented by an additional sampling step allowing for incomplete covariate matrices. The following Chapter introduces the alpha version of a software package which provides the estimation algorithms just discussed.

5 R package 'LaRA: Latent Regression Analysis'

> Perhaps the most important principle for the good algorithm designer is to refuse to be content. (Aho, Hopcroft, & Ullmann, 1974, p. 70)

5.1 General information

This Chapter gives an overview of the current development status of an R package for Windows which will be submitted to the Comprehensive R Archive Network (CRAN). R (R Core Team, 2016) is an open source statistical software freely available under the terms of the GNU General Public License (see, e.g., Dalgaard, 2008, for an introduction to the language). It dates back (at least) to Ihaka and Gentleman (1996) and has recently been ranked as fifth top programming language, outperformed only by the general-purpose languages C, Java, Phython and C++ (Cass, 2016).

R packages allow to distribute program code and data to other users and thereby extend the basic functionality of the R system. Hornik (2012), member of the R development core team, counts 3,425 active R extension packages and 26,152 active as well as archived source package files with a total file size of 17.97GB on the CRAN package repository. Already hosted on CRAN are several functions in existing packages performing Bayesian inference for LRMs, namely the 2PNO LRM for binary items (`MCMCirtHier1d()` from *MCMCpack*; Martin, Quinn, & Park, 2011), the confirmatory factor analysis model conditional on exogenous covariates for binary and ordinal items (`bcfa()` from *blavaan*; Merkle & Rosseel, 2016), the

2PNO IRT model including a person- and a cluster-specific random effect without covariates for binary items (`mcmc.2pno.ml()` from *sirt*; Robitzsch, 2016) and several 2PNO multilevel LRMs for binary and ordinal items (`estmlirt()` from *mlirt*; Fox, 2007, note that *mlirt* is no longer actively maintained and requires an outdated version of R). To my knowledge, there is no function that fits FMLRMs and none of the listed packages is capable of simultaneous parameter estimation and imputation of partially missing person covariates. For these reasons, *LaRA* is an important improvement to psychometric modeling on CRAN.

Aside from the two main and other auxiliary functions, *LaRA* relies on some routines from other R packages, where the latest CRAN version is in use. With respect to random numbers and densities, `rmvn()` and `dmvn()` from *mvnfast* (Fasiolo, 2016) are utilized for simulating and evaluating from multivariate normal distributions, and similarly are `rmvt()` and `dmvt()` from *mvtnorm* (Genz et al., 2016) applied in the case of multivariate t distributions. Further specific functions encompass `ucminf()` for general-purpose unconstrained nonlinear optimization and `rpart()` for an implementation of CART, both contained in packages with the same name as the functions, i.e., *ucminf* by Nielsen and Mortensen (2016) and *rpart* by Therneau et al. (2015). In order to describe the usage of the main estimation routines, the next two Sections give an overview.

5.2 The `fmlrm` function

The `fmlrm` function provides an implementation of ALGORITHM I presented in Section 3.3.1 and thus enables the estimation of the MGLRM and the FMLRM, i.e., a one-dimensional 2PNO IRT model including a multivariate regression equation of person-level predictors on the latent trait. Regression parameters are allowed to vary across either observed groups or a predefined number of mixture components.

Mixture probabilities can also be modeled in terms of concomitant variables. In case of partially observed person covariates, missing values are imputed in each sampling iteration according to the imputation step presented in Section 4.2.

5.2.1 Arguments

`fmlrm` has arguments `Y`, `BG`, `measurement`, `Ymis`, `S`, `nomix`, `BGS` and `BGTheta` defining data input and model structure. The most simple LRM without grouping results from `S` and `nomix` both set to `NULL`. Further, the user can make adjustments regarding the return value and the sequential CART imputations via arguments `eap`, `nopvs`, `mincut` and `mindev`. The function's usage is:

```
fmlrm(Y, BG = NULL, measurement = "2pno", Ymis = "ignore", S = NULL,
  nomix = NULL, BGS = NULL, BGTheta = NULL, eap = FALSE, nopv = NULL,
  mincut = 5, mindev = 1e-04, itermcmc = 12000, burnin = 2000,
  thin = 1, start.gamma = 0, start.sigma2 = 1, start.zeta = 0,
  start.alpha = 1, start.beta = 0, start.kappa = 1, gamma.mu = 0,
  gamma.prec = .01, sigma2.shape = 1, sigma2.scale = 1, zeta.mu = 0,
  zeta.prec = .01, xi.mu = 0, xi.prec = .01, tau.mu = 0,
  tau.prec = .01, ...)
```

In detail, the parameters needed are:

- `Y`, a data frame containing item responses. They can be binary or ordinal items. The responses must be coded starting at `0` or as `NA`. Rows of `Y` correspond to persons and columns correspond to items.

- `BG`, a data frame containing person covariates on the latent trait and on the mixture probabilities. They can be quantitative or factor variables and contain missing values coded as `NA`. An intercept is included. Rows of `BG` correspond to persons and columns correspond to covariates. With `BG = NULL` (default), an empty model will be estimated.

- `measurement`, a character string denoting which measurement model is estimated. `measurement = "2pno"` (default) will estimate a two-parameter normal ogive, `"1pno"` a one-parameter normal ogive and `"mixedprobit"` a mixed effects probit regression model without item parameters.

- `Ymis`, a character string how to treat `NA`s in `Y`. The default method `"ignore"` will omit them element-wise (unbalanced panel structure) and `"incorrect"` will treat them as incorrect answers.

- `S`, a vector of observed individual group membership. A multigroup model with `S` stratifying the sample will be estimated.

- `nomix`, the number of mixture components. A finite mixture model with `nomix` components will be estimated.

- `BGS`, a logical vector indicating which columns of BG serve as concomitant variables on mixture probabilities.

- `BGTheta`, a logical vector specifying which columns of BG serve as covariates on the latent trait. When `BGS` is set to `NULL`, all columns of `BG` serve as covariates on the latent trait.

- `eap`, a logical value deciding whether (`TRUE`) or not (`FALSE`) expected a posteriori scores of the latent trait and their standard deviations are returned.

- `nopv`, the number of plausible values to draw from each respondent's posterior distribution of the latent trait. If there are `NA`s in `BG`, the associated imputed data frames are returned as well. The default setting `nopvs=NULL` will provide no plausible values.

- `mincut`, the minimum number of observations in any terminal tree node during sequential CART-imputation.
- `mindev`, the complexity parameter during sequential CART-imputation. Any split that does not decrease the overall lack of fit by a factor of `mindev` is not attempted.

The length of the MCMC chains and starting values for the single parameter blocks are determined through the arguments `itermcmc`, `burnin`, `thin`, `start.gamma`, `start.sigma2`, `start.zeta`, `start.alpha`, `start.beta` and `start.kappa`, whereas initial values for the latent trait are provided through random draws from a standard normal distribution. Prior distribution parameters can be specified by the arguments `gamma.mu`, `gamma.prec`, `sigma2.shape`, `sigma2.scale`, `zeta.mu`, `zeta.prec`, `xi.mu`, `xi.prec`, `tau.mu` and `tau.prec`, which have same dimensions as the corresponding starting values. These arguments are comprehensively:

- `itermcmc`, the number of MCMC iterations.
- `burnin`, the number of burnin iterations.
- `thin`, the thinning interval. Every `thin`th iteration is retained (itermcmc $\times$ thin and burnin $\times$ thin yields total number of MCMC and burnin iterations).
- `start.gamma`, starting values for regression weights on `BG`. Either a scalar value, a vector with length equal to the number of covariates plus one for all `gamma` or a matrix with column vectors for each group.
- `start.sigma2`, starting values for residual variance of the latent trait. Either a scalar value for all `sigma2` or a vector having length equal to the number of groups.

- `start.zeta`, starting values for regression weights concerning concomitant variables. Either a scalar value, a vector with length equal to the number of concomitant variables plus one for all `zeta` or a matrix with column vectors for each group.

- `start.alpha`, starting values for item discrimination parameters. Either a scalar value for all `alpha` or a vector having length equal to the number of items.

- `start.beta`, starting values for item difficulty parameters. Either a scalar value for all `beta` or a vector with length equal to the number of items.

- `start.kappa`, starting values for item category cutoff parameters. Either a vector having length equal to the number of categories `- 2` for all `kappa` or a list of vectors for each ordinal item. If the default argument is selected, initial values are set to `c(1,...,#categories - 2)`.

- `gamma.mu`, prior mean of `gamma`.

- `gamma.prec`, prior precision of `gamma`.

- `sigma2.shape`, prior shape parameter of `sigma2`.

- `sigma2.scale`, prior scale parameter of `sigma2`.

- `zeta.mu`, prior mean of `zeta`.

- `zeta.prec`, prior precision of `zeta`.

- `xi.mu`, prior mean of `xi=(alpha beta)`.

- `xi.prec`, prior precision of `xi=(alpha beta)`.

- `tau.mu`, prior mean of `tau`.
- `tau.prec`, prior precision of `tau`.

5.2.2 Value

The return value of the `fmlrm` function is an object of class '`LaRA`', a list with elements `mcmcdraws`, `acc.tau`, `acc.zeta`, `eapscores`, `pvs` and `pvsBGimp` containing the produced parameter samples, additional information on the sampler, latent trait scores and imputed background variables. Contained in these list elements are more specifically:

- `mcmcdraws`, a list containing matrices of posterior samples. Columns correspond to parameter blocks and are sorted in the following order: `gamma`, `sigma2`, `zeta`, `alpha`, `beta` and `kappa`. Rows correspond to MCMC iterations.
- `acc.tau`, if estimated, a vector of M-H acceptance rates of category cutoff parameters for ordinal items.
- `acc.zeta`, if estimated, a vector of M-H acceptance rates of regression weights for concomitant variables.
- `eapscores`, if requested, a list containing vectors of respondent's expected a priori scores of the latent trait and their standard deviations.
- `pvs`, if requested, a list of length `nopv` containing vectors of plausible values.
- `pvsBGimp`, if requested, a list of length `nopv` containing imputed versions of data frame `BG` which were used to generate `pvs`.

5.2.3 Examples

The following R code runs the empirical applications using NEPS data from SC4 in Sections 4.4.1 and 4.4.2. After specifying the MCMC run length, the user needs to split the data frames `nepssc4math` and `nepssc4scoff` into objects containing the respective item responses and person covariates:

```
R> library(LaRA)
R> ## Define MCMC stuff
R> ## -----------------
R> itermcmc <- 6000
R> burnin <- 1000
R> thin <- 10
R> ## Prepare data input
R> ## ------------------
R> # EXAMPLE 1: mathematical competencies
R> str(nepssc4math)
'data.frame': 13075 obs. of  28 variables:
$ item01    : int  0 1 1 1 1 1 0 0 0 1 ...
$ item02    : int  0 0 0 1 0 0 1 0 0 1 ...
$ item03    : int  NA 1 2 2 2 1 3 0 2 3 ...
$ item04    : int  0 0 1 0 0 0 0 0 0 0 ...
$ item05    : int  0 1 1 0 1 1 1 0 1 1 ...
$ item06    : int  0 0 1 1 1 0 0 0 1 1 ...
$ item07    : int  0 1 1 1 1 1 1 0 0 1 ...
$ item08    : int  0 1 0 0 0 0 0 0 0 1 ...
$ item09    : int  1 1 1 1 1 1 1 0 0 1 ...
$ item11    : int  1 0 0 1 0 1 1 0 0 0 ...
$ item12    : int  0 1 0 NA 1 0 0 1 0 1 ...
$ item13    : int  0 0 1 1 1 0 0 0 0 1 ...
$ item14    : int  1 0 1 1 1 0 1 1 0 1 ...
$ item15    : int  0 1 0 0 1 0 0 0 0 1 ...
$ item16    : int  1 1 1 0 1 1 1 1 0 1 ...
$ item17    : int  1 0 3 1 3 2 3 1 2 3 ...
$ item18    : int  0 0 0 NA 1 0 0 0 NA 1 ...
$ item19    : int  0 1 1 0 1 0 0 1 0 0 ...
$ item20    : int  1 0 1 1 1 1 1 0 0 1 ...
$ item21    : int  1 0 1 0 1 0 1 1 0 1 ...
$ item22    : int  0 0 0 1 1 0 1 0 0 0 ...
$ item23    : int  0 0 1 0 1 0 0 1 0 1 ...
```

```
$ schooltype: Factor w/ 3 levels "HS","RS","GYM": 3 2 3 3 2 2 2 1 ...
$ school    : int  410 84 438 11 316 249 441 429 245 215 ...
$ female    : Factor w/ 2 levels "male","female": 2 1 2 1 2 2 1 2 ...
$ repeat    : Factor w/ 2 levels "no","yes": 1 2 2 1 1 1 1 1 1 1 ...
$ hisei     : num  5.45 5.74 4.05 7.05 5.99 ...
R> YMC <- nepssc4math[, grep("item", names(nepssc4math), value = T)]
R> BGMCMG <- nepssc4math[, c("female", "repeat", "hisei")]
R> SMG <- as.numeric(nepssc4math[, "schooltype"])
R> # EXAMPLE 2: eating disorders
R> str(nepssc4scoff)
'data.frame': 12460 obs. of  7 variables:
$ scoff01: num  0 0 0 0 0 1 0 1 0 0 ...
$ scoff02: num  0 0 0 0 0 0 0 1 0 0 ...
$ scoff03: num  0 0 0 0 0 1 0 0 0 0 ...
$ scoff04: num  0 0 0 0 0 0 0 1 0 0 ...
$ scoff05: num  0 1 0 1 1 1 1 1 0 1 ...
$ female : Factor w/ 2 levels "male","female": 2 1 2 1 2 2 1 2 1 2 ...
$ bmi    : num  17.8 20.5 20.8 18.5 18.1 ...
R> YED <- nepssc4scoff[, grep("scoff", names(nepssc4scoff),
+    value = T)]
R> BGED <- nepssc4scoff[, c("female", "bmi")]
```

Once the data objects have been created, it is straightforward to fit the various LRMs calling the `fmlrm` function:

```
R> ## Start estimation runs
R> ## ---------------------
R> # mathematical competencies: empty LRM
R> MC1 <- fmlrm(Y = YMC, itermcmc = itermcmc, burnin = burnin,
+    thin = thin)
R> # mathematical competencies: empty MGLRM
R> MC2 <- fmlrm(Y = YMC, S = SMG, itermcmc = itermcmc,
+    burnin = burnin, thin = thin)
R> # mathematical competencies: MGLRM
R> MC3 <- fmlrm(Y = YMC, BG = BGMCMG, S = SMG,
+    itermcmc = itermcmc, burnin = burnin, thin = thin)
R> # eating disorders: empty LRM
R> ED1 <- fmlrm(Y = YED, measurement = "mixedprobit",
+    itermcmc = itermcmc, burnin = burnin, thin = thin)
R> # eating disorders: LRM
R> ED2 <- fmlrm(Y = YED, BG = BGED, measurement = "mixedprobit",
```

```
+    itermcmc = itermcmc, burnin = burnin, thin = thin)
R> # eating disorders: empty 2FMLRM
R> ED3 <- fmlrm(Y = YED, measurement = "mixedprobit", nomix = 2,
+    itermcmc = itermcmc, burnin = burnin, thin = thin)
R> # eating disorders: 2FMLRM
R> ED4 <- fmlrm(Y = YED, BG = BGED, measurement = "mixedprobit",
+    nomix = 2, itermcmc = itermcmc, burnin = burnin, thin = thin)
R> # eating disorders: 2FMLRM with concomitant variables
R> ED5 <- fmlrm(Y = YED, BG = BGED, measurement = "mixedprobit",
+    nomix = 2, BGSi = c(F, T), BGThetai = c(T, T),
+    itermcmc = itermcmc, burnin = burnin, thin = thin)
```

To finally obtain Bayesian point and interval estimates, the list element `mcmcdraws` of the returned 'LaRA' object can be summarized using the `apply` family of functions over array margins, e.g., for calculating mean posterior draws and respective quantiles of the regression weights:

```
R> # mathematical competencies: empty MGLRM
R> MC2postmean.gamma <- apply(MC2$mcmcdraws[-(1:burnin), 1:3], 2,
+    mean)
R> MC2posthdr.gamma <- apply(MC2$mcmcdraws[-(1:burnin), 1:3], 2,
+    quantile, probs = c(0.025, 0.975))
```

5.3 The `rilrm` function

The `rilrm` function provides an implementation of ALGORITHM II presented in Section 3.3.2 and thus allows to estimate the RILRM, i.e., a one-dimensional 2PNO IRT model including a multivariate regression equation of a cluster-level random effect and person-level predictors on the latent trait. In case of partially observed person covariates, missing values are imputed in each sampling iteration. For this purpose the imputation step introduced in Section 4.2 is utilized.

5.3.1 Arguments

`rilrm` can be called in the following manner:

```
rilrm(Y, BG = NULL, measurement = "2pno", Ymis = "ignore", S = NULL,
  eap = FALSE, ri = FALSE, nopv = NULL, mincut = 5, mindev = 1e-04,
  itermcmc = 12000, burnin = 2000, thin = 1, start.gamma = 0,
  start.sigma2 = 1, start.upsilon2 = 1, start.alpha = 1,
  start.beta = 0, start.kappa = 1, gamma.mu = 0, gamma.prec = .01,
  sigma2.shape = 1, sigma2.scale = 1, upsilon2.shape = 1,
  upsilon2.scale = 1, xi.mu = 0, xi.prec = .01, tau.mu = 0,
  tau.prec = .01, ...)
```

Note that except for `ri`, the arguments and inputs defined by the user can be identically specified to `fmlrm()`. New parameters not part of `fmlrm()` are:

- `ri`, a logical value deciding whether (`TRUE`) or not (`FALSE`) posterior means and standard deviations of the random intercepts are returned.
- `start.upsilon2`, starting value for variance of the cluster-specific random effect.
- `upsilon2.shape`, prior shape parameter of `upsilon2`.
- `upsilon2.scale`, prior scale parameter of `upsilon2`.

5.3.2 Value

In line with the function arguments and additional to the output created by `fmlrm()`, the return value of `rilrm` holds a further list `clustereffects`, that, if requested, contains vectors of cluster-specific random intercepts and their standard deviations.

5.3.3 Examples

The remaining steps of the empirical application in Section 4.4.1 are run with the following code:

```
R> ## Prepare data input
R> ## ------------------
R> SRI <- nepssc4math[, "school"]
```

```
R> BGMCRI <- nepssc4math[, c("female", "repeat", "schooltype",
+    "hisei")]
R> ## Start estimation runs
R> ## --------------------
R> # mathematical competencies: empty RILRM
R> MC4 <- rilrm(Y = YMC, S = SRI, itermcmc = itermcmc,
+    burnin = burnin, thin = thin)
R> # mathematical competencies: RILRM
R> MC5 <- rilrm(Y = YMC, BG = BGMCRI, S = SRI, itermcmc = itermcmc,
+    burnin = burnin, thin = thin)
```

The above functions offer a complete open source implementation of nonlinear mixed models with latent variables embedded into a powerful environment for statistical computing and graphics. Nevertheless, much research remains to be done. Some of the questions following from my thesis will be discussed in the next Chapter.

6 Directions for future research

In the near future the R package *LaRA* and its underlying statistical models will be completed to incorporate additional multilevel data structures. The most obvious extensions are introducing further levels, random slopes and cluster-level covariates into the RILRM. In this way, for example school and class effects on latent student abilities could be studied simultaneously and person-level regression coefficients may vary with the value of higher-order covariates like class size. Even more complex models allow for cross-classifications of respondents and membership to more than one cluster at the same time (Browne, Goldstein, & Rasbash, 2001). While Asparouhov and Muthén (2007) show that a combination of the multilevel and finite mixture approach is feasible and highlight the advantages of such a composite model, the authors also note that "the more general and flexible a statistical model is, the bigger the effort on the part of the researcher to interpret the model and the results in a practically meaningful way." (p. 28). I am convinced, too, that the key priority for data analysis is to be only as complex as necessary.

Further work on *LaRA* will concentrate on addressing multidimensional problems, i.e., more than one latent trait variable is involved in modeling. LSAs usually consist of several domains which are expected to be related. For instance, the first wave of NEPS SC4 surveys computer and scientific literacy in addition to mathematical competency. Moreover, as the NEPS explicitely aims at pursuing the different starting cohorts over time, single domains will be retested with advancing age of the study (see again the multicohort sequence design of the NEPS in figure 4.3).

As a result, competency measurements need to be combined over different points in time. Both fields of application illustrate the necessity of multidimensional models which simultaneously estimate multiple traits and their mutual interdependencies. Multidimensional IRT models were developed by Béguin and Glas (2001) and Edwards (2010) amongst others, whereas Azevedo, Fox, and Andrade (2016) recently proposed a longitudinal IRT model to analyze individual differences in educational progress. Note that none of these methodical approaches does yet include any type of background variables or hierarchical structures.

One possibility to approximate a multidimensional LRM which would be comparatively easy to implement, is given through the specification of multiple univariate LRMs that include all other traits as covariates. If, for example, both mathematical and scientific competencies are assessed, this gives two regression equations

$$\begin{aligned} \underline{\theta}_{\text{math}} &= (\mathbf{X}\ \underline{\theta}_{\text{science}})\underline{\gamma}_{\text{math}} + \underline{\epsilon}_{\text{math}} \\ \underline{\theta}_{\text{science}} &= (\mathbf{X}\ \underline{\theta}_{\text{math}})\underline{\gamma}_{\text{science}} + \underline{\epsilon}_{\text{science}} \end{aligned} \tag{6.1}$$

with $\underline{\epsilon}_{\text{math}} \sim \mathcal{N}(0, \sigma^2_{\epsilon,\text{math}}\mathcal{I}_N)$ and $\underline{\epsilon}_{\text{science}} \sim \mathcal{N}(0, \sigma^2_{\epsilon,\text{science}}\mathcal{I}_N)$, where the lower indices denote the domain belonging to the latent ability, regression weights and disturbance vectors. After initialization, a chained Gibbs sampler can be run that subsequently samples from each set of domain specific full conditional distributions. In doing so, the dependence structure between mathematical and scientific literacy is considered through the inclusion of each ability dimension in the respective other population model.

Another urgent question regards the statistical choice between nonnested model specifications. This becomes especially relevant in the case of FMLRMs when the number of mixture components has to be decided. Bayesian estimation using MCMC technology offers a conceptually straightforward way to deal with the comparison of

overall model fit (or more generally hypothesis testing) via Bayes factors (see Kass & Raftery, 1995, for a thorough discussion). The Bayes factor for checking model $\mathcal{M}_1$ against model $\mathcal{M}_2$ based on sample data $\mathbf{D}$ is defined as

$$BF_{12} = \frac{f(\mathbf{D}|\mathcal{M}_1)}{f(\mathbf{D}|\mathcal{M}_2)} = \frac{f(\mathcal{M}_1|\mathbf{D})}{f(\mathcal{M}_2|\mathbf{D})} \; / \; \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}. \tag{6.2}$$

This ratio expresses the evidence in favour of model $\mathcal{M}_1$ compared to model $\mathcal{M}_2$. Thus, the model with the largest marginal likelihood value is chosen among a set of competing specifications. When two models have the same prior probability, i.e., $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2)$, the Bayes factor reduces to their posterior odds $f(\mathcal{M}_1|\mathbf{D})/f(\mathcal{M}_2|\mathbf{D})$. According to Jeffreys (1961), a Bayes factor greater than 3.2 may be interpreted as a substantial evidence against $\mathcal{M}_2$.

As the integrals involved in (6.2) cannot be evaluated analytically in the context of the LRM and its extensions, I will seek to employ the methods of Chib (1995) and Chib and Jeliazkov (2001) for computing the marginal likelihood. After rearranging equation (3.1) and transforming it to the log scale, the right-hand side of this *basic identity* can be evaluated at any value ψ^* (usually a point of high posterior density) to estimate the marginal likelihood, i.e.,

$$\ln\widehat{f}(\mathbf{D}|\mathcal{M}_l) = \ln f(\mathbf{D}|\mathcal{M}_l, \psi_l^*) + \ln\pi(\psi_l^*|\mathcal{M}_l) - \ln\widehat{f}(\psi_l^*|\mathbf{D}, \mathcal{M}_l). \tag{6.3}$$

Whereas prior and likelihood evaluations for the models considered in this thesis are available directly and using simulation techniques respectively (see, e.g., Liesenfeld and Richard (2008) for a review of efficient importance sampling), estimation of the posterior ordinate is nontrivial. Finally, relying on the law of total probability, this quantity is estimated by

$$\ln\widehat{f}(\psi_l^*|\mathbf{D},\mathcal{M}_l) = \sum_{b=1}^{B} \ln\widehat{f}(\psi_b^*|\mathbf{D},\psi_\iota^*(\iota<b)), \tag{6.4}$$

where $\widehat{f}(\psi_b^*|\mathbf{D},\psi_\iota^*(\iota<b))$ are estimates of each posterior component resulting from reduced MCMC simulation runs.

Note that the calculations hold for completely observed covariate data. If item nonresponse is present and hence the DA algorithm includes draws from the corresponding full conditional distributions of missing values, these distributions need to be considered during evaluations too. Adopting the notation introduced in Section 4.1, equation (6.3) may be rewritten as

$$\begin{aligned}\ln\widehat{f}(\mathbf{Y}) = {} & \ln f(\mathbf{Y}|\psi^*,\mathbf{X}_{\text{mis}}^*,\mathbf{X}_{\text{obs}}) + \ln\pi(\psi^*,\mathbf{X}_{\text{mis}}^*|\mathbf{X}_{\text{obs}}) \\ & - \ln\widehat{f}(\psi^*,\mathbf{X}_{\text{mis}}^*|\mathbf{Y},\mathbf{X}_{\text{obs}}).\end{aligned} \tag{6.5}$$

The problem remains that sequential CART only approximates the correct distribution of missing values. However, if their full conditional distributions can be specified, Chib's method applies. This could be accomplished in the following way: For each variable involving missing values, the corresponding entries in $\mathbf{X}_{\text{mis}}^*$ are fixed at the average values across all imputations (like the mean, median or mode). The likelihood function is then calculated based on these values. Regarding the prior densities of missing values in continuous covariates, kernel density estimates are computed from $\mathbf{X}_{\text{obs}}$ and afterwards evaluated at $\mathbf{X}_{\text{mis}}^*$. In the case of incomplete categorical covariates, a multinomial distribution with probabilities equal to the observed frequencies in the final nodes of a CART run serves as a prior density. Lastly, mean densities for the just defined distributions across all MCMC iterations are used for posterior evaluation.

Despite tolerable running speed of the main functions from *LaRA*, C++ pro-

gram code will be integrated for accelerating computations. To achieve this, the packages *Rcpp* (Eddelbuettel, 2013; Eddelbuettel & François, 2011) and *RcppArmadillo* (Eddelbuettel & Sanderson, 2014) will be used. They greatly facilitate the interchange of R objects between R and C++ and connect R with the Armadillo C++ linear algebra library.

7 Conclusions

In this thesis I examine the class of LRMs which are versatile research tools for social sciences applications. They allow the researcher for both, scaling a latent construct and determining its relationship with additional covariates. In applying this type of modeling, measurement error has to be addressed in three different ways: Uncertainty arises from (1) the particular indicators which measure the latent trait, (2) the latent dependencies with respect to the predictors used and (3) the imputation of partially missing covariate data. Further, to meet the challenges of real-world settings, hierarchical data structures need to be considered in the aforementioned issues (1), (2) and (3). The key outcome of my thesis is that the data augmented MCMC procedures suggested therein proved to be capable of handling all these requirements.

Uncertainty due to (1) and (2) is fully reflected over the iterations of a single MCMC run. By providing simulated draws from the joint posterior distribution, a whole range of valid parameter values becomes available after estimation. Regarding (3), the current standard involves a multistage procedure: After applying dummy-variable adjustment (which does not take into account statistical uncertainty at all) or multiple imputation for missing values in background variables, model parameters are estimated separately. In contrast, I propose to sample missing values in person characteristics along with the underlying continuous outcomes, the model parameters and the latent trait. The DA device enables to unify the estimation of all these quantities in a statistically efficient one-step procedure. The uncertainty

stemming from partially missing covariate data is directly incorporated into parameter estimation. At every iteration of the algorithm an imputed version of the covariate data is used to sample from the set of full conditional posterior distributions. Vice versa, the iteratively updated parameter values resulting from posterior sampling can be proximately put into the imputation model. Thus, compared to existing methods the novel method carries out parameter estimation and imputation of missing background variables simultaneously. Taken together, there are several advantages resulting from such an approach:

- It is statistically efficient in the sense that values for the latent trait, item characteristics, and nonresponse imputations are all provided at once,
- all possible sources of uncertainty are taken into account, and
- imputation of latent variables may be conditioned on updated draws from any full conditional distribution inserted into the sampler.

I choose the underlying variable formulation of the outcomes to find a solution for mixed-format tests that include binary and ordinal items. This model derivation further facilitates Bayesian estimation via MCMC techniques, also with regard to a flexible handling of hierarchical data structures. To extend the basic LRM to clustered observations, only some minor modifications of the algorithm are needed. However, considering these structures is an important aspect in analyzing large-scale assessment data. For instance, educational institutions may represent the sampling frame of a study and latent competency scores are likely correlated within schools or universities. School systems with a high degree of horizontal stratification serve as another good example for heterogeneity across respondents. In order to gauge such effects, I developed two Metropolis-within-Gibbs algorithms performing estimation

of the standard LRM as well as a finite mixture and a random intercept specification. The finite mixture LRM assumes a composite population consisting of a few strata in which separate LRMs hold. This approach offers the flexibility to also control for latent population heterogeneity, i.e., the respondents belonging to a stratum is not observed and thus not available prior to analysis. If, on the other hand, individual cluster membership is observed, this results in the multiple group LRM. An alternative modeling strategy is presented in terms of the random intercept LRM which adds another random coefficient to the regression equation due to a large number of second level units.

With regard to missing covariate data in the context of LRMs, sequential CART prove advantageous for treating the nonresponse. Especially the fact that there is no need for specifying an imputation model makes it a robust and valuable imputation tool and opens the door towards large population models. In two simulation studies DAC was capable of adequately recovering the structural parameters of two MGLRMs in the presence of missing background information. The results are valid for the covariates being MCAR and MAR and even when rates of missingness are quite high. Whereas stochastic regression imputation and CC analysis failed to capture the true parameters, my approach showed good accuracy, especially in detecting the nonlinear effects involved in the structural relationships. These findings indicate that DAC generally approximates well the full sample estimates. Its benefits arise in terms of methodological stringency and gains in statistical efficiency. Two empirical examples using the German NEPS reflecting two different research questions further demonstrate the broad applicability of the approach to a wide range of social science topics. In these examples, only a low number of person covariates was selected. An enlarged set of differently scaled background variables, presumably increasing the amount of missing data, would illustrate the usefulness

of DAC even better.

Obviously, the provision of PVs is a convenient application for the methods and the newly developed software package *LaRA* introduced in this thesis. Context questionnaires adapted for LSAs may easily comprise hundreds of background variables. The usability could be twofold. On the one hand, PVs contained in scientific- and public-use-files can be generated as usual by the research institute but now rely on a proper imputation procedure with all its benefits. On the other hand, users can estimate PVs themselves specific to their research question. For that purpose, special trainings could be offered to assist the users in working with *LaRA*. This, in turn, would give the package maintainers valuable insights into user needs and requirements to improve software quality.

Moreover, imputed versions of the partially observed covariate matrices result from each simulation iteration. Nevertheless, PVs and nonresponse imputations have to be determined at the same iteration if one would like to analyze them jointly. Besides permitting the estimation of ability scores and their correlations with the context variables purified from measurement error, any number of completed data sets may also serve as multiple imputations of the missing background information. By incorporating the clustering of observations into the imputation model in terms of the latent trait, individual cluster membership or random intercepts, the DAC sampler provides an answer for handling incomplete multilevel data. In conclusion, my research advocates a unified approach to the simultaneous treatment of diverse types of latent variables by means of MCMC simulation. Maybe it will contribute to rethinking standards in the estimation of latent trait distributions involving covariates.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47-76.

Aho, A. V., Hopcroft, J. E., & Ullmann, J. D. (1974). *The design and analysis of computer algorithms*. Reading, MA: Addison-Wesley.

Aitkin, M., & Aitkin, I. (2011). *Statistical modeling of the national assessment of educational progress*. New York, NY: Springer.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics*, *17*(3), 251-269.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669-679.

Albert, J. H., & Chib, S. (1997). *Bayesian methods for cumulative, sequential and two-step ordinal data regression models*. Bowling Greene, OH: Department of Mathematics and Statistics, Bowling Greene State University.

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The naep 1996 technical report (nces-1999-452)*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40-64.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*(4), 475-496.

Asparouhov, T., & Muthén, B. O. (2007). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models.* Charlotte, NC: Information Age Publishing.

Aßmann, C., & Boysen-Hogrefe, J. (2011). A bayesian approach to model-based clustering for binary panel probit models. *Computational Statistics and Data Analysis*, *55*(1), 261-279.

Aßmann, C., Gaasch, J.-C., Pohl, S., & Carstensen, C. H. (2016). Estimation of plausible values considering partially missing background information: A data augmented mcmc approach. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological issues of longitudinal surveys: The example of the national educational panel study* (p. 503-521). Wiesbaden: Springer VS.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). Sampling designs of the national educational panel study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Education as a lifelong process. The german na-*

tional educational panel study (neps) (p. 51-65). Wiesbaden: VS Verlag für Sozialwissenschaften.

Aßmann, C., Würbach, A., Goßmann, S., Geisser, F., & Bela, A. (2015). Nonparametric multiple imputation for questionnaires with individual skip patterns and constraints: The case of income imputation in the national educational panel study. *Sociological Methods & Reserach*, *published online before print, doi:10.1177/0049124115610346*, 1-34.

Azevedo, C. L. N., Andrade, D. F., & Fox, J.-P. (2012). A bayesian generalized multiple group irt model with model-fit assessment tools. *Computational Statistics and Data Analysis*, *56*(12), 4399-4412.

Azevedo, C. L. N., Fox, J.-P., & Andrade, D. F. (2016). Bayesian longitudinal item response modeling with restricted covariance pattern structures. *Statistics and Computing*, *26*(1), 443-460.

Bayer, M., Goßmann, F., & Bela, D. (2014). *Neps technical report: Generated school type variable t723080_g1 in starting cohorts 3 and 4 (neps working paper no. 46).* (University of Bamberg, Leibniz Institute for Educational Trajectories, National Educational Panel Study)

Béguin, A. A., & Glas, C. A. W. (2001). Mcmc estimation and some model-fit analysis of multidimensional irt models. *Psychometrika*, *66*(4), 541-562.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis* (2nd ed.). New York, NY: Springer Science + Business Media.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Addison-Wesley.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process. the german national educational panel study (neps).* Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in timss: A case study. *Large-scale Assessments in Education*, *1*(4), 1-33.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199-231.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Boca Raton: Chapman & Hall/CRC.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (mmmc) models. *Statistical Modelling*, *1*(2), 103-124.

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070-1076.

Burstein, L. (1980). The analysis of multilevel data in educational research and

evaluation. *Review of Research in Education*, *8*, 158-233.

Cai, L. (2012). Latent variable modeling. *Shanghai Archives of Psychiatry*, *24*(2), 118-120.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York, NY: Cambridge University Press.

Cass, S. (2016, July 26). *The 2016 top programming languages.* Retrieved from `http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages`

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

Chib, S. (2001). Markov chain monte carlo methods: Computation and inference. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, p. 3569-3649). Amsterdam: Elsevier.

Chib, S. (2008). Panel data modeling and inference: A bayesian primer. In L. Mátyás & P. Sevestre (Eds.), *The econometrics of panel data: Fundamentals and recent developments in theory and practice* (Vol. 46). Berlin: Springer.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, *96*(453), 270-281.

Clinton, J., Jackman, S., & Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review*, *98*(2), 355-370.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (1975/2002). *Applied multiple regression/correlation analysis of the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. (1st ed. published in 1975)

Cowles, M. K. (1996). Accelerating monte carlo markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, *6*, 101-111.

Dalgaard, P. (2008). *Introductory statistics with r* (2nd ed.). New York, NY: Springer Science + Business Media.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173-178.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1-38.

Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis*, *72*, 92-104.

Drechsler, J. (2015). Multiple imputation of multilevel missing data—rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, *40*(1), 69-95.

Duchhardt, C., & Gerdes, A. (2013). *Neps technical report for mathematics - scaling results of starting cohort 4 in ninth grade (neps working paper no. 22).* (University of Bamberg, Leibniz Institute for Educational Trajectories, National Educational Panel Study)

Eddelbuettel, D. (2013). *Seamless r and c++ integration.* New York, NY: Springer.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, *40*(8), 1-18.

Eddelbuettel, D., & Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++. *Computational Statistics and Data Analysis*, *71*, 1054-1063.

Edwards, M. C. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474-497.

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

Fasiolo, M. (2016). An introduction to mvnfast. r package version 0.1.5. [Computer software manual]. Retrieved from `http://cran.r-project.org/web/packages/mvnfast/vignettes/mvnfast.html`

Fox, J.-P. (2005). Multilevel irt using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 145-172.

Fox, J.-P. (2007). Multilevel irt modeling in practice with the package mlirt. *Journal of Statistical Software*, *20*(5), 1-16.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer Science + Business Media.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika*, *66*(2), 271-288.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611-631.

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*. New York, NY: Springer Science + Business Media.

Frühwirth-Schnatter, S. (2011). Panel data analysis: A survey on model-based clustering of time series. *Advances in Data Analysis and Classification*, *5*, 251-280.

Frühwirth-Schnatter, S., & Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, *26*(1), 78-89.

Ganzeboom, H. B. G. (2010). *A new international socio-economic index [isei] of occupational staus for the international standard classification of occupation 2008 [isco-08] constructed with data from the issp 2002-2007; with an analysis of quality of occupational measurement in issp.* (Annual Conference of International Social Survey Programme, Lisbon, May 1 2010)

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the

bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721-741.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2016). mvtnorm: Multivariate normal and t distributions [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=mvtnorm` (R package version 1.0-5)

Geweke, J., & Keane, M. (2001). Computationally intensive methods for integration in econometrics. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, p. 3463-3568). Amsterdam: Elsevier.

Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal*, *7*(1), 98-119.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38 - 47.

Hansson, E., Daukantaité, D., & Johnsson, P. (2015). Scoff in a general swedish adolescent population. *Journal of Eating Disorders*, *3*(48), 1-9.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer Science + Business Media.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Hornik, K. (2012). Are there too many r packages? *Austrian Journal of Statistics*, *41*(1), 59-66.

Hox, J. (2010). *Multilevel analysis. techniques and applications* (2nd ed.). New York, NY: Routledge.

Huang, G.-H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, *69*(1), 5-32.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299-314.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Johnson, M. S., & Jenkins, F. (2005). *A bayesian hierarchical model for large-scale educational surveys: An application to the national assessment of educational progress (ets rr-04-38)*. Princeton, NJ: Educational Testing Service.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*(433), 222-230.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York, NY: Guilford Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity

into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*(5), 850-874.

Koskinen, J. H., Robins, G. L., & Pattison, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. *Statistical Methodology*, *7*(3), 366-384.

Krauth, C., Buser, K., & Vogel, H. (2002). How high are the costs of eating disorders - anorexia nervosa and bulimia nervosa - for german society? *The European Journal of Health Economics*, *3*(4), 244-250.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, *95*(2), 391-413.

Lenk, P. J., & DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, *65*(1), 93-119.

Liesenfeld, R., & Richard, J.-F. (2008). Simulation techniques for panels: Efficient importance sampling. In L. Mátyás & P. Sevestre (Eds.), *The econometrics of panel data. fundamentals and recent developments in theory and practice* (3rd ed., p. 419-450). Springer.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: J. Wiley & Sons.

Liu, M., Taylor, J. M. G., & Belin, T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*, *56*(4), 1157-1153.

Lord, F. M. (1952). *A theory of test scores (psychometric monograph no. 7)*. Richmond, VA: Psychometric Corporation.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, *18*(1), 57-75.

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

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

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21-39.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). Mcmcpack: Markov chain monte carlo in r. *Journal of Statistical Software*, *42*(9), 1-21.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McBride, O., McManus, S., Thompson, J., Palmer, R. L., & Brugha, T. (2013). Profiling disordered eating patterns and body mass index (bmi) in the english general population. *Social Psychiatry and Psychiatric Epidemiology*, *48*(5), 783-793.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal*

Statistical Society: Series B, *42*(2), 109-142.

McKelvey, R., & Zavoina, W. (1975). A statistical model for the analysis of ordered level dependent variables. *Journal of Mathematical Sociology*, *4*(1), 103-120.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538-558.

Merkle, E., & Rosseel, Y. (2016). blavaan: Bayesian latent variable analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=blavaan` (R package version 0.1-4)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, *21*(6), 1087-1092.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*(392), 993-997.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*(1), 81-91.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133-161.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (p. 57-74). San Diego, CA: Academic Press.

Mislevy, R. J., Eugene, G. J., & Muraki, E. (1992). Scaling procedures in naep. *Journal of Educational Statistics*, *17*(2), 131-154.

Morgan, J. F., Reid, F., & Lacey, H. (1999). The scoff questionnaire: Assessment of a new screening tool for eating disorders. *British Medical Journal*, *319*(7223), 1467-1468.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *Timss 2015 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muro-Sans, P., Amador-Campos, J. A., & Morgan, J. F. (2008). The scoff-c: Psychometric properties of the catalan version in a spanish adolescent sample. *Journal of Psychosomatic Research*, *64*, 81-86.

Muthén, B. O. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association*, *74*(368), 807-811.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending irt to external variables. In W. Hainer & H. Braun (Eds.), *Test*

validity (p. 213-238). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*(4), 407-419.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online*, *5*(2), 80-109.

Nielsen, H. B., & Mortensen, S. B. (2016). ucminf: General-purpose unconstrained non-linear optimization [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ucminf` (R package version 1.1-4)

OECD. (2013a). *Pisa 2012 results: Excellence through equity: Giving every student the chance to suceed (volume ii)*. Paris: OECD Publishing.

OECD (Ed.). (2013b). *Technical report of the survey of adult skills (piaac)*. Paris: OECD Publishing.

OECD (Ed.). (2014). *Pisa 2012 technical report*. Paris: OECD Publishing.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Peskun, P. H. (1973). Optimum monte-carlo sampling using markov chains. *Biometrika*, *60*(3), 607-612.

Pohl, S., & Carstensen, C. H. (2012). *Neps technical report: Scaling the data of the competence tests (neps working paper no. 14)*. (University of Bamberg, Leibniz Institute for Educational Trajectories, National Educational Panel Study)

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*(3), 423-452.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org`

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*, 85-96.

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

Rässler, S. (2002). *Statistical matching - a frequentist theory, practical applications, and alternative bayesian approaches*. New York, NY: Springer Science

+ Business Media.

Rässler, S., Rubin, D. B., & Zell, E. R. (2008). Incomplete data in epidemiology and medical statistics. In C. R. Rao, J. P. Miller, & D. C. Rao (Eds.), *Handbook of statistics (vol. 27): Epidemiology and medical statistics*. Amsterdam: Elsevier.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, *13*(2), 85-116.

Reiter, J. P. (2005). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441-462.

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptions of multiple imputation. *Journal of the American Statistical Association*, *102*(480), 1462-1471.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185-205.

Robitzsch, A. (2016). sirt: Supplementary item response theory models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=sirt` (R package version 1.12-2)

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rubin, D. B. (1978). Multiple imputation in sample surveys - a pehnomenological bayesian approach to nonresponse. In *Proceedings of the american statistical association* (p. 20-40).

Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, *9*(1), 130-134.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement*, *48*(3), 293-312.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (psychometric monograph no. 17)*. Richmond, VA: Psychometric Corporation.

Schenker, N., & Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, *22*(4), 425-446.

Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale asessment surveys. *Journal of Educational and Behavioral Statistics*, *38*(5), 499-521.

Simon, J., Schmidt, U., & Pilling, S. (2005). The health service use and cost of eating disorders. *Psychological Medicine*, *35*(11), 1543-1551.

Skopek, J., Pink, S., & Bela, D. (2013). *Starting cohort 4: Grade 9 (sc4). suf version 1.1.0. data manual (neps research data paper)*. (University of Bamberg, Leibniz Institute for Educational Trajectories, National Educational Panel Study)

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Mul-*

tilevel, longitudinal, and structural equation models. Boca Raton: Chapman & Hall/CRC.

Smit, A., Kelderman, H., & Van der Flier, H. (1999). Collateral information and mixed rasch models. *Methods of Psychological Research Online*, *4*, 19-32.

Smit, A., Kelderman, H., & Van der Flier, H. (2000). The mixed birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online*, *5*, 31-43.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis. an introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, *15*(2), 201-292.

Steinhausen, R.-M. C., H.-C., & Seidel, R. (1991). Follow-up studies of anorexia nervosa: A review of four decades of outcome research. *Psychological Medicine*, *21*(2), 447-454.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, *62*(4), 795-809.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-549.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications. advantages, disadvantages and some neglected topics. *Medical Care*, *44*(11), S152-S170.

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rpart` (R package version 4.1-10)

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, *22*(4), 1701-1728.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations. *Journal of Statistical Software*, *45*(3), 1-67.

van Dyk, D. A., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1-50.

von Davier, M., & Carstensen, C. H. (Eds.). (2007). *Multivariate and mixture distribution rasch models*. New York, NY: Springer Science + Business Media.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the national assessment of educational progress (naep): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (vol. 26): Psychometrics* (p. 1039-1055). Amsterdam: Elsevier.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H.

(2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Education as a lifelong process. The german national educational panel study (neps)* (p. 67-86). Wiesbaden: VS Verlag für Sozialwissenschaften.

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education*, *2*(9), 1-18.

Wilson, M., & De Boeck, P. (2004). *Explanatory item response models*. New York, NY: Springer.

Xie, Y. (2013). Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6262-6268.

Zinn, S. (2013). *An imputation model for multilevel binary data (neps working paper no. 31)*. Bamberg: University of Bamberg, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Zwinderman, A. H. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, *56*(4), 589-600.

A Tables

Table 3.1 Prior specifications and starting values for the MGLRM, FMLRM and RILRM

Model	Parameter	Probability distribution	$\psi^{(0)}$
MGLRM	$\underline{\gamma}_g$	$\mathcal{N}(\underline{\nu}_{\gamma_g}, \Omega_{\gamma_g})$	0
	$\sigma^2_{\epsilon,g}$	$\mathcal{IG}(a^0_{\sigma^2_{\epsilon,g}}, b^0_{\sigma^2_{\epsilon,g}})$	1
	$\tilde{\underline{\xi}}_j = (\tilde{\alpha}_j, \tilde{\beta}_j)'$	$\mathcal{N}(\underline{\nu}_{\tilde{\xi}_j}, \Omega_{\tilde{\xi}_j})\mathbf{1}(\tilde{\alpha}_j > 0)$	$(1, 0)$
	$\underline{\tau}_j$	$\mathcal{N}(\underline{\nu}_{\tau_j}, \Omega_{\tau_j})$	0
FMLRM	$\underline{\boldsymbol{\zeta}}$	$\mathcal{N}(\underline{\nu}_{\zeta}, \Omega_{\zeta})$	0
	$\underline{\gamma}_g$	$\mathcal{N}(\underline{\nu}_{\gamma_g}, \Omega_{\gamma_g})$	0
	$\sigma^2_{\epsilon,g}$	$\mathcal{IG}(a^0_{\sigma^2_{\epsilon,g}}, b^0_{\sigma^2_{\epsilon,g}})$	1
	$\tilde{\underline{\xi}}_j = (\tilde{\alpha}_j, \tilde{\beta}_j)'$	$\mathcal{N}(\underline{\nu}_{\tilde{\xi}_j}, \Omega_{\tilde{\xi}_j})\mathbf{1}(\tilde{\alpha}_j > 0)$	$(1, 0)$
	$\underline{\tau}_j$	$\mathcal{N}(\underline{\nu}_{\tau_j}, \Omega_{\tau_j})$	0
RILRM	$\underline{\gamma}$	$\mathcal{N}(\underline{\nu}_{\gamma}, \Omega_{\gamma})$	0
	σ^2_{ϵ}	$\mathcal{IG}(a^0_{\sigma^2_{\epsilon}}, b^0_{\sigma^2_{\epsilon}})$	1
	σ^2_{ω}	$\mathcal{IG}(a^0_{\sigma^2_{\omega}}, b^0_{\sigma^2_{\omega}})$	1
	$\tilde{\underline{\xi}}_j = (\tilde{\alpha}_j, \tilde{\beta}_j)'$	$\mathcal{N}(\underline{\nu}_{\tilde{\xi}_j}, \Omega_{\tilde{\xi}_j})\mathbf{1}(\tilde{\alpha}_j > 0)$	$(1, 0)$
	$\underline{\tau}_j$	$\mathcal{N}(\underline{\nu}_{\tau_j}, \Omega_{\tau_j})$	0

Notes: MGLRM = multigroup latent regression item response model; FMLRM = finite mixture latent regression item response model; RILRM = random intercept latent regression item response model.

Table 4.1 Simulation studies, scenario 1—true parameter values, mean posterior means and standard deviations over 200 replications obtained from BD, DAC and DAR

Data: 2MGLRM, estimates: 2MGLRM							
		$\overline{\text{Mean}}$			$\overline{\text{Sd}}$		
Parameter	True	BD	DAC	DAR	BD	DAC	DAR
Regression weight							
$\gamma_{1,0}$	−0.500	−0.498	−0.496	−0.468	0.029	0.030	0.031
$\gamma_{1,1}$	0.200	0.199	0.200	0.207	0.014	0.015	0.016
$\gamma_{1,2}$	0.200	0.201	0.198	0.162	0.014	0.015	0.016
$\gamma_{2,0}$	1.000	1.005	1.004	1.008	0.024	0.025	0.027
$\gamma_{2,1}$	0.400	0.402	0.397	0.362	0.013	0.014	0.015
$\gamma_{2,2}$	−0.200	−0.202	−0.195	−0.163	0.012	0.013	0.014
Variance							
$\sigma^2_{\epsilon,1}$	0.490	0.494	0.498	0.559	0.030	0.030	0.033
$\sigma^2_{\epsilon,2}$	0.250	0.259	0.269	0.350	0.020	0.021	0.025
Item discrimination							
α_1	1.017	1.014	1.014	1.014	0.046	0.046	0.046
α_2	0.964	0.968	0.968	0.968	0.044	0.044	0.044
α_3	1.326	1.330	1.330	1.330	0.062	0.062	0.062
α_4	1.080	1.078	1.078	1.078	0.050	0.051	0.051
α_5	0.867	0.866	0.866	0.866	0.041	0.041	0.041
α_6	0.979	0.981	0.980	0.980	0.047	0.047	0.047
α_7	0.775	0.778	0.778	0.778	0.038	0.038	0.038
α_8	1.095	1.097	1.097	1.097	0.049	0.049	0.049
α_9	0.850	0.852	0.852	0.852	0.040	0.040	0.040
α_{10}	1.164	1.169	1.169	1.168	0.052	0.052	0.052
α_{11}	1.111	1.118	1.118	1.118	0.051	0.051	0.052
α_{12}	0.784	0.788	0.788	0.787	0.042	0.042	0.042
α_{13}	1.107	1.120	1.120	1.120	0.049	0.049	0.049
α_{14}	1.412	1.415	1.415	1.415	0.064	0.064	0.064
α_{15}	0.917	0.922	0.923	0.923	0.042	0.042	0.042
α_{16}	0.779	0.781	0.780	0.781	0.039	0.039	0.039
α_{17}	0.841	0.839	0.839	0.840	0.040	0.040	0.040
α_{18}	1.119	1.121	1.121	1.121	0.049	0.049	0.049
α_{19}	0.865	0.862	0.862	0.862	0.033	0.033	0.033
α_{20}	1.261	1.259	1.260	1.260	0.043	0.043	0.044

Notes: $N = 2000$; $J = 20$. BD = before deletion; DAC = data augmentation using sequential CART imputation; DAR = data augmentation using sequential stochastic regresion imputation; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.1 Simulation studies, scenario 1—true parameter values, mean posterior means and standard deviations over 200 replications obtained from BD, DAC and DAR

Data: 2MGLRM, estimates: 2MGLRM							
		$\overline{\text{Mean}}$			$\overline{\text{Sd}}$		
Parameter	True	BD	DAC	DAR	BD	DAC	DAR
Item difficulty							
β_1	−0.070	−0.072	−0.072	−0.072	0.036	0.036	0.036
β_2	−0.082	−0.084	−0.084	−0.084	0.036	0.036	0.036
β_3	−0.196	−0.200	−0.200	−0.200	0.039	0.039	0.039
β_4	−0.375	−0.382	−0.382	−0.382	0.037	0.037	0.037
β_5	−0.237	−0.238	−0.238	−0.238	0.035	0.035	0.035
β_6	−0.466	−0.467	−0.467	−0.467	0.036	0.036	0.036
β_7	−0.327	−0.329	−0.329	−0.329	0.034	0.034	0.034
β_8	0.867	0.870	0.870	0.871	0.048	0.048	0.048
β_9	−0.166	−0.164	−0.164	−0.165	0.035	0.035	0.035
β_{10}	0.008	0.007	0.007	0.007	0.038	0.038	0.038
β_{11}	−0.252	−0.253	−0.253	−0.253	0.037	0.037	0.037
β_{12}	−0.644	−0.646	−0.646	−0.646	0.036	0.036	0.036
β_{13}	0.522	0.531	0.531	0.531	0.042	0.042	0.042
β_{14}	0.858	0.861	0.861	0.861	0.052	0.052	0.053
β_{15}	0.032	0.032	0.032	0.032	0.036	0.036	0.036
β_{16}	−0.340	−0.338	−0.338	−0.338	0.034	0.034	0.034
β_{17}	0.887	0.890	0.890	0.891	0.045	0.045	0.045
β_{18}	0.301	0.301	0.302	0.301	0.040	0.039	0.040
β_{19}	0.101	0.099	0.099	0.100	0.035	0.035	0.035
β_{20}	−0.412	−0.419	−0.419	−0.419	0.039	0.038	0.039
Item category cutoff							
$\kappa_{19,1}$	0.500	0.500	0.500	0.500	0.028	0.028	0.028
$\kappa_{19,2}$	1.000	1.005	1.005	1.005	0.037	0.037	0.037
$\kappa_{20,1}$	0.700	0.699	0.699	0.699	0.039	0.039	0.039
$\kappa_{20,2}$	1.400	1.403	1.403	1.404	0.051	0.051	0.051

Notes: $N = 2000$; $J = 20$. BD = data set before deletion; DAC = data augmentation using sequential CART imputation; DAR = data augmentation using sequential stochastic regresion imputation; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.2 Simulation studies, scenario 1—RMSEs and coverage ratios over 200 replications obtained from BD, DAC and DAR

Data: 2MGLRM, estimates: 2MGLRM						
	RMSE			Coverage		
Parameter	BD	DAC	DAR	BD	DAC	DAR
Regression weight						
$\gamma_{1,0}$	0.032	0.032	0.052	0.930	0.925	0.755
$\gamma_{1,1}$	0.014	0.014	0.019	0.970	0.965	0.890
$\gamma_{1,2}$	0.015	0.015	0.043	0.945	0.945	0.405
$\gamma_{2,0}$	0.023	0.025	0.031	0.965	0.960	0.910
$\gamma_{2,1}$	0.013	0.014	0.043	0.950	0.960	0.360
$\gamma_{2,2}$	0.012	0.014	0.039	0.960	0.930	0.235
Variance						
$\sigma^2_{\epsilon,1}$	0.029	0.030	0.094	0.950	0.950	0.550
$\sigma^2_{\epsilon,2}$	0.022	0.029	0.108	0.925	0.890	0.040
Item discrimination						
α_1	0.042	0.042	0.043	0.970	0.980	0.965
α_2	0.040	0.041	0.041	0.970	0.965	0.970
α_3	0.061	0.062	0.062	0.940	0.930	0.930
α_4	0.055	0.055	0.055	0.925	0.930	0.930
α_5	0.048	0.048	0.048	0.905	0.905	0.910
α_6	0.048	0.048	0.048	0.955	0.955	0.955
α_7	0.037	0.038	0.037	0.950	0.950	0.955
α_8	0.052	0.053	0.053	0.940	0.935	0.950
α_9	0.042	0.042	0.042	0.955	0.945	0.945
α_{10}	0.055	0.054	0.054	0.925	0.935	0.935
α_{11}	0.055	0.055	0.055	0.940	0.940	0.940
α_{12}	0.042	0.042	0.042	0.910	0.910	0.915
α_{13}	0.053	0.053	0.054	0.930	0.930	0.920
α_{14}	0.069	0.069	0.069	0.925	0.915	0.925
α_{15}	0.041	0.041	0.041	0.965	0.965	0.965
α_{16}	0.039	0.039	0.039	0.955	0.940	0.945
α_{17}	0.040	0.040	0.040	0.970	0.960	0.955
α_{18}	0.050	0.051	0.050	0.940	0.930	0.930
α_{19}	0.030	0.030	0.030	0.955	0.960	0.955
α_{20}	0.041	0.041	0.041	0.940	0.955	0.950

Notes: $N = 2000$; $J = 20$. RMSE = root mean square error; BD = data set before deletion; DAC = data augmentation using sequential CART imputation; DAR = data augmentation using sequential stochastic regresion imputation; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.2 Simulation studies, scenario 1—RMSEs and coverage ratios over 200 replications obtained from BD, DAC and DAR

Data: 2MGLRM, estimates: 2MGLRM						
	RMSE			Coverage		
Parameter	BD	DAC	DAR	BD	DAC	DAR
Item difficulty						
β_1	0.036	0.036	0.036	0.935	0.945	0.935
β_2	0.034	0.034	0.034	0.965	0.970	0.970
β_3	0.039	0.039	0.039	0.955	0.960	0.960
β_4	0.036	0.036	0.036	0.955	0.960	0.960
β_5	0.036	0.036	0.035	0.955	0.960	0.960
β_6	0.038	0.038	0.038	0.940	0.930	0.940
β_7	0.034	0.034	0.034	0.960	0.955	0.960
β_8	0.054	0.054	0.054	0.925	0.925	0.920
β_9	0.036	0.036	0.036	0.930	0.935	0.920
β_{10}	0.039	0.039	0.039	0.935	0.935	0.940
β_{11}	0.037	0.037	0.037	0.950	0.950	0.955
β_{12}	0.034	0.034	0.034	0.955	0.950	0.950
β_{13}	0.048	0.048	0.049	0.910	0.920	0.915
β_{14}	0.056	0.056	0.057	0.915	0.915	0.915
β_{15}	0.037	0.037	0.037	0.945	0.945	0.940
β_{16}	0.033	0.033	0.033	0.960	0.960	0.955
β_{17}	0.046	0.047	0.047	0.965	0.960	0.960
β_{18}	0.038	0.038	0.038	0.945	0.940	0.945
β_{19}	0.035	0.035	0.035	0.945	0.950	0.955
β_{20}	0.039	0.039	0.039	0.950	0.955	0.945
Item category cutoff						
$\kappa_{19,1}$	0.027	0.027	0.028	0.940	0.930	0.930
$\kappa_{19,2}$	0.037	0.037	0.038	0.940	0.945	0.935
$\kappa_{20,1}$	0.041	0.041	0.041	0.930	0.930	0.925
$\kappa_{20,2}$	0.055	0.055	0.055	0.940	0.945	0.935

Notes: $N = 2000$; $J = 20$. RMSE = root mean square error; BD = data set before deletion; DAC = data augmentation using sequential CART imputation; DAR = data augmentation using sequential stochastic regresion imputation; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.3 Simulation studies, scenario 2—true parameter values, mean posterior means and standard deviations over 200 replications obtained from BD, DAC and CC

Data: 2MGLRM, estimates: 2MGLRM							
		$\overline{\text{Mean}}$			$\overline{\text{Sd}}$		
Parameter	True	BD	DAC	CC	BD	DAC	CC
Regression weight							
$\gamma_{1,0}$	−0.500	−0.501	−0.507	−0.230	0.036	0.038	0.044
$\gamma_{1,1}$	0.200	0.202	0.207	0.171	0.015	0.017	0.017
$\gamma_{1,2}$	0.200	0.200	0.194	0.171	0.015	0.017	0.017
$\gamma_{1,3}$	0.300	0.300	0.293	0.251	0.055	0.062	0.061
$\gamma_{2,0}$	1.000	1.005	0.989	1.035	0.029	0.030	0.030
$\gamma_{2,1}$	0.400	0.402	0.397	0.383	0.013	0.014	0.014
$\gamma_{2,2}$	−0.200	−0.202	−0.196	−0.192	0.012	0.013	0.013
$\gamma_{2,3}$	−0.500	−0.502	−0.488	−0.475	0.045	0.048	0.048
Variance							
$\sigma^2_{\epsilon,1}$	0.490	0.496	0.515	0.421	0.030	0.033	0.031
$\sigma^2_{\epsilon,2}$	0.250	0.257	0.277	0.245	0.019	0.020	0.019
Item discrimination							
α_1	1.017	1.022	1.022	1.020	0.047	0.047	0.060
α_2	0.964	0.971	0.971	0.970	0.045	0.045	0.058
α_3	1.326	1.327	1.327	1.314	0.062	0.062	0.078
α_4	1.080	1.083	1.084	1.076	0.052	0.052	0.067
α_5	0.867	0.867	0.868	0.867	0.042	0.042	0.055
α_6	0.979	0.980	0.981	0.979	0.048	0.049	0.064
α_7	0.775	0.779	0.780	0.784	0.040	0.040	0.053
α_8	1.095	1.095	1.094	1.105	0.051	0.051	0.060
α_9	0.850	0.853	0.853	0.851	0.041	0.041	0.054
α_{10}	1.164	1.172	1.171	1.176	0.053	0.054	0.067
α_{11}	1.111	1.120	1.120	1.120	0.053	0.053	0.068
α_{12}	0.784	0.788	0.790	0.782	0.043	0.043	0.059
α_{13}	1.107	1.109	1.108	1.112	0.050	0.050	0.060
α_{14}	1.412	1.408	1.406	1.418	0.065	0.065	0.075
α_{15}	0.917	0.922	0.922	0.924	0.043	0.043	0.055
α_{16}	0.779	0.780	0.781	0.787	0.040	0.040	0.053
α_{17}	0.841	0.838	0.838	0.843	0.042	0.042	0.050
α_{18}	1.119	1.121	1.120	1.126	0.050	0.050	0.062
α_{19}	0.865	0.863	0.863	0.870	0.035	0.035	0.043
α_{20}	1.261	1.258	1.258	1.260	0.044	0.044	0.055

Notes: $N = 2000$; $J = 20$. BD = data set before deletion; DAC = data augmentation using sequential CART imputation; CC = complete cases analysis; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.3 Simulation studies, scenario 2—true parameter values, mean posterior means and standard deviations over 200 replications obtained from BD, DAC and CC

Data: 2MGLRM, estimates: 2MGLRM							
		$\overline{\text{Mean}}$			$\overline{\text{Sd}}$		
Parameter	True	BD	DAC	CC	BD	DAC	CC
Item difficulty							
β_1	−0.070	−0.071	−0.070	−0.073	0.036	0.036	0.046
β_2	−0.082	−0.080	−0.080	−0.080	0.035	0.035	0.046
β_3	−0.196	−0.194	−0.195	−0.201	0.038	0.038	0.048
β_4	−0.375	−0.380	−0.380	−0.384	0.036	0.036	0.046
β_5	−0.237	−0.234	−0.234	−0.236	0.034	0.034	0.045
β_6	−0.466	−0.470	−0.469	−0.472	0.036	0.036	0.046
β_7	−0.327	−0.327	−0.326	−0.327	0.034	0.034	0.044
β_8	0.867	0.873	0.872	0.880	0.047	0.047	0.058
β_9	−0.166	−0.162	−0.162	−0.166	0.034	0.034	0.045
β_{10}	0.008	0.003	0.003	0.006	0.037	0.037	0.048
β_{11}	−0.252	−0.253	−0.253	−0.253	0.036	0.036	0.046
β_{12}	−0.644	−0.646	−0.645	−0.651	0.035	0.035	0.046
β_{13}	0.522	0.523	0.523	0.521	0.042	0.041	0.052
β_{14}	0.858	0.855	0.854	0.862	0.051	0.051	0.062
β_{15}	0.032	0.030	0.030	0.031	0.035	0.035	0.046
β_{16}	−0.340	−0.342	−0.341	−0.338	0.034	0.034	0.044
β_{17}	0.887	0.888	0.887	0.887	0.044	0.044	0.055
β_{18}	0.301	0.299	0.299	0.304	0.039	0.039	0.050
β_{19}	0.101	0.101	0.101	0.103	0.035	0.035	0.044
β_{20}	−0.412	−0.414	−0.415	−0.414	0.038	0.038	0.047
Item category cutoff							
$\kappa_{19,1}$	0.500	0.503	0.503	0.504	0.027	0.027	0.030
$\kappa_{19,2}$	1.000	0.999	0.999	0.999	0.037	0.037	0.040
$\kappa_{20,1}$	0.700	0.699	0.699	0.697	0.037	0.037	0.043
$\kappa_{20,2}$	1.400	1.399	1.400	1.397	0.049	0.049	0.055

Notes: $N = 2000$; $J = 20$. BD = data set before deletion; DAC = data augmentation using sequential CART imputation; CC = complete cases analysis; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.4 Simulation studies, scenario 2—RMSEs and coverage ratios over 200 replications obtained from BD, DAC and CC

Data: 2MGLRM, estimates: 2MGLRM						
	RMSE			Coverage		
Parameter	BD	DAC	CC	BD	DAC	CC
Regression weight						
$\gamma_{1,0}$	0.038	0.039	0.273	0.935	0.960	0.000
$\gamma_{1,1}$	0.015	0.019	0.034	0.940	0.900	0.610
$\gamma_{1,2}$	0.015	0.018	0.034	0.940	0.930	0.555
$\gamma_{1,3}$	0.054	0.061	0.077	0.950	0.945	0.875
$\gamma_{2,0}$	0.030	0.033	0.047	0.945	0.950	0.770
$\gamma_{2,1}$	0.014	0.014	0.022	0.940	0.940	0.730
$\gamma_{2,2}$	0.012	0.013	0.015	0.955	0.950	0.910
$\gamma_{2,3}$	0.042	0.047	0.051	0.955	0.965	0.935
Variance						
$\sigma^2_{\epsilon,1}$	0.030	0.041	0.075	0.965	0.880	0.445
$\sigma^2_{\epsilon,2}$	0.021	0.035	0.022	0.920	0.715	0.890
Item discrimination						
α_1	0.051	0.051	0.063	0.915	0.925	0.940
α_2	0.051	0.051	0.060	0.925	0.930	0.935
α_3	0.064	0.063	0.084	0.940	0.940	0.940
α_4	0.052	0.052	0.063	0.950	0.945	0.960
α_5	0.042	0.042	0.055	0.945	0.935	0.950
α_6	0.046	0.046	0.066	0.950	0.955	0.940
α_7	0.039	0.039	0.054	0.965	0.955	0.945
α_8	0.053	0.053	0.065	0.925	0.930	0.920
α_9	0.039	0.039	0.054	0.965	0.970	0.945
α_{10}	0.052	0.052	0.065	0.970	0.970	0.970
α_{11}	0.058	0.058	0.074	0.920	0.905	0.920
α_{12}	0.040	0.040	0.060	0.965	0.965	0.930
α_{13}	0.051	0.051	0.059	0.950	0.940	0.965
α_{14}	0.061	0.061	0.071	0.965	0.965	0.965
α_{15}	0.043	0.044	0.060	0.950	0.935	0.930
α_{16}	0.037	0.037	0.053	0.955	0.955	0.940
α_{17}	0.045	0.045	0.053	0.920	0.915	0.945
α_{18}	0.051	0.051	0.062	0.940	0.940	0.940
α_{19}	0.036	0.036	0.043	0.940	0.945	0.970
α_{20}	0.046	0.047	0.056	0.940	0.915	0.930

Notes: $N = 2000$; $J = 20$. RMSE = root mean square error; BD = data set before deletion; DAC = data augmentation using sequential CART imputation; CC = complete cases analysis; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.4 Simulation studies, scenario 2—RMSEs and coverage ratios over 200 replications obtained from BD, DAC and CC

Data: 2MGLRM, estimates: 2MGLRM						
	RMSE			Coverage		
Parameter	BD	DAC	CC	BD	DAC	CC
Item difficulty						
β_1	0.035	0.035	0.047	0.965	0.965	0.965
β_2	0.034	0.034	0.044	0.970	0.965	0.960
β_3	0.037	0.037	0.047	0.965	0.960	0.960
β_4	0.035	0.035	0.050	0.940	0.935	0.905
β_5	0.036	0.036	0.046	0.930	0.935	0.915
β_6	0.037	0.037	0.046	0.935	0.925	0.950
β_7	0.033	0.033	0.045	0.955	0.960	0.955
β_8	0.048	0.047	0.063	0.935	0.945	0.925
β_9	0.034	0.034	0.049	0.965	0.955	0.930
β_{10}	0.036	0.036	0.046	0.955	0.945	0.960
β_{11}	0.034	0.034	0.046	0.975	0.975	0.960
β_{12}	0.035	0.035	0.049	0.960	0.960	0.945
β_{13}	0.043	0.044	0.051	0.940	0.930	0.960
β_{14}	0.047	0.047	0.057	0.980	0.970	0.970
β_{15}	0.037	0.037	0.048	0.935	0.935	0.970
β_{16}	0.033	0.032	0.046	0.955	0.960	0.935
β_{17}	0.047	0.047	0.055	0.935	0.940	0.955
β_{18}	0.039	0.039	0.050	0.955	0.955	0.945
β_{19}	0.036	0.036	0.044	0.925	0.920	0.945
β_{20}	0.035	0.035	0.040	0.960	0.965	0.990
Item category cutoff						
$\kappa_{19,1}$	0.030	0.030	0.033	0.925	0.925	0.935
$\kappa_{19,2}$	0.038	0.038	0.041	0.930	0.935	0.940
$\kappa_{20,1}$	0.039	0.039	0.043	0.940	0.940	0.950
$\kappa_{20,2}$	0.049	0.050	0.054	0.950	0.955	0.955

Notes: $N = 2000$; $J = 20$. RMSE = root mean square error; BD = data set before deletion; DAC = data augmentation using sequential CART imputation; CC = complete cases analysis; 2MGLRM = two-group multigroup latent regression item response model.

Table 4.5 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—variable information, response format and frequency distribution of the test items

			Response				
Variable	Name	Format	0	1	2	3	% missing
item01	mag9q071_c	MC	0.37	0.62	–	–	0.01
item02	mag9v131_c	MC	0.49	0.50	–	–	0.01
item03	mag9v13s_c	CMC	0.10	0.13	0.35	0.34	0.08
item04	mag9r261_c	MC	0.86	0.10	–	–	0.03
item05	mag9r111_c	MC	0.33	0.64	–	–	0.02
item06	mag9d171_c	MC	0.49	0.49	–	–	0.01
item07	mag9d151_c	MC	0.24	0.76	–	–	0.01
item08	mag9r051_c	MC	0.58	0.41	–	–	0.01
item09	mag9v011_c	MC	0.32	0.67	–	–	0.01
item10	mag9v012_c	MC	0.45	0.53	–	–	0.02
item11	mag9q161_c	MC	0.67	0.31	–	–	0.03
item12	mag9d201_c	MC	0.54	0.45	–	–	0.01
item13	mag9r191_c	MC	0.33	0.66	–	–	0.01
item14	mag9v121_c	MC	0.73	0.26	–	–	0.01
item15	mag9q181_c	MC	0.14	0.86	–	–	0.00
item16	mag9r25s_c	CMC	0.07	0.29	0.20	0.36	0.09
item17	mag9r061_c	SCR	0.53	0.26	–	–	0.21
item18	mag9q081_c	MC	0.54	0.43	–	–	0.03
item19	mag9q101_c	MC	0.33	0.62	–	–	0.05
item20	mag9q021_c	MC	0.51	0.44	–	–	0.05
item22	mag9v091_c	MC	0.41	0.54	–	–	0.04
item23	mag9q211_c	MC	0.49	0.47	–	–	0.05

Notes: $N = 13075$. MC = simple multiple-choice; CMC = complex multiple-choice; SCR = short constructed response. All items are stored in data file xTargetCompetencies.

Table 4.6 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—variable information and description of background variables

Variable	Data file	Name	Content
schooltype	CohortProfile	t723080_g1	type of secondary school according to the German education system: $\{HS, RS, GYM\}$
school	CohortProfile	ID_i	unique number assigned to each school
female	pTarget	t700031	dichotomous variable indicating whether the student is female
repeat	pTarget	p725020	dichotomous variable indicating whether the student ever repeated a school year
hisei	pTarget	p731422_g14, p731472_g14	highest occupational status of parents according to ISEI-08

Table 4.7 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—summary statistics of background variables

	Complete cases summary				Missing	
Variable	Min	Max	Mean	Sd	No.	%
schooltype:RS	0	1	0.32	–	0	0.0
schooltype:GYM	0	1	0.40	–	0	0.0
school	1	511	–	–	0	0.0
female	0	1	0.50	–	28	0.2
repeat	0	1	0.19	–	346	2.6
hisei	1.16	8.90	5.13	2.07	2538	19.4

Notes: $N = 13075$. Minimum and maximum values, means, standard deviations, absolute and relative counts are reported.

Table 4.8 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—structural parameter estimates obtained from $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$

Dependent variable: *mathematical competency*

			$g =$"HS"		$g =$"RS"		$g =$"GYM"	
Model	Parameter		Mean	Sd	Mean	Sd	Mean	Sd
$\mathcal{M}_1$	γ_0	(*constant*)	0.181*	0.006	–	–	–	–
	σ^2_ϵ		0.409*	0.007	–	–	–	–
$\mathcal{M}_2$	$\gamma_{g,0}$	(*constant*)	−0.317*	0.008	0.036*	0.008	0.678*	0.009
	$\sigma^2_{\epsilon,g}$		0.148*	0.006	0.201*	0.007	0.333*	0.009
$\mathcal{M}_3$	$\gamma_{g,0}$	(*constant*)	−0.260*	0.023	0.118*	0.025	0.610*	0.031
	$\gamma_{g,1}$	(*female*)	−0.204*	0.016	−0.324*	0.016	−0.336*	0.018
	$\gamma_{g,2}$	(*repeat*)	−0.114*	0.016	−0.056*	0.019	−0.338*	0.031
	$\gamma_{g,3}$	(*hisei*)	0.018*	0.005	0.019*	0.005	0.046*	0.005
	$\sigma^2_{\epsilon,g}$		0.134*	0.005	0.173*	0.006	0.289*	0.008

Notes: $N = 13075$; $J = 22$. Means and standard deviations of the posterior distributions are reported. * indicate the 95% highest density region does not include zero.

Table 4.9 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—structural parameter estimates obtained from $\mathcal{M}_4$ and $\mathcal{M}_5$

Dependent variable: *mathematical competency*				
Model	Parameter		Mean	Sd
$\mathcal{M}_4$	γ_0	(*constant*)	0.093*	0.021
	σ^2_ϵ		0.180*	0.004
	υ^2_ω		0.223*	0.015
$\mathcal{M}_5$	γ_0	(*constant*)	−0.224*	0.023
	γ_1	(*female*)	−0.279*	0.009
	γ_2	(*repeat*)	−0.107*	0.012
	γ_3	(*schooltype:RS*)	0.306*	0.026
	γ_4	(*schooltype:GYM*)	0.892*	0.027
	γ_5	(*hisei*)	0.020*	0.003
	σ^2_ϵ		0.159*	0.003
	υ^2_ω		0.054*	0.004

Notes: $N = 13075$; $J = 22$. Means and standard deviations of the posterior distributions are reported. * indicate the 95% highest density region does not include zero.

Table 4.10 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—item parameter estimates obtained from $\mathcal{M}_1$, $\mathcal{M}_3$ and $\mathcal{M}_5$

	Model					
	$\mathcal{M}_1$		$\mathcal{M}_3$		$\mathcal{M}_5$	
Parameter	Mean	Sd	Mean	Sd	Mean	Sd
Item discrimination						
α_1	0.922	0.024	0.928	0.024	0.925	0.023
α_2	0.963	0.024	0.966	0.023	0.958	0.023
α_3	0.928	0.020	0.952	0.020	0.946	0.020
α_4	1.117	0.034	1.051	0.030	1.112	0.033
α_5	1.046	0.026	1.035	0.026	1.029	0.025
α_6	0.708	0.021	0.698	0.020	0.702	0.020
α_7	1.318	0.033	1.379	0.034	1.326	0.032
α_8	0.844	0.022	0.845	0.022	0.850	0.022
α_9	1.235	0.029	1.243	0.029	1.211	0.028
α_{10}	1.165	0.027	1.165	0.025	1.164	0.025
α_{11}	0.644	0.021	0.652	0.020	0.653	0.021
α_{12}	1.002	0.024	1.021	0.023	1.010	0.023
α_{13}	0.778	0.023	0.785	0.023	0.783	0.022
α_{14}	0.848	0.023	0.828	0.022	0.833	0.022
α_{15}	0.968	0.032	0.997	0.034	0.976	0.031
α_{16}	0.743	0.019	0.748	0.019	0.733	0.018
α_{17}	1.251	0.031	1.164	0.028	1.230	0.029
α_{18}	0.969	0.024	0.938	0.023	0.950	0.023
α_{19}	1.510	0.035	1.499	0.033	1.476	0.032
α_{20}	1.135	0.026	1.141	0.025	1.149	0.025
α_{21}	1.338	0.030	1.364	0.029	1.365	0.029
α_{22}	1.086	0.025	1.121	0.025	1.129	0.025

Notes: $N = 13075$; $J = 22$. Means and standard deviations of the posterior distributions are reported. * indicate the 95% highest density region does not include zero.

Table 4.10 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—item parameter estimates obtained from $\mathcal{M}_1$, $\mathcal{M}_3$ and $\mathcal{M}_5$

	Model					
	$\mathcal{M}_1$		$\mathcal{M}_3$		$\mathcal{M}_5$	
Parameter	Mean	Sd	Mean	Sd	Mean	Sd
Item difficulty						
β_1	−0.211	0.012	−0.207	0.012	−0.211	0.012
β_2	0.138	0.012	0.141	0.012	0.137	0.012
β_3	−1.246	0.016	−1.242	0.016	−1.248	0.016
β_4	1.736	0.026	1.707	0.024	1.727	0.025
β_5	−0.306	0.012	−0.303	0.012	−0.306	0.012
β_6	0.127	0.012	0.128	0.012	0.125	0.012
β_7	−0.700	0.014	−0.704	0.014	−0.702	0.013
β_8	0.392	0.013	0.396	0.012	0.393	0.012
β_9	−0.380	0.013	−0.378	0.013	−0.379	0.013
β_{10}	0.078	0.013	0.079	0.013	0.078	0.012
β_{11}	0.638	0.013	0.644	0.013	0.639	0.013
β_{12}	0.288	0.012	0.294	0.012	0.289	0.013
β_{13}	−0.353	0.012	−0.348	0.012	−0.352	0.012
β_{14}	0.866	0.014	0.864	0.014	0.860	0.014
β_{15}	−1.101	0.015	−1.098	0.015	−1.103	0.015
β_{16}	−1.414	0.017	−1.407	0.017	−1.409	0.017
β_{17}	0.883	0.018	0.856	0.018	0.879	0.018
β_{18}	0.337	0.013	0.335	0.012	0.334	0.012
β_{19}	−0.290	0.013	−0.288	0.013	−0.287	0.013
β_{20}	0.311	0.013	0.316	0.013	0.317	0.013
β_{21}	−0.002	0.012	0.001	0.012	0.003	0.013
β_{22}	0.207	0.013	0.216	0.013	0.216	0.013
Item category cutoff						
$\kappa_{3,2}$	0.632	0.014	0.634	0.014	0.635	0.014
$\kappa_{3,3}$	1.808	0.020	1.817	0.019	1.815	0.020
$\kappa_{16,2}$	1.249	0.018	1.248	0.018	1.244	0.018
$\kappa_{16,3}$	1.858	0.020	1.858	0.020	1.851	0.020

Notes: $N = 13075$; $J = 22$. Means and standard deviations of the posterior distributions are reported. * indicate the 95% highest density region does not include zero.

Table 4.11 NEPS GRADE 9, EATING DISORDERS—variable information, item wording and frequency distribution of the SCOFF questionnaire

			Response	
Variable	Name	Item wording	0	1
scoff01	t526300	"Do you make yourself **S**ick because you feel uncomfortably full?"	0.94	0.06
scoff02	t526301	"Do you worry you have lost **C**ontrol over how much you eat?"	0.75	0.25
scoff03	t526302	"Have you recently lost more than **O**ne stone in a three month period?"	0.88	0.12
scoff04	t526303	"Do you believe yourself to be **F**at when others say you are too thin?"	0.80	0.20
scoff05	t526304	"Would you say that **F**ood dominates your life?"	0.68	0.32

Notes: $N = 12460$. All items are stored in data file pTarget.

Table 4.12 NEPS GRADE 9, EATING DISORDERS—variable information and description of background variables

Variable	Data file	Name	Content
female	pTarget	t700031	dichotomous variable indicating whether the student is female
bmi	pTarget	t520000_g1, t520001_g1	body mass index in units of kg/m^2

Table 4.13 NEPS GRADE 9, EATING DISORDERS—summary statistics of background variables

	Complete cases summary				Missing	
Variable	Min	Max	Mean	Sd	No.	%
female	0	1	0.49	–	24	0.2
bmi	10.01	49.38	20.81	3.24	1327	10.7

Notes: $N = 12460$. Minimum and maximum values, means, standard deviations, absolute and relative counts are reported.

Table 4.14 NEPS GRADE 9, EATING DISORDERS—structural parameter estimates obtained from $\mathcal{M}_6$ and $\mathcal{M}_7$

Dependent variable: *Eating disorder*				
Model	Parameter		Mean	Sd
$\mathcal{M}_6$	γ_0	$(constant)$	-1.020^*	0.009
	σ^2_ϵ		0.341^*	0.013
$\mathcal{M}_7$	γ_0	$(constant)$	-2.425^*	0.055
	γ_1	$(female)$	0.430^*	0.016
	γ_2	(bmi)	0.057^*	0.002
	σ^2_ϵ		0.268^*	0.012

Notes: $N = 12460$; $J = 5$. Means and standard deviations of the posterior distributions are reported. * indicate the 95% highest density region does not include zero.

Table 4.15 NEPS GRADE 9, EATING DISORDERS—structural parameter estimates obtained from $\mathcal{M}_8$, $\mathcal{M}_9$ and $\mathcal{M}_{10}$

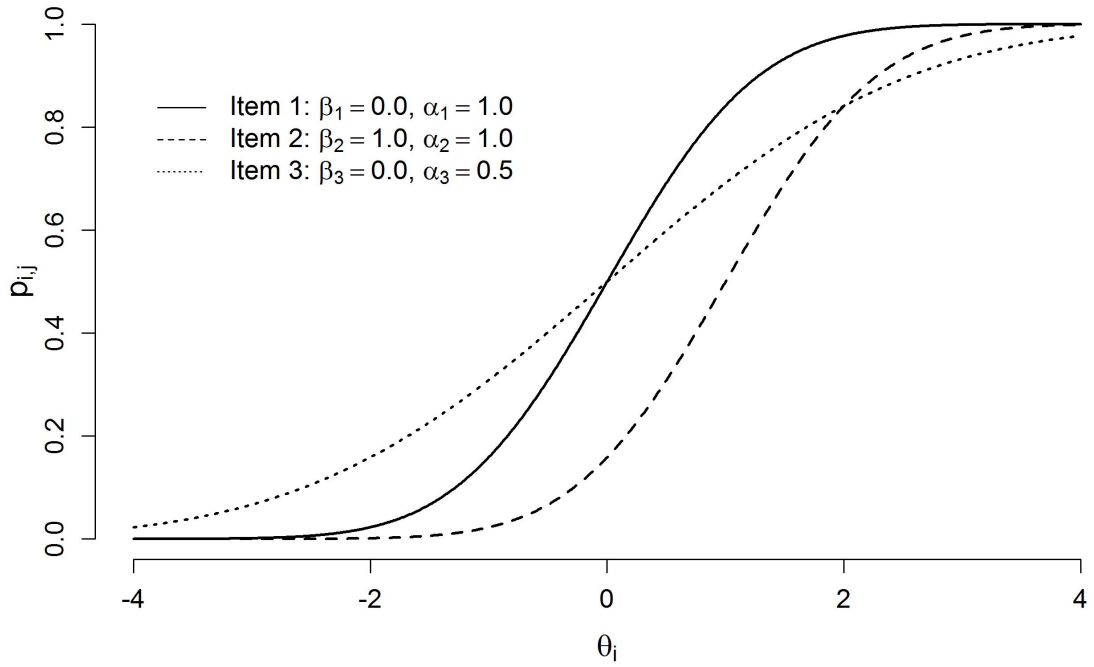
Dependent variable: *Eating disorder*

			$g = 1$		$g = 2$	
Model	Parameter		Mean	Sd	Mean	Sd
$\mathcal{M}_8$	$\zeta_{g,0}$	$(constant)$	-0.035	0.194	0.000	0.000
	$\gamma_{g,0}$	$(constant)$	-1.038^*	0.016	-1.003^*	0.017
	$\sigma^2_{\epsilon,g}$		0.346^*	0.022	0.340^*	0.021
$\mathcal{M}_9$	$\zeta_{g,0}$	$(constant)$	0.080	0.259	0.000	0.000
	$\gamma_{g,0}$	$(constant)$	-2.591	0.128	-2.243	0.144
	$\gamma_{g,1}$	$(female)$	0.430^*	0.041	0.429^*	0.044
	$\gamma_{g,2}$	(bmi)	0.065^*	0.006	0.049^*	0.006
	$\sigma^2_{\epsilon,g}$		0.264^*	0.020	0.273^*	0.020
$\mathcal{M}_{10}$	$\zeta_{g,0}$	$(constant)$	0.228	0.442	0.000	0.000
	$\zeta_{g,1}$	(bmi)	-0.007	0.015	0.000	0.000
	$\gamma_{g,0}$	$(constant)$	-2.610^*	0.135	-2.230^*	0.153
	$\gamma_{g,1}$	$(female)$	0.431^*	0.040	0.429^*	0.043
	$\gamma_{g,2}$	(bmi)	0.066^*	0.006	0.049^*	0.007
	$\sigma^2_{\epsilon,g}$		0.264^*	0.020	0.273^*	0.021

Notes: $N = 12460$; $J = 5$. Means and standard deviations of the posterior distributions are reported. * indicate the 95% highest density region does not include zero.

B Figures

Figure 2.1 Graphical representation of three IRFs in the 2PNO IRT model



Notes: IRF = item response function; 2PNO IRT = two-parameter normal ogive item response theory.

Figure 2.2 Threshold mechanism for an ordinal four-category item

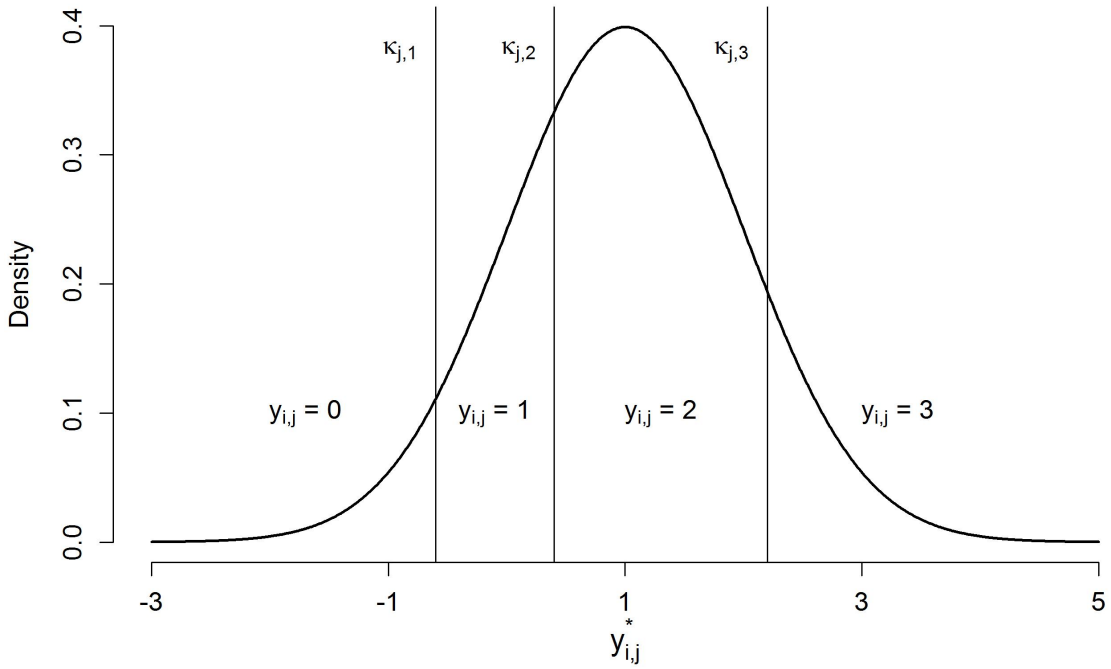


Figure 2.3 Path diagram of a latent trait variable explained by two person covariates affecting the response to three items

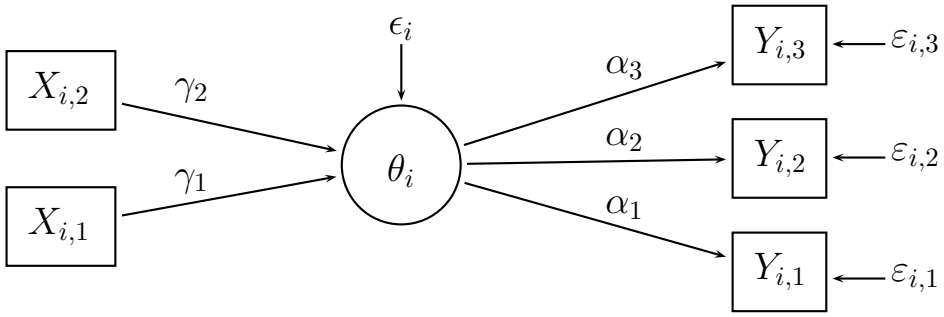


Figure 4.1 Classification tree applied to kyphosis data

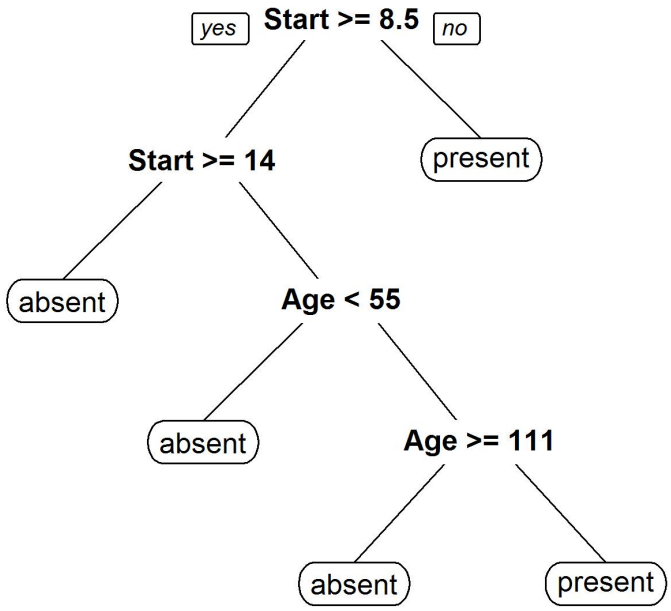


Figure 4.2 Pillars and stages of the NEPS. Adapted from "The National Educational Panel Study: Milestones of the years 2006 to 2015," by J. von Maurice, H.-P. Blossfeld and H.-G. Roßbach, in H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological issues of longitudinal surveys: The example of the national educational panel study* (p. 8), 2016, Wiesbaden: Springer VS. Copyright 2016 by Leibniz Institute for Educational Trajectories. Adapted with permission.

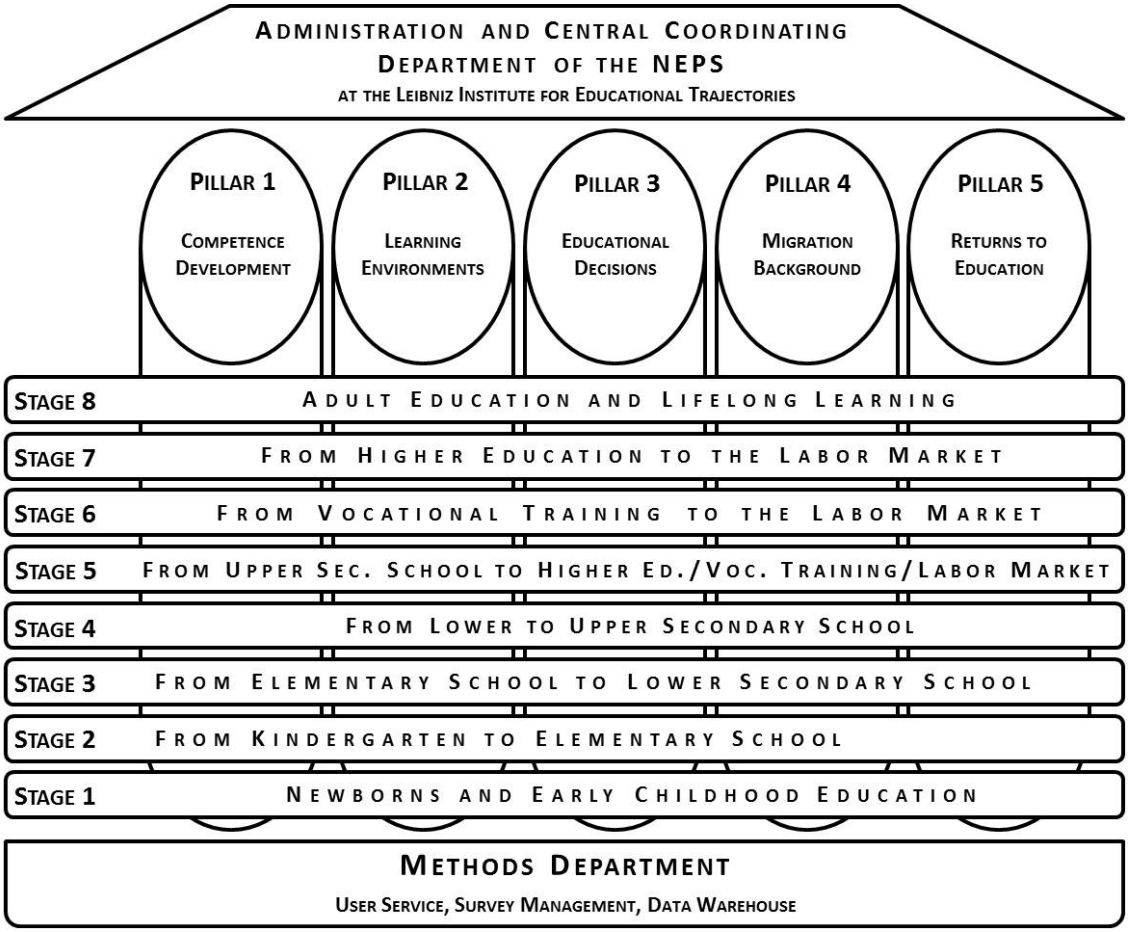


Figure 4.3 Multicohort sequence design of the NEPS. Adapted from "The National Educational Panel Study: Milestones of the years 2006 to 2015," by J. von Maurice, H.-P. Blossfeld and H.-G. Roßbach, in H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological issues of longitudinal surveys: The example of the national educational panel study* (p. 9), 2016, Wiesbaden: Springer VS. Copyright 2016 by Leibniz Institute for Educational Trajectories. Adapted with permission.

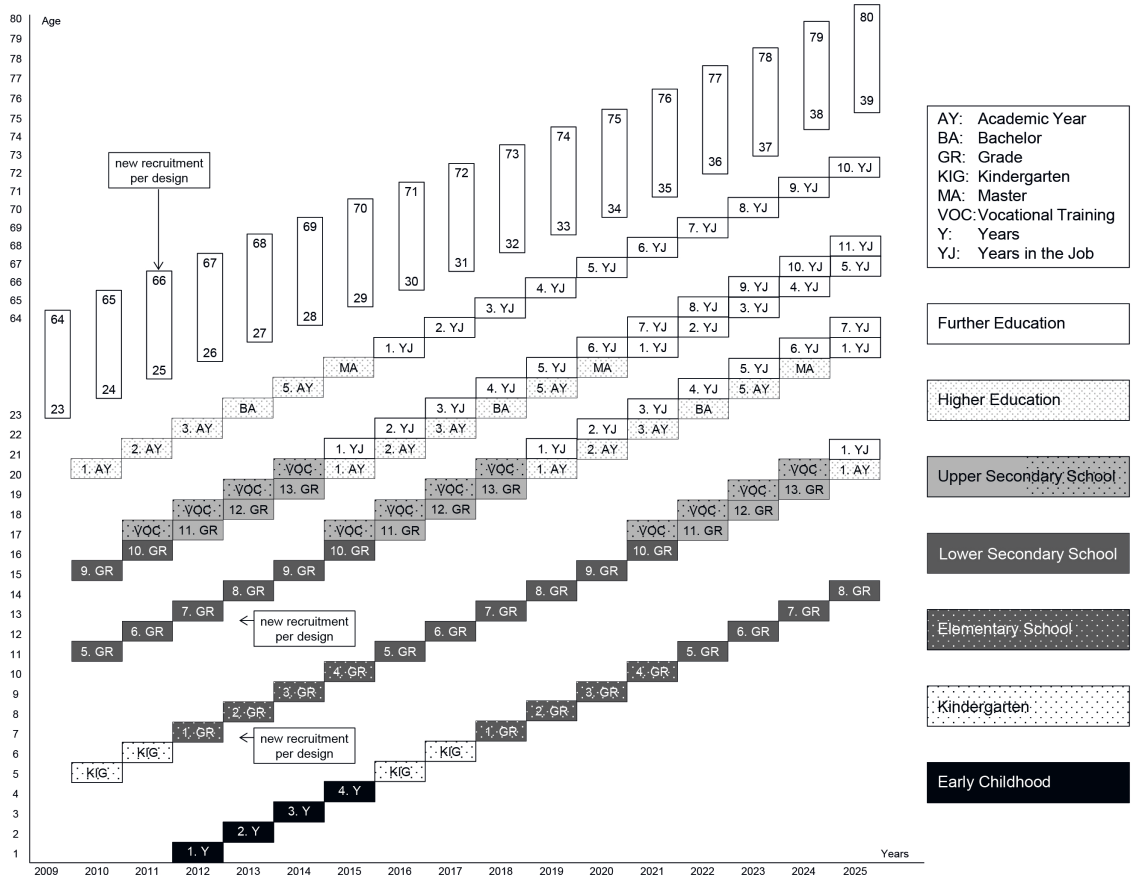
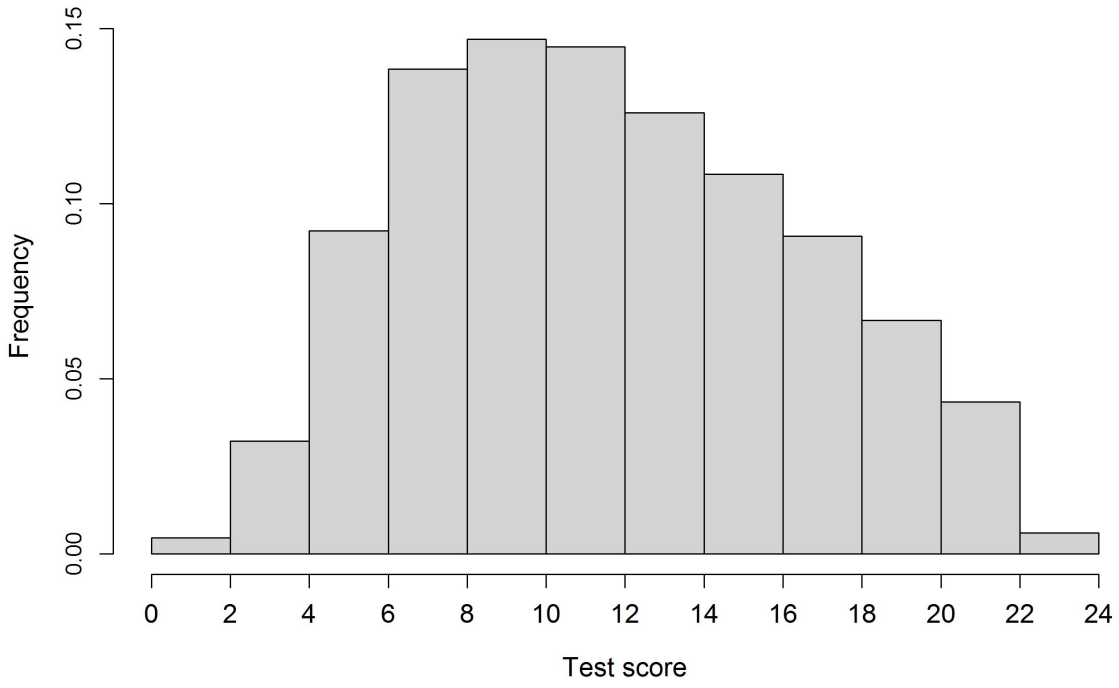


Figure 4.4 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—histogram of grouped test scores



Notes: Ordered polytomous items are scored $\tilde{y}_{i,j} = 0.5y_{i,j}$.

Figure 4.5 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—trace plots and cumulative means for $\mathcal{M}_5$

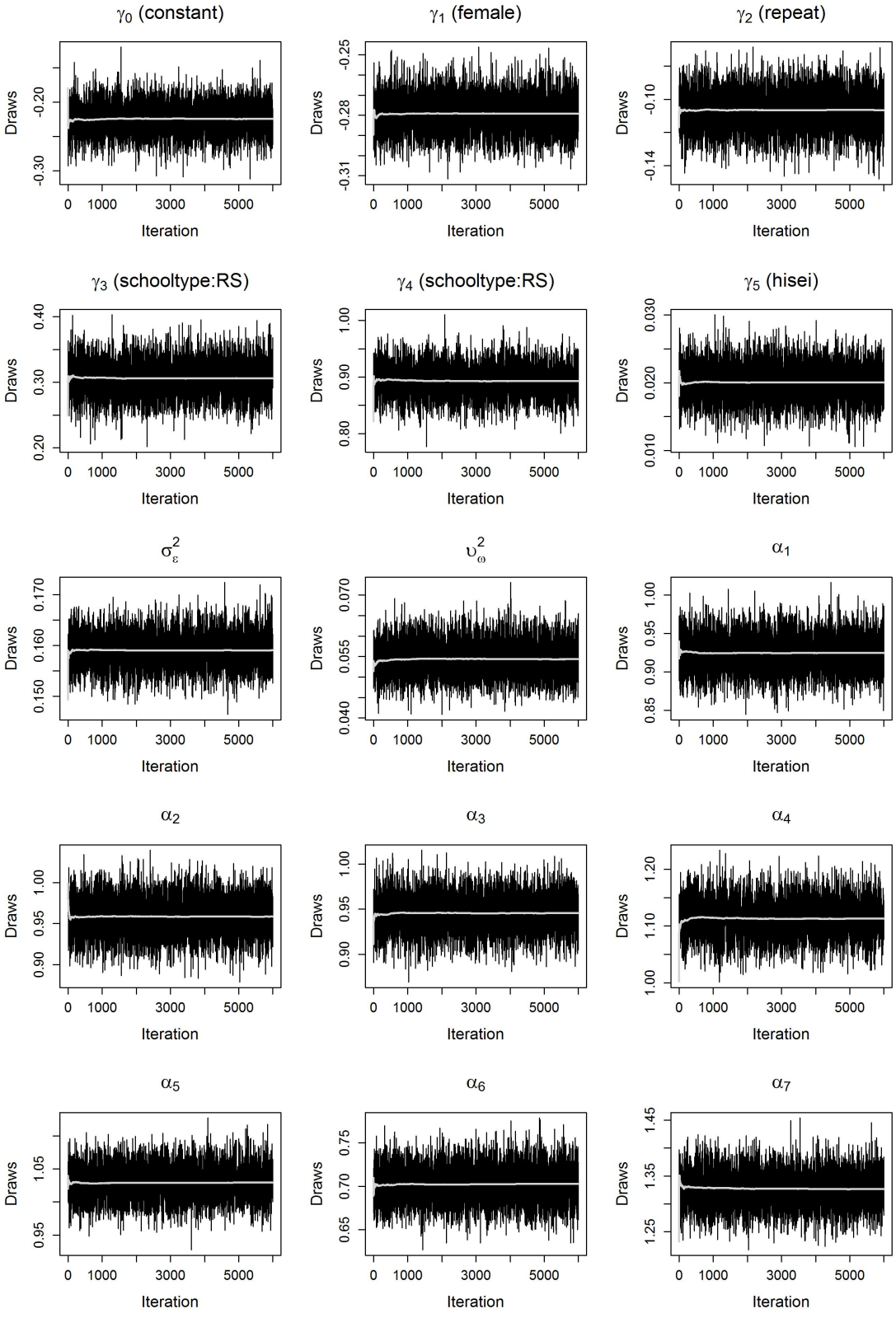


Figure 4.5 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—trace plots and cumulative means for $\mathcal{M}_5$

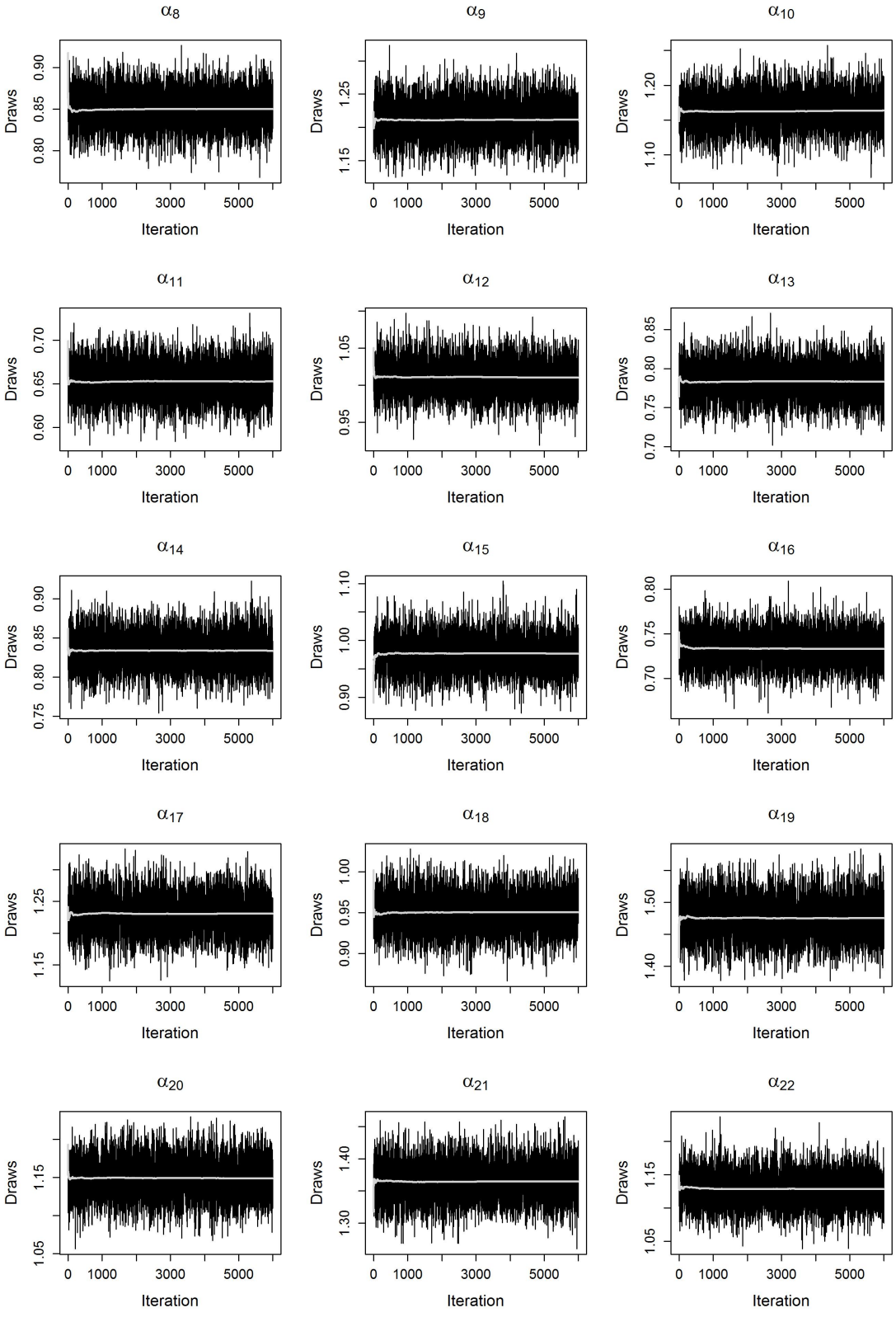


Figure 4.5 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—trace plots and cumulative means for $\mathcal{M}_5$

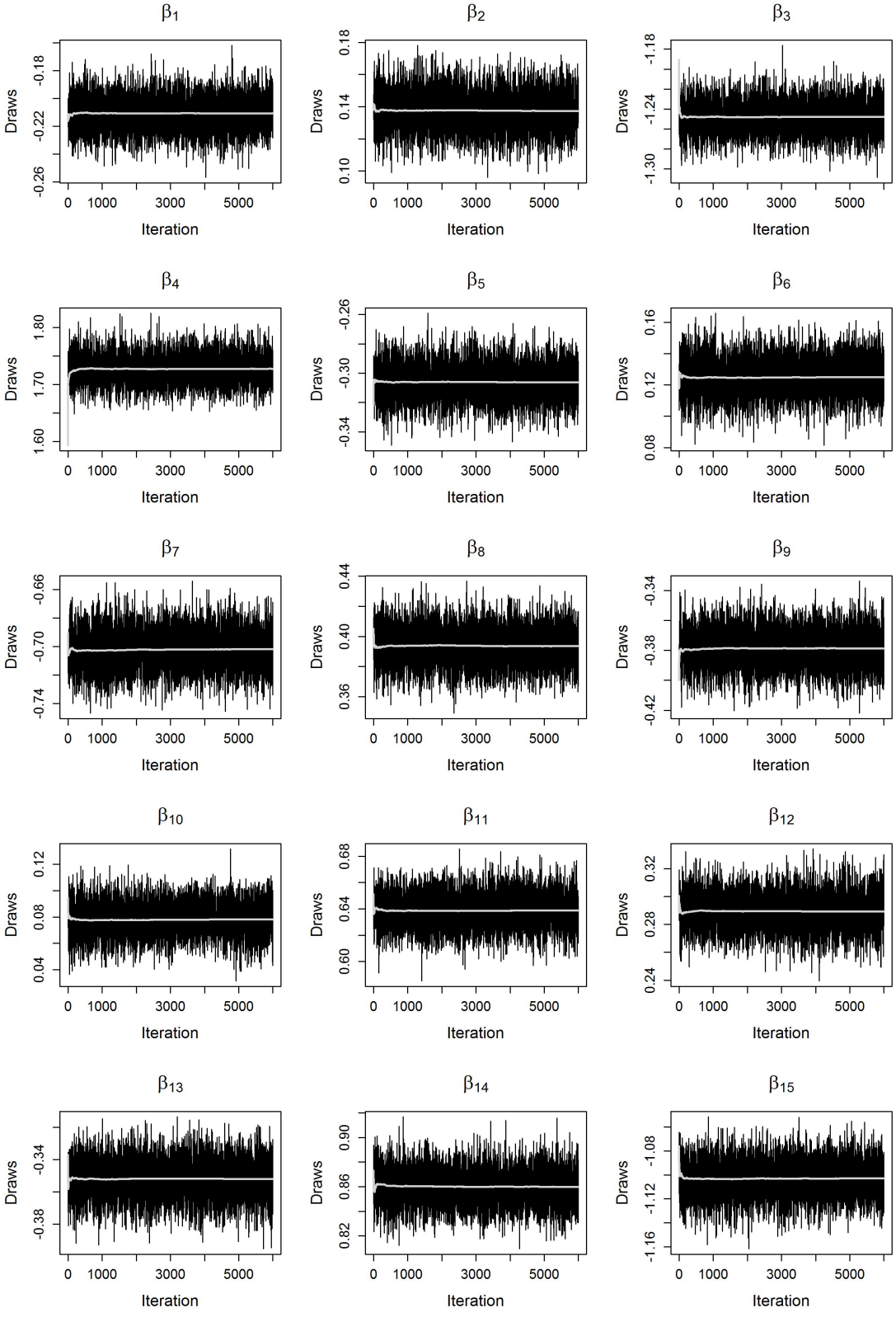


Figure 4.5 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—trace plots and cumulative means for $\mathcal{M}_5$

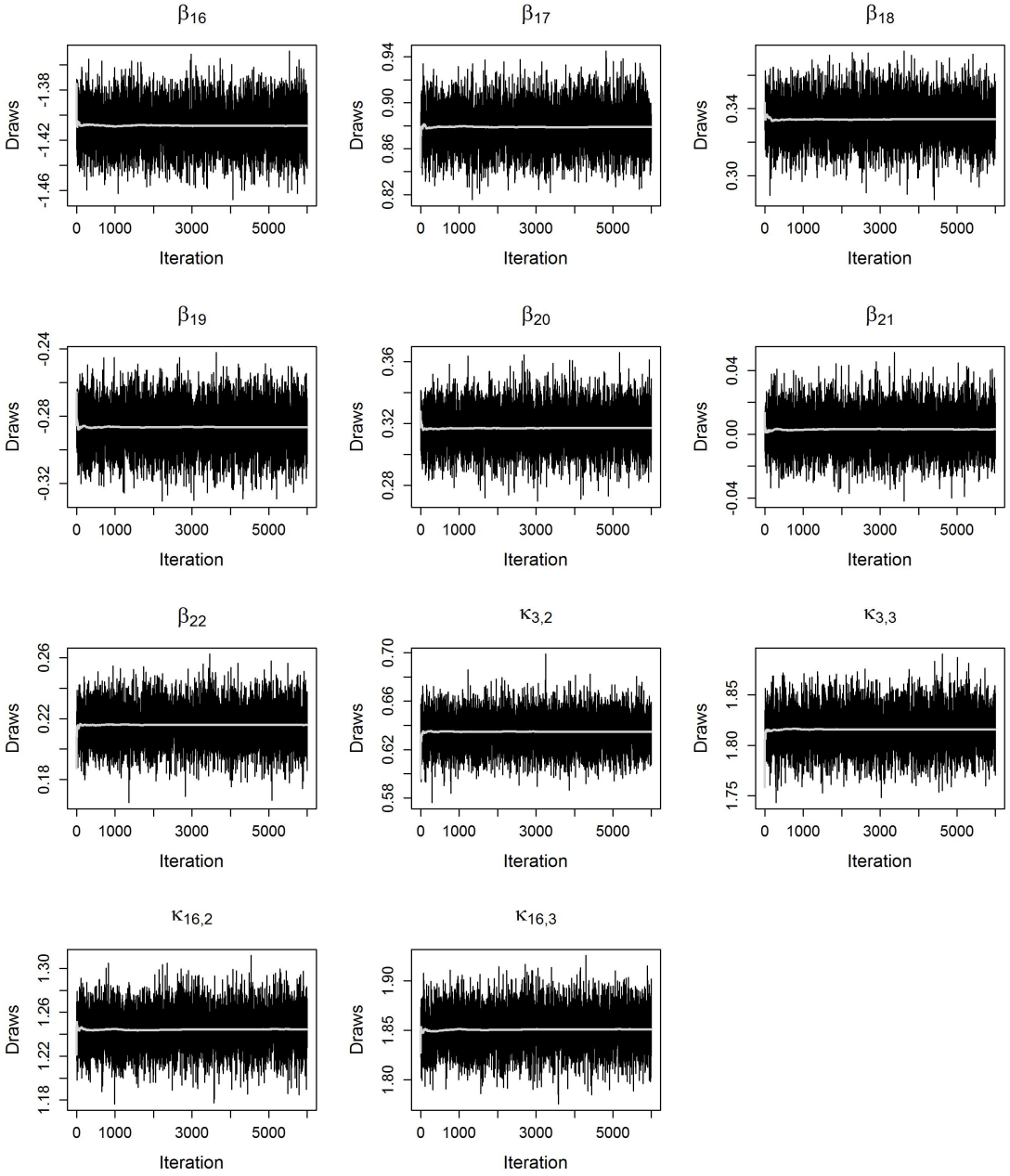


Figure 4.6 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—post burn-in lag-1 autocorrelation functions for $\mathcal{M}_5$

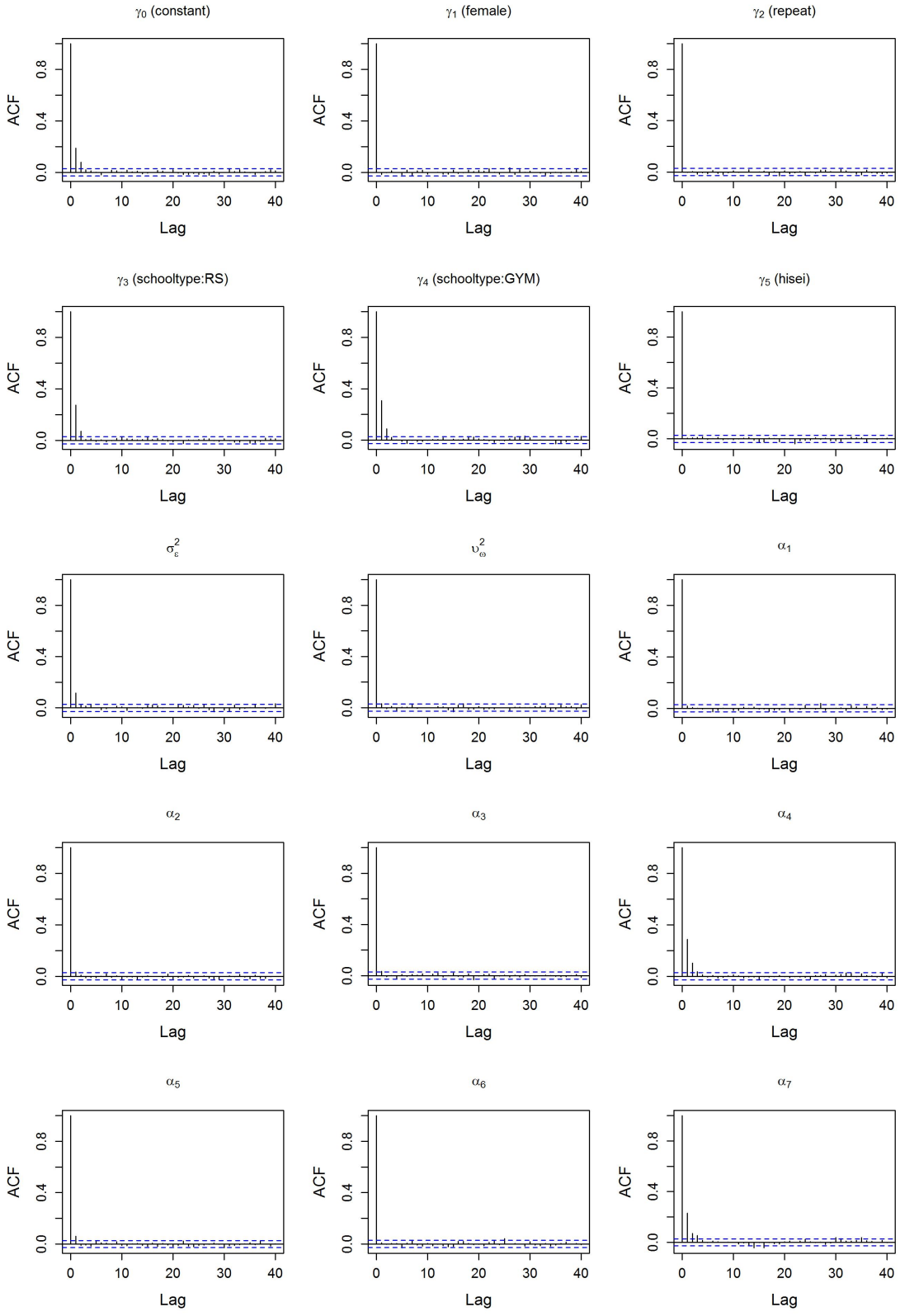


Figure 4.6 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—post burn-in lag-1 autocorrelation functions for $\mathcal{M}_5$

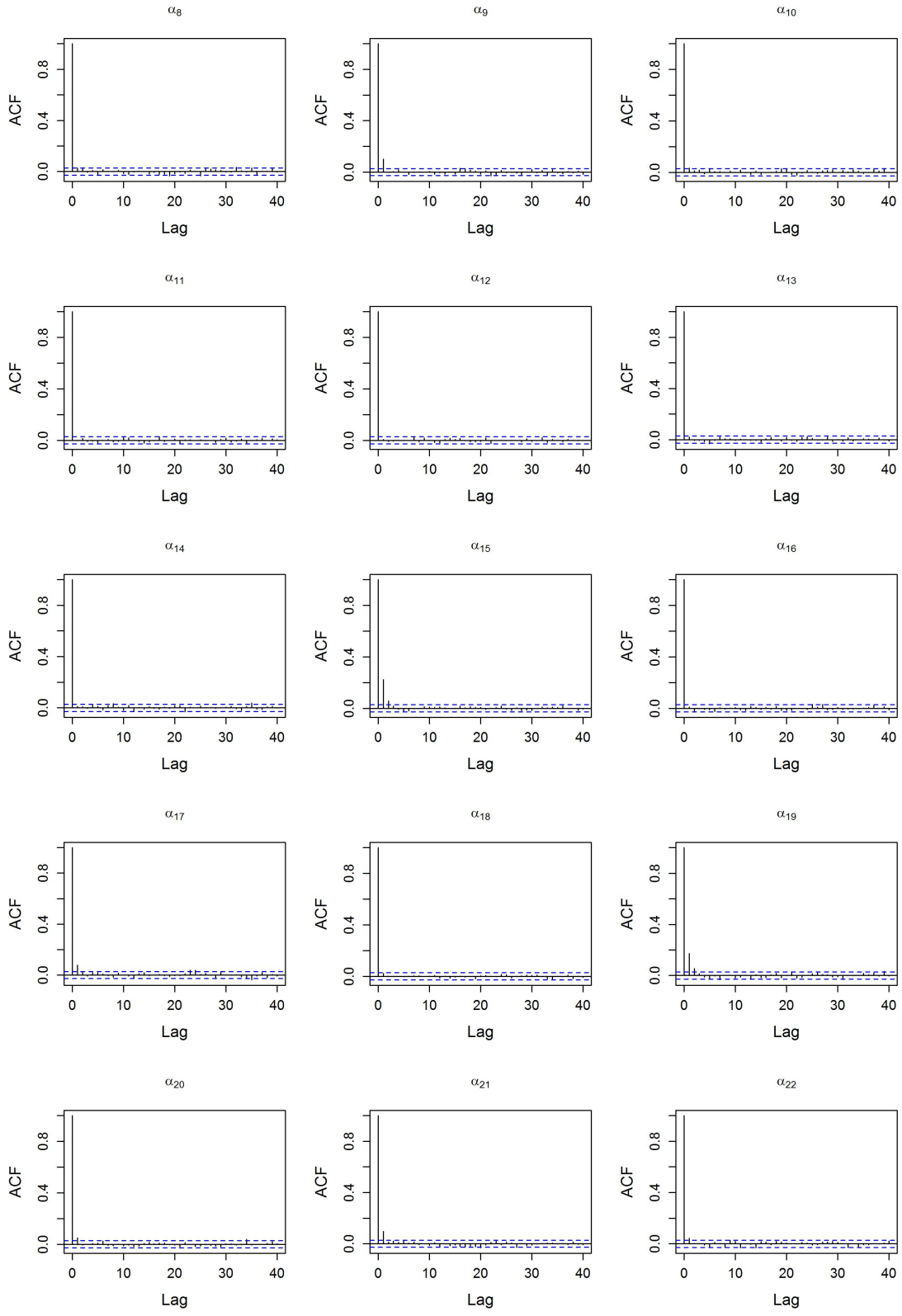


Figure 4.6 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—post burn-in lag-1 autocorrelation functions for $\mathcal{M}_5$

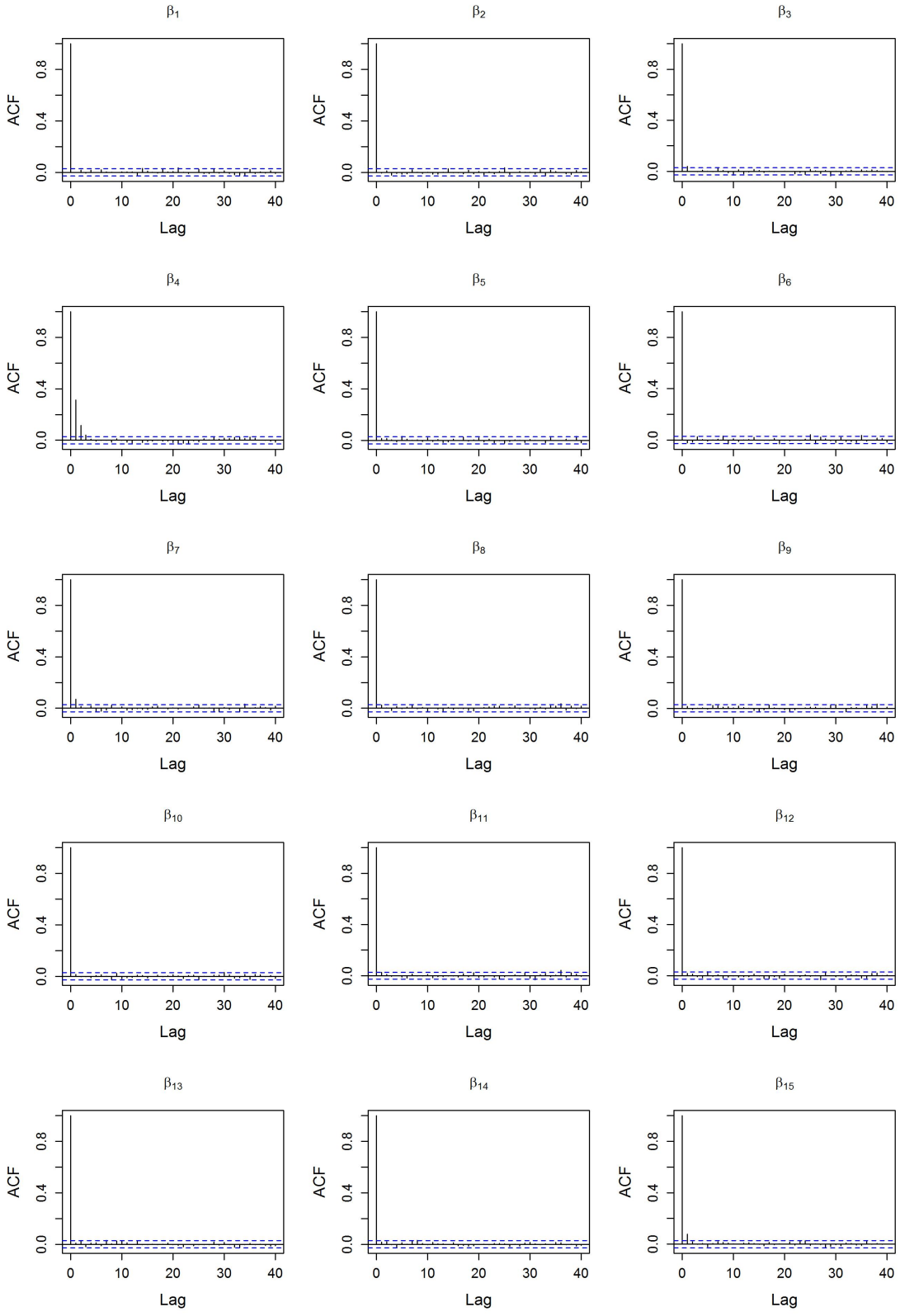


Figure 4.6 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—post burn-in lag-1 autocorrelation functions for $\mathcal{M}_5$

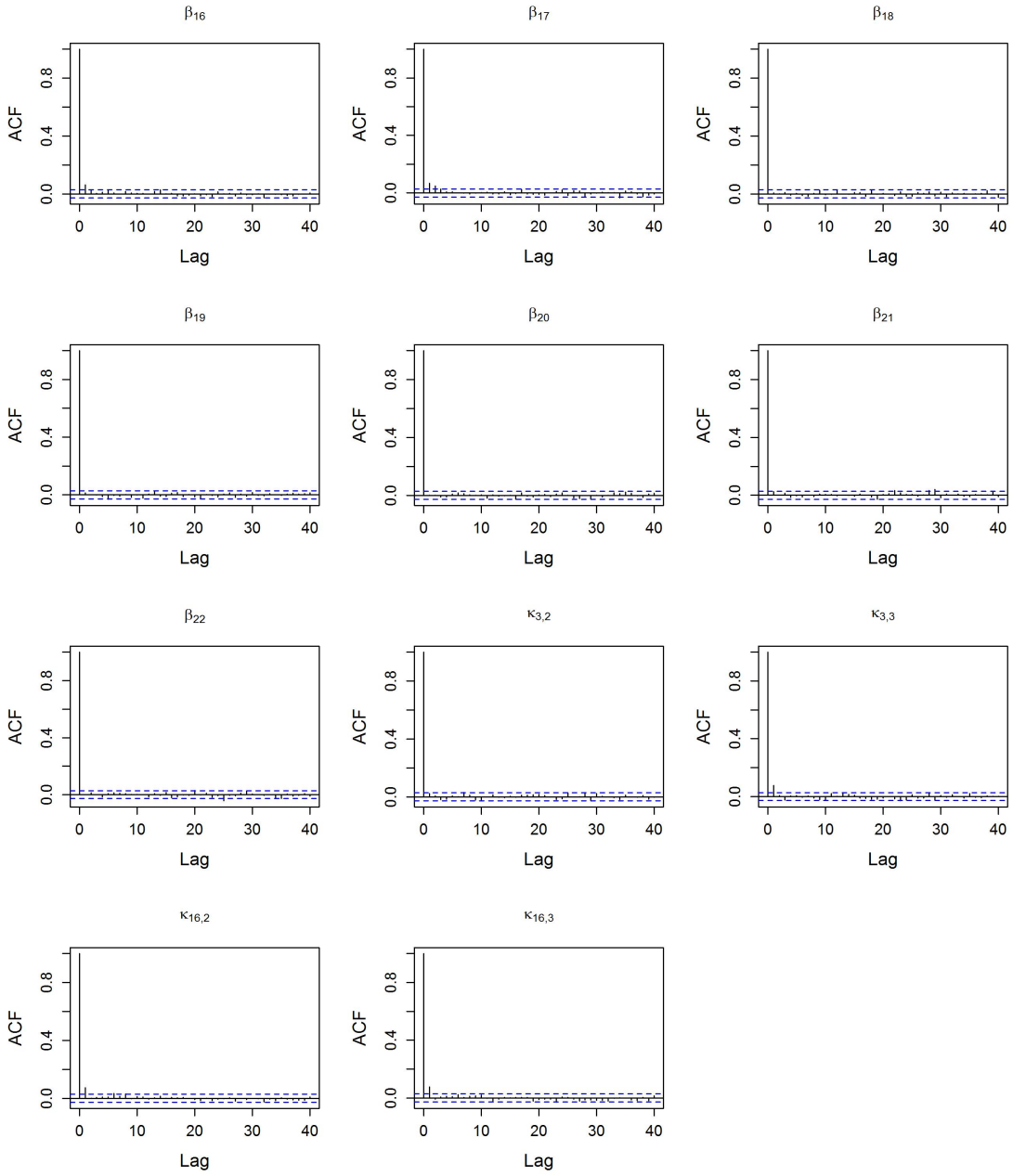


Figure 4.7 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—kernel density estimates for the set of expected a posteriori estimates obtained from $\mathcal{M}_3$

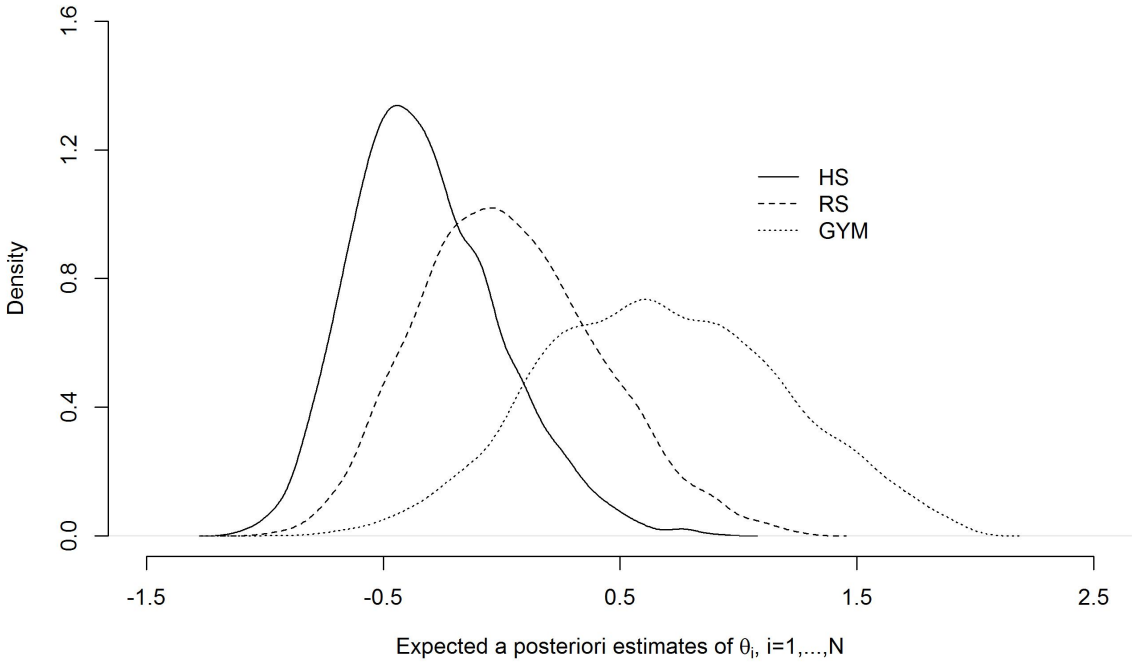
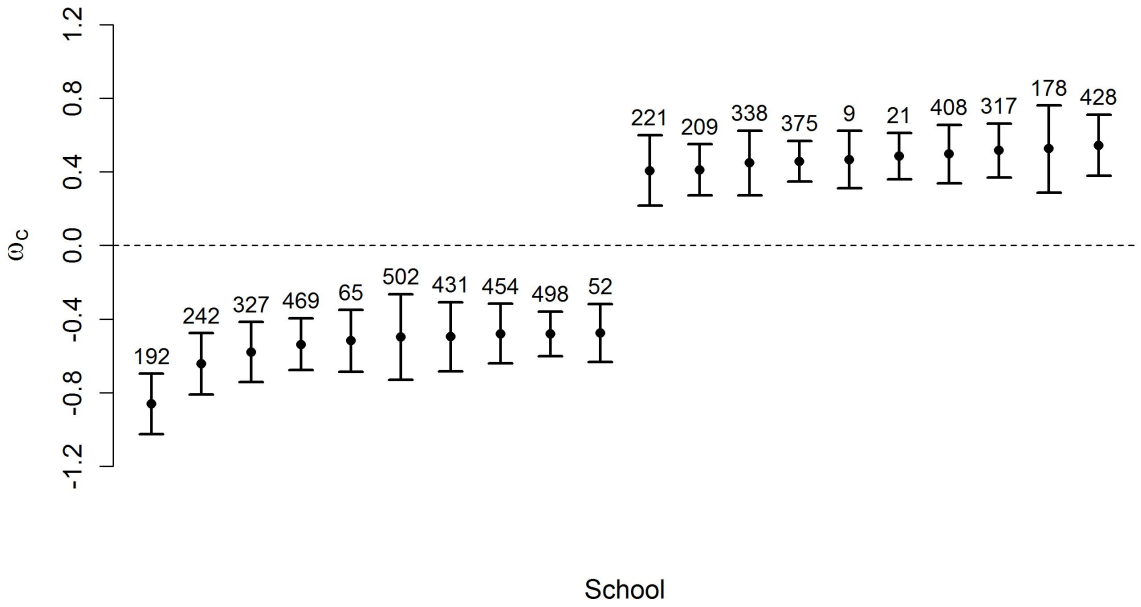


Figure 4.8 NEPS GRADE 9, MATHEMATICAL COMPETENCIES—posterior means and 95% HDR of ten smallest and ten largest random intercepts obtained from $\mathcal{M}_5$



Notes: HDR = highest density region.

Figure 4.9 NEPS GRADE 9, EATING DISORDERS—barplot of SCOFF scores

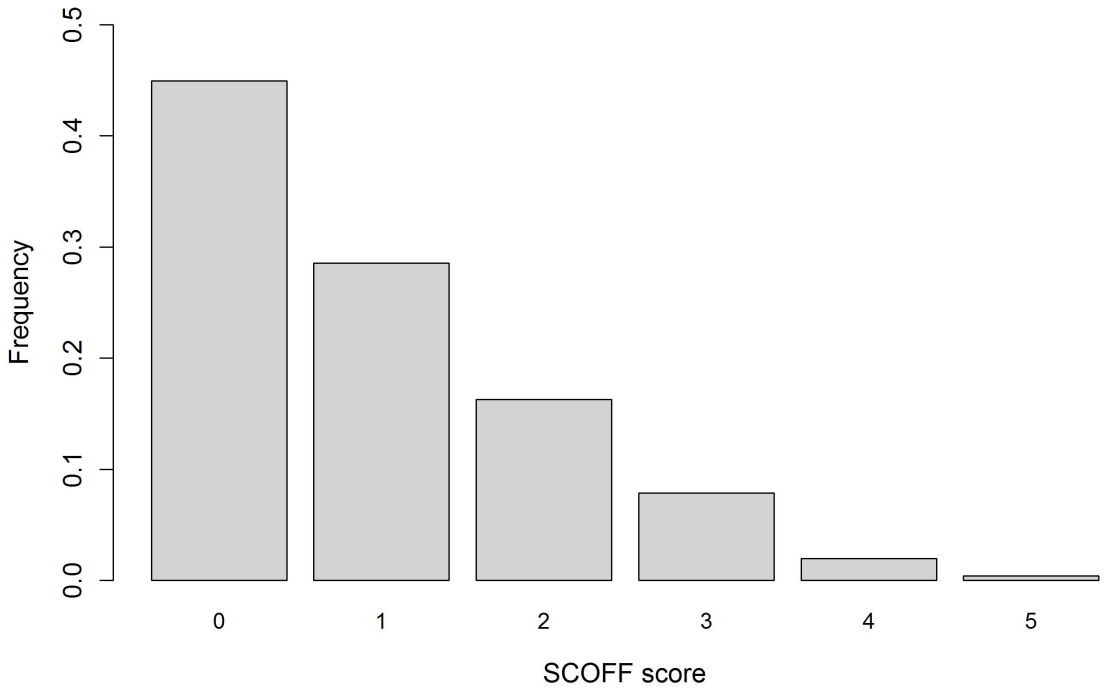


Figure 4.10 NEPS GRADE 9, EATING DISORDERS—barplot of SCOFF screening results

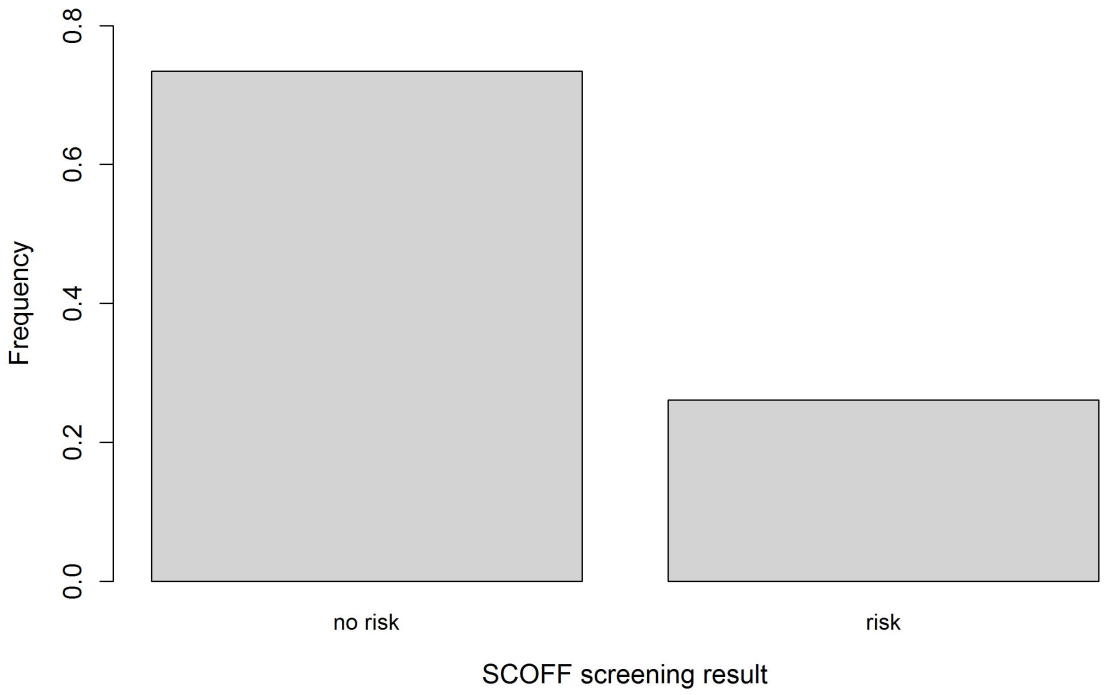


Figure 4.11 NEPS GRADE 9, EATING DISORDERS—trace plots and cumulative means for $\mathcal{M}_{10}$

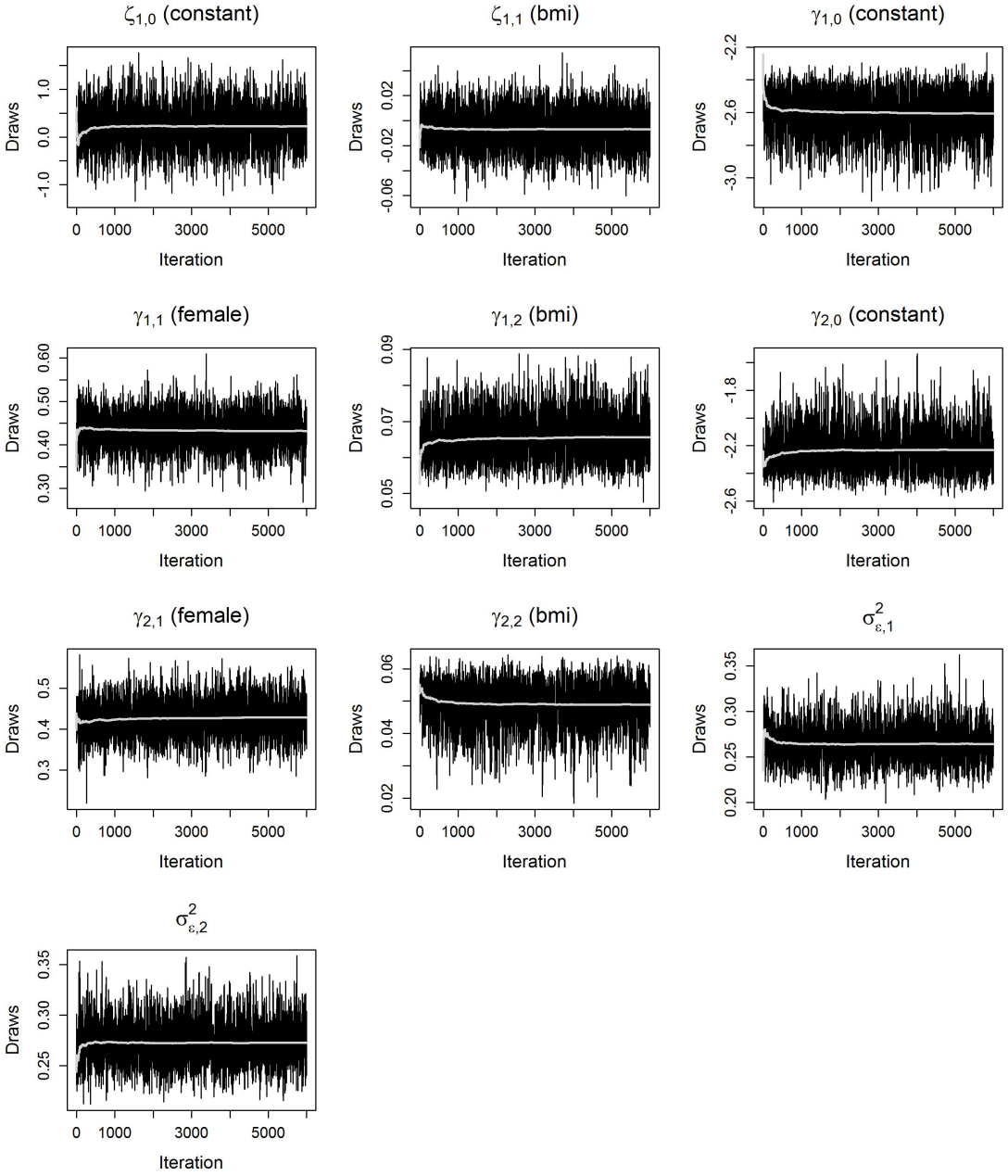
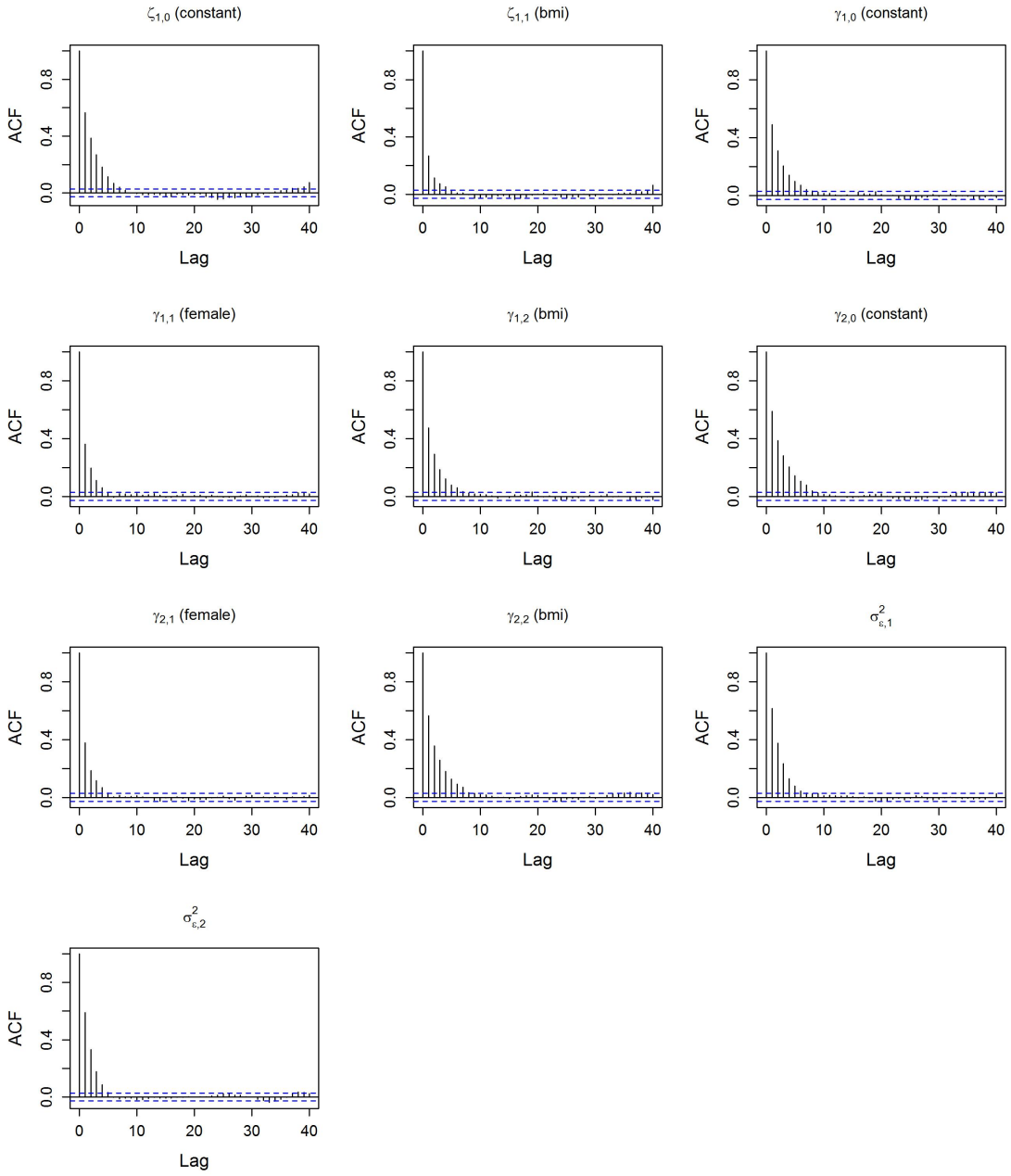


Figure 4.12 NEPS GRADE 9, EATING DISORDERS—post burn-in lag-1 autocorrelation functions for $\mathcal{M}_{10}$



C Program code

```
## ------------------------------------------------------------------------ ##
## ++ Chapter 4                                                             ##
## ------------------------------------------------------------------------ ##


## -- 4.3 Simulation studies                                                ##
## ------------------------------------------------------------------------ ##
# load required packages
library(foreach)
library(doMC)
library(mvnfast)
library(mvtnorm)
library(ucminf)
library(rpart)


sim1 <- function(
  seed
){

  # Run scenario 1.
  #
  # args
  # ----
  #   seed : set the random number seed for each simulation replication.
  #
  # returns
  # -------
  #   a three column matrix [1) estimation before deletion, 2) CART
  #   imputation and 3) stochastic regression imputation] of posterior
  #   means, sds and HDRs.
```

```
32
33    set.seed(seed)
34    N <- 2000
35    S <- c(rep(1, 1000), rep(2, 1000))
36    itermcmc <- 6000
37    burnin <- 1000
38    thin <- 2
39    data <- simdgp(N = N, S = S, scenario = "one")
40    Y <- data[, 1:20]
41    X <- data[, 21:22]
42    XMCAR <- X
43    pX1NA <- 0.1
44    pX2NA <- 0.1
45    misIndX <- cbind(rbinom(N, 1, prob = pX1NA), rbinom(N, 1, prob = pX2NA))
46    XMCAR[misIndX == 1] <- NA
47    est1 <- simmcmc(Y = Y, X = X, S = S,
48      itermcmc = itermcmc, burnin = burnin, thin = thin)
49    est2 <- simmcmc(Y = Y, X = XMCAR, S = S,
50      itermcmc = itermcmc, burnin = burnin, thin = thin)
51    est3 <- simmcmc(Y = Y, X = XMCAR, S = S,
52      itermcmc = itermcmc, burnin = burnin, thin = thin, regrimp = TRUE)
53    results <- cbind(est1, est2, est3)
54    return(results)
55
56  }
57
58  sim2 <- function(
59    seed
60  ){
61
62    # Run scenario 2.
63    #
64    # args
65    # ----
66    #   seed : set the random number seed for each simulation replication.
67    #
68    # returns
69    # -------
```

```
  #   a three column matrix [1) estimation before deletion, 2) CART
  #   imputation and 3) complete cases analysis] of posterior means, sds
  #   and HDRs.

  set.seed(seed)
  N <- 2000
  S <- c(rep(1, 1000), rep(2, 1000))
  itermcmc <- 6000
  burnin <- 1000
  thin <- 2
  data <- simdgp(N = N, S = S, scenario = "two")
  Y <- data[, 1:20]
  X <- data[, 21:23]
  Theta <- data[, 24]
  XMAR <- X
  pX1 <- pnorm(-1 - .5*Theta)
  pX2 <- pnorm(-1.7 - Theta)
  pX3 <- pnorm(-1.7 - Theta)
  misIndX <- cbind(rbinom(N, 1, prob = pX1), rbinom(N, 1, prob = pX2),
    rbinom(N, 1, prob = pX3))
  XMAR[misIndX == 1] <- NA
  YCC <- Y[complete.cases(XMAR), ]
  XCC <- XMAR[complete.cases(XMAR), ]
  SCC <- S[complete.cases(XMAR)]
  est1 <- simmcmc(Y = Y, X = X, S = S,
    itermcmc = itermcmc, burnin = burnin, thin = thin)
  est2 <- simmcmc(Y = Y, X = XMAR, S = S,
    itermcmc = itermcmc, burnin = burnin, thin = thin)
  est3 <- simmcmc(Y = YCC, X = XCC, S = SCC,
    itermcmc = itermcmc, burnin = burnin, thin = thin)
  results <- cbind(est1, est2, est3)
  return(results)

}

simdgp <- function(
  N,
  S,
```

```
  Gamma = matrix(c(-0.5, 0.2, 0.2, 1, 0.4, -0.2), nrow = 3, ncol = 2),
  Sigma2 = c(0.7^2, 0.5^2),
  INDITEMBIN = c(rep(T, 18), rep(F, 2)),
  Alpha = c(0.931, 0.883, 1.214, 0.989, 0.794, 0.896, 0.710, 1.003, 0.778,
    1.066, 1.017, 0.718, 1.014, 1.293, 0.840, 0.713, 0.770, 1.025, 0.792,
    1.155),
  Beta = c(-0.146, -0.158, -0.272, -0.451, -0.313, -0.542, -0.403, 0.791,
    -0.242, -0.068, -0.328, -0.720, 0.446, 0.782, -0.044, -0.416, 0.811,
    0.225, 0.025, -0.488),
  Kappa = list(c(0.5, 1), c(0.7, 1.4)),
  scenario = c("one", "two")
){

  # Data generating process scenarios 1 and 2 following the multigroup
  # latent regression model for two groups. Test data comprises J = 20
  # items (18 x binary and 2 x ordinal).
  #
  # args
  # ----
  #   N          : number of observations.
  #   S          : integer vector of individual group membership.
  #   Gamma      : matrix of group specific regression coefficients
  #                affecting latent abilities. Number of columns
  #                corresponds to number of groups.
  #   Sigma2     : vector of group specific error variances. Length
  #                corresponds to number of groups.
  #   INDITEMBIN : logical vector indicating which items are binary.
  #   Alpha      : vector of item discrimination parameters.
  #   Beta       : vector of item difficulty parameters.
  #   Kappa      : list with elements corresponding to item category cutoff
  #                parameters for ordinal items.
  #   scenario   : the simulation scenario to be used. scenario = "one"
  #                simulates two continuous person covariates and
  #                scenario = "two" simulates two continuous and one binary
  #                person covariates
  #
  # returns
  # -------
```

```
#   a data frame containing test data and covariates.

whichscenario <- match.arg(scenario)
if(whichscenario == "one"){
  SigmaX <- matrix(2, nrow = 2, ncol = 2)
  diag(SigmaX) <- rep(4, 2)
  X <- rmvn(N, rep(1, 2), SigmaX)
  X <- data.frame(X)
}else{
  SigmaX <- matrix(c(4, 2, 1, 2, 4, 1, 1, 1, 1), nrow = 3, ncol = 3)
  X <- rmvn(N, c(1, 1, 0), SigmaX)
  X[, 3] <- ifelse(X[, 3] > 0, 2, 1)
  X <- data.frame(X1 = X[, 1], X2 = X[, 2], X3 = factor(X[, 3]))
  Gamma <- matrix(c(-0.5, 0.2, 0.2, 0.3, 1, 0.4, -0.2, -0.5), nrow = 4,
    ncol = 2)
}
XDC <- model.matrix(~., X)
Theta <- numeric(length = N)
for(i in 1:N){
  Theta[i] <- XDC[i, ]%*%Gamma[, S[i]] + rnorm(1, 0,
    sqrt(Sigma2[S[i]]))
}
J <- length(INDITEMBIN)
Betar <- Beta - sum(Beta)/J
Kappar <- vector("list", J)
Kappar[INDITEMBIN] <- lapply(Kappar[INDITEMBIN], function(x){
  return(c(-1e+05, 0, 1e+05))
})
Kappar[!INDITEMBIN] <- lapply(Kappa, function(x){
  return(c(-1e+05, 0, x, 1e+05))
})
Alphar <- Alpha*(1/prod(Alpha))^(1/J)
Ylat <- matrix(0, nrow = N, ncol = J)
Y <- matrix(0, nrow = N, ncol = J)
for(j in 1:J){
  Ylat[, j] <- Alphar[j]*Theta - Betar[j] + rnorm(N, 0, 1)
  Y[, j] <- as.numeric(cut(Ylat[, j], breaks = Kappar[[j]])) - 1
}
```

```
  testdata <- cbind(Y, X, Theta)
  return(testdata)

}

simmcmc <- function(
  Y,
  X,
  S = NULL,
  itermcmc,
  burnin,
  thin,
  impute = c("CART", "regression")
){

  # Estimate multigroup latent regression model via data augmented
  # Metropolis-within-Gibbs sampler. Partially observed background
  # variables are imputed in each sampling iteration.
  #
  # args
  # ----
  #   Y        : a data frame containing test data.
  #   X        : a data frame containing person-level predictors for
  #              latent abilities.
  #   S        : integer vector of individual group membership.
  #   itermcmc : number of MCMC iterations.
  #   burnin   : number of burnin iterations.
  #   thin     : thinning interval (if argument is used, itermcmc*thin and
  #              burnin*thin yields total number of MCMC and burnin
  #              iterations).
  #   impute   : the method to be used for imputation of missing values in
  #              X. impute = "CART" executes sequential CART imputations
  #              and impute = "regression" executes sequential stochastic
  #              regression imputations.
  #
  # returns
  # -------
  #   vector of posterior means, sds and HDRs.
```

```
222
223   Y <- data.matrix(Y)
224   N <- nrow(Y)
225   J <- ncol(Y)
226   Qj <- apply(Y, 2, function(x){
227     length(unique(x[!is.na(x)]))
228   })
229   INDITEMBIN <- ifelse(Qj == 2, T, F)
230   WHICHITEMORD <- which(Qj != 2)
231   whichimpute <- match.arg(impute)
232   INDXNA <- any(is.na(X))
233   if(INDXNA){
234     # INITIALIZE MISSINGS
235     varmis <- colSums(is.na(X)) > 0
236     miscol <- varmis*c(1:ncol(X))
237     miscol <- miscol[miscol > 0]
238     colOrder <- order(colSums(is.na(X[, miscol, drop = FALSE])))
239     IndNA <- lapply(X[, miscol[colOrder], drop = FALSE], function(x){
240       which(is.na(x))
241     })
242     for(j in 1:length(IndNA)){
243       yvar <- names(IndNA[j])
244       indNA <- IndNA[[j]]
245       X[indNA, yvar] <- sample(X[-indNA, yvar], length(indNA),
246         replace = TRUE)
247     }
248   }
249   XDC <- model.matrix(~., X)
250   KX <- ncol(XDC)
251   XX <- crossprod(XDC)
252   if(is.null(S)){
253     S <- rep(1, N)
254   }
255   G <- length(unique(S))
256   Ng <- table(S)
257   INDG <- matrix(nrow = G, ncol = N)
258   for(g in 1:G){
259     INDG[g, ] <- ifelse(S == g, TRUE, FALSE)
```

```
}
YLAT <- matrix(0, nrow = N, ncol = J)
THETA <- rnorm(N)
GAMMA <- matrix(0, nrow = KX, ncol = G)
SIGMA2 <- rep(1, G)
ALPHA <- rep(1, J)
BETA <- rep(0, J)
XI <- cbind(ALPHA, BETA)
TAU <- lapply(Qj, function(x){
  if(x == 2){
    NULL
  }else{
    rep(0, x - 2)
  }
})
KAPPA <- lapply(Qj, function(x){
  if(x == 2){
    c(-1e+05, 0, 1e+05)
  }else{
    c(-1e+05, 0, cumsum(exp(rep(0, x - 2))), 1e+05)
  }
})
Gamma <- matrix(0, nrow = itermcmc, ncol = G*KX)
Sigma2 <- matrix(0, nrow = itermcmc, ncol = G)
Alpha <- matrix(0, nrow = itermcmc, ncol = J)
Beta <- matrix(0, nrow = itermcmc, ncol = J)
Kappa <- matrix(0, nrow = itermcmc, ncol = sum(Qj[!INDITEMBIN] - 2))
accTau <- rep(0, sum(!INDITEMBIN))
muGamma0 <- rep(0, KX)
covGamma0 <- 100*diag(KX)
precGamma0 <- solve(covGamma0)
shapeSigma20 <- 1
scaleSigma20 <- 1
scaleSigma20inv <- 1/scaleSigma20
shapeSigma2 <- Ng/2 + shapeSigma20
precXi0 <- solve(100*diag(2))
tdf <- 10
# MCMC
```

```
298  for (ii in 1:itermcmc) {
299  for (iii in 1:thin) {
300    # (1) LATENT VARIABLE
301    for (j in 1:J) {
302      mu <- ALPHA[j]*THETA - BETA[j]
303      FA <- pnorm(KAPPA[[j]][Y[, j] + 1] - mu)
304      FB <- pnorm(KAPPA[[j]][Y[, j] + 2] - mu)
305      YLAT[, j] <- mu + qnorm(runif(length(mu))*(FB - FA) + FA)
306    }
307    # (2) ITEM PARAMETERS
308    Xitem <- cbind(THETA, -1)
309    covitem <- solve(crossprod(Xitem) + precXi0)
310    for(j in 1:J){
311      mitem <- covitem%*%crossprod(Xitem, YLAT[, j])
312      XI[j, 1] <- 0
313      while(XI[j, 1] <= 0){
314        XI[j, ] <- rmvn(1, mitem, covitem)
315      }
316    }
317    BETA <- XI[, 2] - sum(XI[, 2])/J
318    ALPHA <- XI[, 1]*(1/prod(XI[, 1]))^(1/J)
319    # (3) TRANSFORMED ITEM CATEGORY CUTOFFS
320    for(j in WHICHITEMORD){
321      propmaxTau <- ucminf(par = TAU[[j]], fn = posttau, Y = Y[, j],
322        qj = Qj[j], alpha = ALPHA[j], beta = BETA[j], Theta = THETA,
323        hessian = 1)
324      prophatTau <- propmaxTau$par
325      propinvhessTau <- solve(propmaxTau$hessian)
326      TAUC <- rmvt(1, delta = prophatTau, sigma = propinvhessTau, df = tdf)
327      ratio <- min(1, exp(
328        -posttau(TAUC, Y[, j], Qj[j], ALPHA[j], BETA[j], THETA) +
329        posttau(TAU[[j]], Y[, j], Qj[j], ALPHA[j], BETA[j], THETA) -
330        dmvt(TAUC, delta = prophatTau, sigma = propinvhessTau, df = tdf,
331          log = T) +
332        dmvt(TAU[[j]], delta = prophatTau, sigma = propinvhessTau,
333          df = tdf, log = T)))
334      if(is.nan(ratio)){
335        cat(paste("TAU *Hessian* in itermcmc ", (ii - 1)*thin + iii, "\n",
```

```
336           sep = ""))
337         propinvhessTau <- propinvhessTauOld
338         TAUC <- rmvt(1, delta = prophatTau, sigma = propinvhessTau,
339           df = tdf)
340         ratio <- min(1, exp(
341           -posttau(TAUC, Y[, j], Qj[j], ALPHA[j], BETA[j], THETA) +
342           posttau(TAU[[j]], Y[, j], Qj[j], ALPHA[j], BETA[j], THETA) -
343           dmvt(TAUC, delta = prophatTau, sigma = propinvhessTau, df = tdf,
344             log = T) +
345           dmvt(TAU[[j]], delta = prophatTau, sigma = propinvhessTau,
346             df = tdf, log = T)))
347       }else{
348         propinvhessTauOld <- propinvhessTau
349       }
350       if(runif(1) < ratio){
351         accTau[which(WHICHITEMORD == j)] <- accTau[which(
352           WHICHITEMORD == j)] + 1
353         TAU[[j]] <- TAUC
354         KAPPA[[j]][3:Qj[j]] <- cumsum(exp(TAUC))
355       }
356     }
357     # (4) PERSON ABILITIES
358     for(i in 1:N){
359       vtheta <- 1/(crossprod(ALPHA) + 1/SIGMA2[S[i]])
360       mtheta <- vtheta*(crossprod(ALPHA, YLAT[i, ] + BETA) +
361         XDC[i, ]%*%GAMMA[, S[i]]/SIGMA2[S[i]])
362       THETA[i] <- rmvn(1, mtheta, vtheta)
363     }
364     # (5) FIXED EFFECTS and (6) POPULATION VARIANCE
365     for(g in 1:G){
366       covgamma <- solve(crossprod(XDC[INDG[g, ], , drop = FALSE])/
367         SIGMA2[g] + precGamma0)
368       mgamma <- covgamma%*%crossprod(XDC[INDG[g, ], , drop = FALSE],
369         THETA[INDG[g, ]])/SIGMA2[g]
370       GAMMA[, g] <- rmvn(1, mgamma, covgamma)
371       scaleSigma2 <- 0.5*crossprod(THETA[INDG[g, ]] -
372         XDC[INDG[g, ], , drop = F]%*%GAMMA[, g]) + scaleSigma20inv
373       SIGMA2[g] <- 1/rgamma(1, shape = shapeSigma2[g], rate = scaleSigma2)
```

```
374     }
375     # (7) IMPUTATION
376     if(INDXNA){
377       if(whichimpute == "regression"){
378         W1 <- cbind(XDC[, 3], THETA, S - 1)
379         LM1 <- lm(XDC[, 2] ~ W1)
380         B1 <- as.numeric(coef(LM1))
381         var1 <- sum(resid(LM1)^2)/LM1$df
382         XDC[IndNA[[1]], 2] <- rnorm(length(IndNA[[1]]),
383           cbind(1, W1[IndNA[[1]], ])%*%B1, sqrt(var1))
384         W2 <- cbind(XDC[, 2], THETA, S - 1)
385         LM2 <- lm(XDC[, 3] ~ W2)
386         B2 <- as.numeric(coef(LM2))
387         var2 <- sum(resid(LM2)^2)/LM2$df
388         XDC[IndNA[[2]], 3] <- rnorm(length(IndNA[[2]]),
389           cbind(1, W2[IndNA[[2]], ])%*%B2, sqrt(var2))
390         XX <- crossprod(XDC)
391       }else{
392         X <- seqcart(data.frame(X, THETA, S = factor(S)), IndNA)
393         X <- X[, -c(ncol(X) - 1, ncol(X)), drop = FALSE]
394         XDC <- model.matrix(~., X)
395         XX <- crossprod(XDC)
396       }
397     }
398   }
399     Gamma[ii, ] <- c(GAMMA)
400     Sigma2[ii, ] <- SIGMA2
401     Alpha[ii, ] <- ALPHA
402     Beta[ii, ] <- BETA
403     Kappa[ii, ] <- unlist(lapply(KAPPA[!INDITEMBIN], function(x){
404       return(x[-c(1, 2, length(x))])
405     }))
406   }
407   Draws <- cbind(Gamma, Sigma2, Alpha, Beta, Kappa)
408   Drawsbi <- Draws[-(1:burnin), ]
409   postm <- apply(Drawsbi, 2, mean)
410   postsd <- apply(Drawsbi, 2, sd)
411   posthdr <- apply(Drawsbi, 2, quantile, probs = c(.025, .975))
```

```
  out <- c(postm, postsd, posthdr)
  return(out)

}

posttau <- function(
  Tau,
  Yj,
  qj,
  alpha,
  beta,
  Theta
){

  # Log posterior of item category cutoff parameters.
  #
  # args
  # ----
  #   Tau   : vector of transformed item category cutoff parameters.
  #   Yj    : vector of item responses.
  #   qj    : number of item categories.
  #   alpha : item discrimination parameter.
  #   beta  : item difficulty parameter.
  #   Theta : vector of latent abilities.
  #
  # returns
  # -------
  #   log posterior value.

  Kappa <- c(-1e+05, 0, cumsum(exp(Tau)), 1e+05)
  ll <- sum(log(pnorm(alpha*Theta - (beta + Kappa[Yj + 1])) -
    pnorm(alpha*Theta - (beta + Kappa[Yj + 2]))))
  lprior <- dmvn(Tau, mu = rep(0, qj - 2), sigma = 100*diag(qj - 2),
    log = T)
  lpost <- lprior + ll
  return(-lpost)

}
```

```
450
451 seqcart <- function(
452   dataimp,
453   IndNA,
454   minCut = 5,
455   minDev = 0.0001
456 ){
457
458   # Single imputation via sequential CART.
459   #
460   # args
461   # ----
462   #   dataimp : a data frame including initial values for the partially
463   #             observed variables.
464   #   IndNA   : list with elements corresponding to vectors of missing
465   #             observations for every partially observed variable in
466   #             dataimp.
467   #   minCut  : minimum number of observations in any terminal tree node
468   #             during CART-imputation cycles.
469   #   minDev  : complexity parameter. Any split that does not decrease the
470   #             overall lack of fit by a factor of minDev is not attempted.
471   #
472   # returns
473   # -------
474   #   the imputed data frame.
475
476   for(j in 1:length(IndNA)){
477     yvar <- names(IndNA[j])
478     indNA <- IndNA[[j]]
479     yobs <- dataimp[-indNA, yvar]
480     xobs <- subset(dataimp, subset = !(1:nrow(dataimp) %in% indNA),
481       select = !(names(dataimp) %in% yvar))
482     xmis <- subset(dataimp, subset = (1:nrow(dataimp) %in% indNA),
483       select = !(names(dataimp) %in% yvar))
484     cartmethod <- ifelse(is.factor(yobs), "class", "anova")
485     treeimp <- rpart(yobs ~ ., data = cbind(yobs, xobs),
486       method = cartmethod, control = rpart.control(minbucket = minCut,
487       cp = minDev))
```

```
488     leafdonor <- floor(as.numeric(row.names(treeimp$frame[treeimp$where,
489       ])))
490     treeimp$frame$yval <- as.numeric(row.names(treeimp$frame))
491     leafmis <- predict(object = treeimp, newdata = xmis, type = "vector")
492     donor <- lapply(leafmis, function(x){
493       yobs[leafdonor == x]
494     })
495     imputes <- sapply(1:length(donor), function(x){
496       bb(donor[[x]])
497     })
498     dataimp[indNA, yvar] <- imputes
499   }
500   return(dataimp)
501
502 }
503
504 bb <- function(
505   donorpool
506 ){
507
508   # Performs bayesian bootstrap before sampling an observation.
509   #
510   # args
511   # ----
512   #   donorpool : sample.
513   #
514   # returns
515   # -------
516   #   the sampled observation.
517
518   di <- sort(runif(length(donorpool) - 1))
519   obs <- sample(x = donorpool, size = 1, prob = (c(di, 1) - c(0, di)))
520   return(obs)
521
522 }
```

D Computer software and hardware

All computations and graphics were done with R version 3.2.5 Patched ("Very, Very Secure Dishes") on an Intel(R) Core (TM) i7-860 processor (8M Cache, 2.80 GHz). R and all packages used are available from the CRAN at

`https://cran.r-project.org`.

Additionally, simulation studies 4.3 and examples 4.4 were run as shared memory parallel jobs on one 28-way Haswell-EP node at the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities Linux Cluster, where R was available in version 3.2.0 ("Full of Ingredients"). For more information, see

`https://www.lrz.de/services/compute/linux-cluster/`.