Statistics for Social and Behavioral Sciences

Bayesian Item Response Modeling

Theory and Applications

Springer

Statistics for Social and Behavioral Sciences

Advisors:

S.E. Fienberg
W.J. van der Linden

For other titles published in this series, go to
http://www.springer.com/series/3463

Jean-Paul Fox

Bayesian Item Response Modeling

Theory and Applications

Springer

Jean-Paul Fox
Department of Research Methodology,
Measurement, and Data Analysis
Faculty of Behavioral Sciences
University of Twente
7500 AE Enschede
The Netherlands

ISBN 978-1-4419-0741-7 e-ISBN 978-1-4419-0742-4
DOI 10.1007/978-1-4419-0742-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010927930

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Jasmijn and Kae

Preface

The modeling of item response data is governed by item response theory, also referred to as modern test theory. The field of inquiry of item response theory has become very large and shows the enormous progress that has been made. The mainstream literature is focused on frequentist statistical methods for estimating model parameters and evaluating model fit. However, the Bayesian methodology has shown great potential, particularly for making further improvements in the statistical modeling process.

The Bayesian approach has two important features that make it attractive for modeling item response data. First, it enables the possibility of incorporating nondata information beyond the observed responses into the analysis. The Bayesian methodology is also very clear about how additional information can be used. Second, the Bayesian approach comes with powerful simulation-based estimation methods. These methods make it possible to handle all kinds of priors and data-generating models.

One of my motives for writing this book is to give an introduction to the Bayesian methodology for modeling and analyzing item response data. A Bayesian counterpart is presented to the many popular item response theory books (e.g., Baker and Kim 2004; De Boeck and Wilson, 2004; Hambleton and Swaminathan, 1985; van der Linden and Hambleton, 1997) that are mainly or completely focused on frequentist methods. The usefulness of the Bayesian methodology is illustrated by discussing and applying a range of Bayesian item response models.

Complex Assessments

The recognition of the complexity of the processes leading to responses on test items stimulated the development of more realistic response models. The flexibility of the Bayesian modeling framework makes it particularly useful for making proper adjustments when the response data violate common model assumptions. Such violations might appear due to flaws in the data collection procedure or the complexity of the sample design.

Individuals' performances are best measured under controlled conditions such that other sources of variation are under control. An experimenter will choose a test for measuring a construct consisting of a set of items that minimizes individual variation and emphasize differences between subjects. Variability between respondents due to factors other than the construct under consideration is not desirable. Measurement models can become quite complex when they are adjusted in such a way that all sources of variation are taken into account. Flaws in experimental setups may engender the need for more complex measurement models. The impact of context effects on assessment results may further complicate the modeling process. Context effects appear when items function differently due to factors like item positioning or other material correlated with the item. Test data that violate assumptions of common item response models require a more flexible model that accounts for the violations.

Other complexities may arise due to the fact that besides response information other kinds of information are known (e.g., individual or group characteristics, response times, item characteristics). Different sampling designs can be used to sample respondents and/or items (e.g., adaptive item sampling, multistage sampling, randomized response sampling, simple random sampling, stratified sampling). Different response formats (e.g., multiple choice, binary, polytomous) and the presence of clusters of items may further stimulate the use of a more complex measurement model.

Besides introducing the Bayesian methodology, my aim has been to write a book to introduce Bayesian modeling of response data from complex assessments. Often information beyond the observed response data is available that can be used. The Bayesian modeling framework will prove to be very flexible, allowing simultaneous estimation of model parameters, computation of complex statistics, and simultaneous hypothesis testing.

In the 1990s, Bayesian inference became feasible with the introduction of Bayesian computational methods such as computer simulation and Monte Carlo techniques. The development of powerful computational simulation techniques induced a tremendous positive change in the applicability of Bayesian methodology. This led to the development of more flexible statistical models for test theory but also different strategies with respect to parameter estimation and hypothesis testing. In this book, the Bayesian way of item response modeling combined with the development of powerful numerical simulation techniques that led to a new research area in modern test theory is outlined.

Outside the Scope of This Book

Designing tests and testing whether tests are suited for the intended purpose are very complex subjects. Various questions need to be answered with respect to the response format of the test, the purpose of the test, and the construction of test materials, among others. The tests developed should also be reliable

and valid; that is, consistently result in scores that reflect the construct level of each respondent and measure what they are supposed to measure. Good tests are discriminating in the sense that they show differences in the construct level of respondents. There are a number of sources where this information is readily available. For classical test theory, see, for example, Gulliksen (1950), and Lord and Novick (1968), and for item response theory, see, for example, Lord and Novick (1968) and Lord (1980). A manual of standards for the construction and use of tests has been prepared by a joint committee of the American Educational Research Association, American Psychological Association and National Council of Measurement in Education (2000).

Overview

Statistical computations are necessary for applying the Bayesian methodology, and some programming skills are needed. That is, some familiarity with a statistical software package like R or S+ is needed to perform Bayesian analysis. On the one hand, this book aims to serve those who just want to apply the models, and they can use the software implemented in R packages and S+ programs (see Section 1.5). On the other hand, others may want to learn via programming and/or implement codes by themselves to extend models or adjust priors. For them, the mathematical details of the estimation procedures are discussed in the book, and the computer codes are provided via a website associated with the book. To understand the material, a basic background in probability and statistics is needed, including some familiarity with matrix algebra at the undergraduate level. The contents as well as the algorithms with their implementations make this book self-contained. Hopefully, it will provide an introduction to the essential features of Bayesian item response modeling as well as a better understanding of more advanced topics. The contents, programs, and codes will hopefully help readers implement their own algorithms and build their own set of tools for Bayesian item response modeling.

The book is organized as follows. In Chapter 1, the typical structure of item response data and the common item response models are discussed. Basic elements of Bayesian response modeling are introduced together with the basic building blocks for making Bayesian statistical inferences. WinBUGS is used to illustrate the Bayesian modeling approach. Chapter 2 presents a hierarchical modeling approach that supports the pooling of information, which becomes important when typically limited information is observed about many individuals. The Bayesian hierarchical modeling approach is outlined, which has tremendous potential with the current developments in statistical computing. Before discussing various sampling-based estimation methods for Bayesian item response models, which will be discussed in Chapter 4, in Chapter 3 a more general introduction is given to sampling-based estimation methods, testing hypotheses, and methods for model selection. Chapter 5 discusses methods for testing hypotheses and for model selection for the Bayesian item response models described in Chapter 4.

In Chapters 6–9, more advanced item response models are discussed for response data from complex assessments, response and response time data, and responses from complex sampling designs. In Chapter 6, respondents are assumed to be nested in groups (e.g., schools, countries). A hierarchical population model for respondents is defined to account for the within- and between-group dependencies. In Chapter 7, models for relaxing common measurement invariance restrictions are discussed. Chapter 8 introduces the multivariate analysis of responses and response times for measuring the speed and accuracy of working. Chapter 9 introduces models for (randomized) response data that are masked before they are observed to invite respondents to answer honestly when asked sensitive questions. Several empirical examples are presented to illustrate the methods and the usefulness of the Bayesian approach.

Acknowledgments

I would like to thank numerous people for their assistance and/or input during the writing of this book. I acknowledge the input from collaborators on earlier research projects that were addressed in the book. The cooperation of Rinke Klein Entink, Cees Glas, Wim van der Linden, Martijn de Jong, and Jan-Benedict Steenkamp has been greatly appreciated. I am indebted to Jim Albert, who provided very helpful comments on earlier drafts. Cheryl Wyrick has kindly provided data for a randomized response application. I thank my colleagues at the Department of Research Methodology, Measurement, and Data Analysis at the University of Twente. My colleagues Rinke Klein Entink and Josine Verhagen read drafts of the book, and their suggestions and comments led to its substantial improvement. I also thank John Kimmel for his confidence and assistance during the preparation of the book. The VIDI grant of the Netherlands Organization for Scientific Research (NWO) supported the writing of this book, for which I am most grateful.

Finally, I thank my wife, Miranda, for her support, encouragement, and patience during the writing of this book.

University of Twente, Enschede *Jean-Paul Fox*
March 2010

Contents

1

Introduction to Bayesian Response Modeling

1.1 Introduction

In modern society, tests are used extensively in schools, industry, and government. Test results can be of value in counseling, treatment, and selection of individuals. Tests can have a variety of functions, and often a broad classification is made in cognitive (tests as measures of ability) versus affective tests (tests designed to measure interest, attitudes, and other noncognitive aspects).

The urge for testing increased in different fields, and tests were used more and more for various purposes, like evaluating the efficiency of educational systems and students' learning progress, besides measuring persons and individual differences. Parallel to this development, an increasing public awareness of the importance, limitations, and impact of testing led to much criticism. Both stimulated the development of better tests and the improvement of statistical methods for analyzing test scores. Dealing with common test problems such as constructing tests and analyzing and interpreting results also encouraged the development of modern test theory or item response theory (IRT). In the second half of the twentieth century, item-based statistical models were used for the measurement of subjective states like intelligence, arithmetic ability, customer satisfaction, or neuroticism.

Originally, the item response models developed in the 1970s and 1980s were mainly meant for analyzing item responses collected under standardized conditions in real test situations. Assigning a score, determining its accuracy, and comparing the results were the main targets. Today, these item response models are widely applied for their well-known scaling and measurement properties, supported by various commercial and noncommercial software packages, where different types of response data require different item response models.

However, together with an increasing demand for testing, in the 1980s and 1990s new possibilities in computer technology made it possible to collect more data, improving the quality of data and the efficiency of data collection. The introduction of computer-assisted interviewing, Web-based surveys,

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_1,

and statistical data centers, among other methods, has made the collection of (response) data faster and more accurate. Nowadays, high-level data like background data, data extracted from public registries, and administrative data have become available via Web-enabled statistical databases. Typically, in survey-based studies, more and more data are available, while the amount of information is often limited at the individual level. Inferences are to be made at the individual level and other finer levels of aggregation, taking the level of uncertainty into account.

Parallel to the availability of high-quality data, new questions arose that were focused on addressing challenges such as complex response behavior (e.g., guessing, extreme responding), missingness and nonresponse, and complex sampling designs. Data from large-scale assessments are often hierarchically structured, where subjects are nested in groups, responses nested in subjects, or items nested within units. The nesting leads to more complicated dependency structures, with sources of variation at the different levels of hierarchy. The recognition of hierarchically structured data led to new challenges, like accurately measuring subject differences and cross-level relationships when accounting for nested sources of variation.

The increasing complexity of situations in which response data are collected also posed new issues. For example, cross-national response observations are difficult to interpret when the test characteristics are not invariant. Cross-national differences can possibly be explained by social background differences and measurement characteristic differences, but it is difficult to identify the real effects with a test that operates differently across countries. This problem gets more complicated when socially desirable answers to sensitive survey questions (e.g., about consumption of alcohol or use of illicit drugs) are obtained where respondents intentionally distort or edit their item responses. Tests are often used as surveys such that the performance on the test does not yield direct consequences for the respondent, with the effect that the amount of nonresponse increases significantly. In more complex survey studies, basic item response model assumptions are often violated and threaten the statistical inferences. One of the challenges is to account for respondent heterogeneity, cross-classified hierarchical structures, and uncertainty at different hierarchical levels while at the same time making accurate inferences at a disaggregate level.

To meet these challenges, a Bayesian approach to item response modeling was started in the 1980s by Mislevy (1986), Rigdon and Tsutakawa (1983), and Swaminathan and Gifford (1982, 1985), among others. The Bayesian modeling framework supports in a natural way extensions of common item response models. The response model parameters are described via prior models at separate levels to account for different sources of uncertainty, complex dependencies, and other sources of information. This flexibility in defining prior models for the item response model parameters is one of the strengths of Bayesian modeling that makes it possible to handle for example more complex sampling designs comprising complex dependency structures.

In the 1980s, the conceptual elegance of the Bayesian approach had been recognized, but major breakthroughs in computation were needed to make a Bayesian modeling approach possible and attractive. Improved computational methods were needed to support a novel flexible modeling approach that, among other things, acts upon the discrete nature of response data and handles relationships with higher-level data where standard distributional assumptions do not apply. This breakthrough was accomplished with the introduction of Markov chain Monte Carlo (MCMC) methods, which stimulated in a profound way a Bayesian item response modeling approach. Since the early 1990s, response modeling issues and problems of making inferences from response data have been attacked in a completely Bayesian way without computational obstacles. A key element was that the MCMC methods for simultaneous estimation remained straightforward as model complexity increased.

Specific problems related to the modeling of response data make certain Bayesian methods very useful. However, before discussing the attractiveness of Bayesian methods, typical characteristics of item response data and the use of latent variables are discussed.

1.1.1 Item Response Data Structures

Response data can be characterized in different ways, but a prominent feature is that they come from respondents. The heterogeneity between respondents is a typical source of variation in response data that needs to be accounted for in a statistical response model. Generally, differences between respondents are modeled via a probability distribution known as a respondents' population distribution, and inferences about respondents are always made with respect to a population distribution, which will receive special attention throughout this book.

Hierarchically Structured Data

In standard situations, respondents are assumed to be sampled independently from each other. This standard sampling design is simple random sampling with replacement from an infinite population. In many situations, respondents are clustered and the population of interest consists of subpopulations. The observations are correlated within clusters and reflect that the clusters differ in certain ways. The observations are said to be hierarchically structured when nested in clusters. Typically, response observations within each cluster are not independently distributed, in contrast to (nonnested) observations from different clusters.

There are various examples of clustered (response) data. Longitudinal data are hierarchically structured when subjects are measured repeatedly on the same outcome at several points in time. When the number of measurements and spacing of time points vary from subject to subject, the observations are viewed as nested within subjects. A slightly more general term is repeated

measurements, which refers to data on subjects measured repeatedly at different times or different conditions. The term clustered data, which characterizes the hierarchical structured nature of the data, is often used when observations are nested in geographical, political, or administrative units, or when respondents are nested under an interviewer or within schools. In educational research, response data are often doubly nested when observations are nested within individuals and are in turn nested within organizations. Multivariate data also contain a hierarchical structure since for each subject multiple outcomes are measured that are nested within the subject.

There are different terms used in the literature to characterize hierarchically structured data. The lowest level of the hierarchy is referred to as the level-1, stage-1, micro-, or observational level. One higher level of the hierarchy is referred to as level-2, stage-2, macro-, or cluster level. The terminology chosen is that most appropriate to the context, and in the absence of a particular context two levels of hierarchy are denoted by level 1 and level 2.

The heterogeneity between respondents is often of a complex nature, where respondents (level 2) are nested in groups (level 3; e.g., schools, countries) and responses (level 1) nested within individuals. Inferences have to be made at different levels of aggregation, and therefore a statistical model has to comprise the different levels of analysis. The responses at the observational or within-respondent level are explicitly modeled via a conditional likelihood where typically conditional independence is assumed given a person parameter. At a higher (hierarchical) level, a between-respondent model defines the heterogeneity between respondents. Later on, it will be shown that hierarchically structured response data can be analyzed in a unified treatment of all different levels of analysis via a Bayesian modeling approach.

Response data are often sparse at the respondent level but are linked to many respondents. This sparsity complicates an estimation procedure for obtaining reliable estimates of individual effects. By borrowing strength from the other individuals' response data nested in the same group, improved estimates of individual effects can be obtained. In the same way, more accurate estimates can be obtained at an aggregate level using the within-individual data.

Response data are often integer-valued, where responses can be obtained as correct or incorrect or are obtained on a five- or seven-point scale. The lumpy nature of response data requires a special modeling approach since the standard distributional assumptions do not apply.

Response data are often obtained in combination with other input variables. For example, response data are obtained from respondents together with school information, and the object is to make joint inferences about individual and school effects given an outcome variable. In a Bayesian framework, different sources of information can be handled efficiently, accounting for their level of uncertainty. It will be shown that the flexibility of a Bayesian modeling approach together with the powerful computational methods will offer an attractive set of tools for analyzing response data.

1.1.2 Latent Variables

Various definitions of latent variables are given in the literature. In this book, a latent variable is defined as a random variable whose realizations cannot be observed directly. It is obvious that a latent variable cannot be measured directly or even in principle when it represents a hypothetical construct like intelligence or motivation (e.g., Torgerson, 1958). An operational definition states that the construct is related to the observable data. The relationship is often defined in such a way that item responses serve as indicators for the measurement of the underlying construct. For example, common item response models define a mathematical relationship between a person's item responses and a latent variable that represents the property of the person that the items measure. In common situations, a latent variable appears as a continuous random variable. It is also possible that a latent variable is defined to be categorical such that respondents are assigned to one of a set of categories that may be ordered or unordered. Bartholomew and Knott (1999) and Skrondal and Rabe-Hesketh (2004), among others, give a general overview of latent variables and their uses in different social science applications.

For various reasons, latent variables play an important role in the statistical modeling of response data, especially in behavioral and social research. First, as mentioned, the item responses are often assumed to be indicators of an underlying construct or latent variable, and interest is focused on its measurement. IRT defines a relationship between item responses and respondents' latent variable values. Second, the direct specification of a joint distribution of the random observations is often extremely difficult, and some sort of summarization is needed to identify the interrelationships of the many random observations. Latent variables can be used to define an underlying structure to reduce the dimensionality of the data, and relationships can be specified for a smaller set of variables. Third, discrete response outcomes are often observed that can be regarded as a partial observation of an underlying continuous variable. For example, it is often assumed that realizations from a latent continuous variable are not observed but a censoring mechanism produces discrete responses on a fixed point scale. For binary responses, a positive response is observed when an underlying continuous variable surpasses a threshold value, and a negative response is observed otherwise. The latent continuous response formulation is very flexible and can handle almost all sorts of discrete responses, and it will be used extensively in subsequent chapters. Then, other advantages of the latent response formulation will be revealed.

In the following sections, some traditional item response models are reviewed from which extended Bayesian response models can be built. Then, a general Bayesian response modeling framework is introduced that is used throughout the rest of the book.

1.2 Traditional Item Response Models

The literature on the development, description, and applications of item response models for item-based tests is very rich and will not be repeated here. Only a short overview of some popular item response models will be given, including their assumptions. This introduction will also be used to introduce the notation. The classic book of Lord and Novick (1968) is often cited as the beginning of model-based statistical inference in educational and psychological measurement. However, the development of item response models has a longer history. A general and historical overview of item response theory can be found in Baker and Kim (2004), Bock (1997), Embretson and Reise (2000), and van der Linden and Hambleton (1997), among others. Item response models are sometimes introduced as an answer to shortcomings of classical test theory (e.g., Hambleton, Swaminathan and Rogers, 1991; Thissen and Wainer, 2001).

IRT is concerned with the measurement of a hypothetical construct that is latent and can only be measured indirectly via the measurement of other manifest variables. This hypothetical construct is a latent variable and often represents the ability, skill, or more generally a latent person characteristic that the items measure. Throughout the entire book, the latent variable will also be called an ability parameter as a generic name for the latent construct that is measured by the items and will usually be denoted as θ. When the latent variable refers to a person characteristic such as ability or proficiency, it will also be called a person parameter.

Item response models have several desirable features. Most of these features result from the fact that a common scale is defined for the latent variable. Item characteristic(s) and respondents' characteristic(s) are both separately parameterized within an item response model and are both invariant. This means that the corresponding estimates are not test-dependent. Latent variable estimates from different sets of items measuring the same underlying construct are comparable and differ only due to measurement error. Estimates of item characteristics from responses of different samples of individuals from the same population are comparable and differ only due to sampling error.

There are two key assumptions involved in IRT. The first assumption states that a change in the latent variable leading to a change in the probability of a specified response is completely described by the item characteristic curve (ICC), item characteristic function, or trace line. This ICC specifies how the probability of an item response changes due to changes in the latent variable. Different mathematical forms of the item characteristic curves lead to different item response models. For dichotomous responses (correct or in agreement), the probability of a success is modeled as a function of item and person parameters. The second assumption states that responses to a pair of items are statistically independent when the underlying latent variable (the items measure a unidimensional latent variable) is held constant. In that case, only one (unidimensional) latent variable influences the item responses and local

independence holds when the assumption of unidimensionality is true. The assumption of local independence is easily generalized to a multidimensional latent variable that states that responses to a pair of items are statistically independent when the multidimensional latent variable is held constant.

A random vector of K responses is denoted as $\mathbf{Y}_i$, with observed values $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iK})$ of an individual indexed i with ability parameter θ_i. Then the assumption of local independence can be stated as

$$P(\mathbf{y}_i \mid \theta_i) = P(y_{i1} \mid \theta_i)P(y_{i2} \mid \theta_i) \ldots P(y_{iK} \mid \theta_i) = \prod_{k=1}^{K} P(y_{ik} \mid \theta_i). \qquad (1.1)$$

There is one latent variable underlying the observed responses when local independence holds, and after conditioning on this latent variable the observed responses are assumed to be independent. Therefore, the assumption of local independence is also known as conditional independence.

There are two points of view on the meaning that (1.1) gives the conditional probability that person i with ability θ_i will produce response pattern $\mathbf{y}_i$ (Holland, 1990; Molenaar, 1995). In the stochastic subject view, it is assumed that subjects are stochastic in nature, which makes it meaningful to say that a person with ability θ_i has a probability of producing a correct response. The idea is that each person gives small response variations when confronting the respondent with the same item over and over again and brainwashing the person after each confrontation. Lord and Novick (1968) defined a so-called propensity distribution that describes similar variations in the total test scores in classical test theory. Holland (1990) mentioned that the stochastic subject view may suggest that there is no need to consider a population model for the respondents (examinee population), but the effect of the population will always be there (e.g., person and item parameters will always be estimated with respect to a population). This leads to the other point of view, which is based on the concept of sampling respondents from a population. In this so-called random sampling view, each probability on the right-hand side of (1.1) is the proportion of respondents with ability θ_i giving a correct response. This viewpoint makes the population of respondents part of the probability model for each response. The random sampling view for the meaning of the conditional probability of a correct response is adopted. Throughout this book, specific populations of respondents and items are included in the model since their effects cannot be ignored.

1.2.1 Binary Item Response Models

The Rasch Model

The Rasch model (Rasch, 1960), the one-parameter logistic response model, is one of the simplest and the most widely used item response model. In the

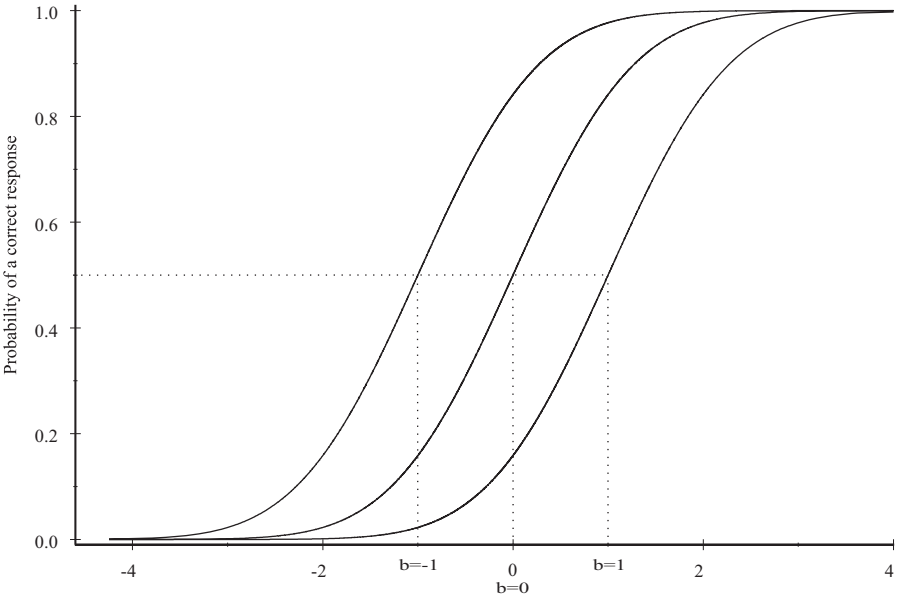

Fig. 1.1. Item characteristic curves of the one-parameter IRT model corresponding to three difficulty levels.

one-parameter response model, the probability of a correct response is given by

$$P(Y_{ik} = 1 \mid \theta_i, b_k) = \frac{\exp(\theta_i - b_k)}{1 + \exp(\theta_i - b_k)} = \big(1 + \exp(b_k - \theta_i)\big)^{-1} \qquad (1.2)$$

for individual i with ability level θ_i and item difficulty parameter b_k. In Figure 1.1, three ICCs corresponding to Equation (1.2) are plotted with different item difficulties. Each ICC describes the item-specific relationship between the ability level and the probability of a correct response. The difficulty parameter b_k is the point on the ability scale that corresponds to a probability of a correct response of $1/2$. An item is said to be easier when the probability of success is higher in comparison with another item given the same ability level. In Figure 1.1, the plotted ICCs from the left to the right have increasing item difficulty parameters. It can be seen that to maintain a probability of success of $1/2$ on each item one should increase its ability level from -1 to 1, starting from the left ICC to the rightmost ICC. An important feature of ICCs corresponding to the Rasch model is that the ICCs are parallel to one another. This means that for these items an increase in ability leads to the same increase in the probability of success. It is said that the items discriminate in the same way between success probabilities for related ability levels.

Rasch (1960) presented the dependent variable as the log odds or logit of passing an item, which equals the ability parameter minus the item difficulty

parameter. The Rasch model has some desirable features. The probability distribution is a member of the exponential family of distributions. As a result, the Rasch model shares the nice mathematical and statistical properties of exponential family models (see, e.g., Lehmann and Casella, 2003, pp. 23–32). The structure of the Rasch model allows algebraic separation of the ability and item parameters. Therefore, in the estimation of the item parameters, the ability parameters can be eliminated through the use of conditional maximum likelihood (CML) estimation. This can be achieved when the response space is partitioned according to the raw sum scores, which are sufficient statistics for the ability parameters. In the same way, the item scores are sufficient statistics for the item difficulties.

It can be seen from Equation (1.2) that a response probability can be increased by adding a constant to the ability parameter or subtracting this constant from the item difficulty parameter. Both parameters are defined in the same metric, and the metric is only defined up to a linear shift. This identification problem is solved by specifying the constraint in such a way that the location of the metric is known. This is usually done by adding the restriction that the sum of the difficulty parameters equals zero or by restricting the mean of the scale to zero.

A limitation of the Rasch model is that all items are assumed to discriminate between respondents in the same way and, as a result, items only differ in item difficulty. It is desirable from a practical point of view to parameterize item difficulties and item discriminations. Thissen (1982) developed an estimation procedure (marginal maximum likelihood, MML) for the one-parameter logistic model where all discrimination parameters are equal but not restricted to be one.

Two-Parameter Model

In the two-parameter logistic model, a discrimination parameter is added to the model, which leads to

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \frac{\exp(a_k\theta_i - b_k)}{1 + \exp(a_k\theta_i - b_k)} = \big(1 + \exp(b_k - a_k\theta_i)\big)^{-1}. \quad (1.3)$$

As a result, the item characteristic curve (ICC) has a slope parameter a_k and the items are no longer equally related to the ability parameter. In Figure 1.2, three ICCs for the two-parameter IRT model with the same difficulty parameter ($b_k = 0$) are plotted. The slope of each ICC is characterized by the discrimination parameter a_k.

The three ICCs have discrimination parameter values of 2, 1, and 1/2. The ICC with $a_k = 2$ has the steepest slope. The higher (lower) the discrimination parameter, the (less) better the item is capable of discriminating between low and high ability levels. Note that the item's discrimination value is strongly related to the item's difficulty value. An item of high discrimination is only useful in the area of the item's difficulty level that corresponds to a certain

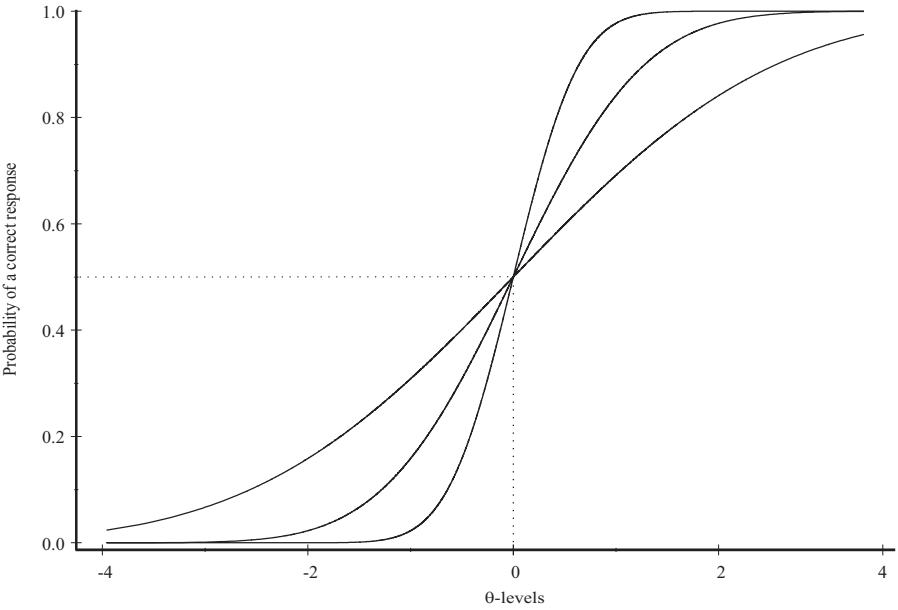

Fig. 1.2. Item characteristic curves of the two-parameter IRT model corresponding to three discrimination levels and an equal level of difficulty.

region of the ability scale. In Figure 1.2, it can be seen that the item with the steepest slope is useful in the region between -1 and 1 of the ability scale, whereas the item with the flattest ICC is useful in the region between -2 and 2.

There is no sufficient statistic for the ability parameters, and as a result conditional maximum likelihood estimation is not possible. Bock and Lieberman (1970) and Bock and Aitkin (1981) developed an estimation procedure based on MML for the two-parameter model. The item parameters are estimated from the marginal distribution by first integrating over the ability distribution and thus removing the ability parameters from the likelihood function.

A probit version of the two-parameter model is defined in the literature as the normal ogive model (e.g., Lord and Novick, 1968, pp. 365–384) in which the ICC is based on a cumulative normal distribution,

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k) = \int_{-\infty}^{a_k\theta_i - b_k} \phi(z)dz, \qquad (1.4)$$

where $\Phi(.)$ and $\phi(.)$ are the cumulative normal distribution function and the normal density function,[1] respectively. The logistic ICC and the normal ogive

[1] Random variable Z is normally distributed with mean μ and variance σ^2 when its probability density function equals $\phi(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(z-\mu)^2\right)$. The standard normal density function is defined by $\mu = 0$ and $\sigma = 1$.

ICC in Equations (1.3) and (1.4) closely resemble each other when the logistic item parameter values are multiplied with a constant scaling factor $d = 1.7$. Then, for different values of the ability parameter, the response probabilities of the two-parameter logistic and the normal ogive differ in absolute value by less than .01 (Hambleton et al., 1991, p. 15). The item parameters will also be denoted by $\boldsymbol{\xi}_k$, with $\boldsymbol{\xi}_k = (a_k, b_k)^t$.

The term $a_k\theta_i - b_k$ in Equations (1.3) and (1.4) is often presented as $a_k\,(\theta_i - b_k^*)$. The b_k^* are defined on the same scale as the latent variable. That is, as in the Rasch model, the b_k^* is the point on the ability scale where an examinee has a probability of success on the item of 1/2. The reparameterization $b_k = a_k \cdot b_k^*$ relates both parameters with each other. In subsequent chapters, it is shown that the term $a_k\theta_i - b_k$ (without parentheses) will be useful and the (estimated) difficulty parameters are easily transformed to another scale. In Figure 1.2, the difficulty levels of the items are zero, and in that case both parameterizations lead to the same difficulty level. The metric of the ability parameters is known from item response data only up to a linear transformation. The metric can be identified by fixing a discrimination and difficulty parameter or by adding constraints that the sum of item difficulties and the product of item parameter values equals, for instance, zero and one, respectively, or by fixing the mean and variance of the population distribution of ability parameters. Note that the choice of the identifying restrictions can lead to numerical problems in the estimation of the parameters.

Three-Parameter Model

The two-parameter normal ogive model can be extended to allow for guessing by introducing a nonzero lower asymptote for the ICC; that is,

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k, c_k) = c_k + (1 - c_k)\Phi(a_k\theta_i - b_k) \tag{1.5}$$

$$= \Phi(a_k\theta_i - b_k) + c_k\left(1 - \Phi(a_k\theta_i - b_k)\right), \tag{1.6}$$

where c_k is known as the guessing parameter of item k. The probability of a correct response is given by a guessing parameter plus a second term representing the probability of a correct response depending on item parameter values and the ability level of respondent i. The logistic version becomes

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k, c_k) = c_k + \frac{1 - c_k}{1 + \exp(b_k - a_k\theta_i)} \tag{1.7}$$

$$= \frac{1}{1 + \exp(b_k - a_k\theta_i)} + \frac{c_k}{1 + \exp(a_k\theta_i - b_k)}.$$

The item parameters of both models differ by a constant scaling factor (see also Section 4.3.2). When $c_k = 0$, the three-parameter model resembles the two-parameter model. For $c_k > 0$, the interpretation of b_k is changed. In the three-parameter model, the proportion responding correctly at b_k/a_k equals $1/2 + c_k$, and in the two-parameter model b_k/a_k is the value of θ_i at which a respondent has a probability of 1/2 of responding correctly.

In Figure 1.3, three ICCs of the three-parameter model are plotted with the same discrimination and difficulty level but with three different levels of guessing, low (.05), medium (.10), and high (.20). The height of the lower asymptote is the guessing level of the item and corresponds to the probability of success when guessing the response. It can be seen that for high-ability respondents the effect of guessing on the success probability is very small since the three ICCs are almost identical at the higher end of the ability scale.

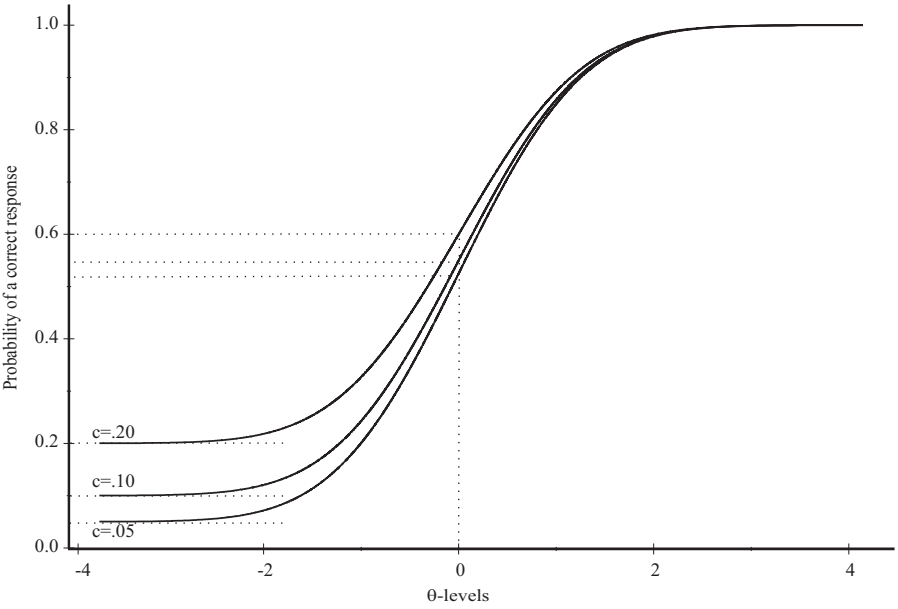

Fig. 1.3. Item characteristic curves of the three-parameter IRT model corresponding to three different levels of guessing and an equal level of discrimination and difficulty.

1.2.2 Polytomous Item Response Models

Measurement items are often presented with multiple categories: rating scale items such as Likert-type items, multiple-choice items where each response category is scored separately, and items that assign partial credit for partially correct answers, among others. Most polytomous models are based on ordered polytomous items, which are those items where the response categories can be ordered with respect to the ability parameter. Responses to ordered polytomous items are also referred to as graded responses. Although polytomous item response models contain more item parameters, more precise information about the ability level can be obtained when more than two scoring categories

are used. The measurement information will be reduced when dichotomizing polytomous response data. Cohen (1983) showed an increase in statistical information from polytomous IRT models in comparison with dichotomous item response models. A general overview and a historical discussion of polytomous item response models can be found in Ostini and Nering (2006) and Embretson and Reise (2000).

In this section, two commonly used polytomous item response models are presented for ordinal response data. The partial credit model (PCM; Masters, 1982) was developed for test items. It requires multiple steps, and partial credit is assigned for completing each step. The probability of a response in a particular category c ($c = 1, \ldots, C_k$) of item k is defined directly as

$$P(Y_{ik} = c \mid \theta_i, \boldsymbol{\kappa}_k) = \frac{\exp \sum_{l=1}^{c} (\theta_i - \kappa_{k,l})}{\sum_{r=1}^{C_k} \left(\exp \sum_{l=1}^{r} (\theta_i - \kappa_{k,l})\right)},$$

where $\kappa_{k,l}$ is the item step difficulty parameter and $\sum_{l=1}^{1} (\theta_i - \kappa_{k,l}) \equiv 0$. The number of categories per item may differ. The PCM model simplifies to the Rasch model for an item with only two categories. The item parameters are not subject to an order constraint since each item parameter is defined locally with respect to two adjacent categories instead of taking into account all categories simultaneously. Muraki (1992, 1993) developed the generalized partial credit model that allows the items to have different slope parameters.

In the PCM, the cumulative probabilities are not modeled directly but are the result of summing the category response functions. In the graded response model (Samejima, 1997), the cumulative probabilities are modeled directly. The probability of scoring in a specific category is modeled by the probability of responding in (or above) this category minus the probability of responding in (or above) the next category. Let C_k denote the number of response categories of item k. Then there are $C_k - 1$ thresholds between the response options. The graded response model has the mathematical representation

$$\begin{aligned} P(Y_{ik} = c \mid \theta_i, \boldsymbol{\kappa}_k) &= P(Y_{ik} \geq c - 1 \mid \theta_i, \boldsymbol{\kappa}_k) - P(Y_{ik} \geq c \mid \theta_i, \boldsymbol{\kappa}_k) \\ &= \int_{\kappa_{k,c-1}}^{\infty} \psi\left(z; a_k\theta_i\right) \mathrm{d}z - \int_{\kappa_{k,c}}^{\infty} \psi\left(z; a_k\theta_i\right) \mathrm{d}z \\ &= \Psi\left(a_k\theta_i - \kappa_{k,c-1}\right) - \Psi\left(a_k\theta_i - \kappa_{k,c}\right) \\ &= \frac{\exp(a_k\theta_i - \kappa_{k,c-1})}{1 + \exp(a_k\theta_i - \kappa_{k,c-1})} - \frac{\exp(a_k\theta_i - \kappa_{k,c})}{1 + \exp(a_k\theta_i - \kappa_{k,c})}, \end{aligned} \tag{1.8}$$

where ψ and Ψ are the logistic density[2] and logistic cumulative distribution function, respectively. The probability of scoring in or above the lowest category is one and the probability of scoring above the highest category is zero.

[2] Random variable Z is logistically distributed with mean μ and variance $\sigma^2\pi^2/3$ when its probability density function equals $\psi(z; \mu, \sigma^2) = \frac{\exp((z-\mu)/\sigma)}{\sigma(1+\exp((z-\mu)/\sigma))^2}$. The standard logistic density function is defined by $\mu = 0$ and $\sigma = 1$.

Note that $\kappa_{k,c}$ is the upper grade threshold parameter for category c. The ordering of the response categories is displayed as $-\infty = \kappa_{k,0} < \kappa_{k,1} \leq \kappa_{k,2}, \ldots, < \kappa_{k,C_k} = \infty$, where there are C_k categories.

The graded response model can also be written in cumulative normal response probabilities; that is,

$$\begin{aligned} P(Y_{ik} = c \mid \theta_i, \boldsymbol{\kappa}_k) &= \int_{\kappa_{k,c-1}}^{\kappa_{k,c}} \phi\left(z; a_k\theta_i\right) \mathrm{d}z \\ &= \Phi\left(\kappa_{k,c} - a_k\theta_i\right) - \Phi\left(\kappa_{k,c-1} - a_k\theta_i\right), \end{aligned}$$

which is the normal ogive version of the graded response model. Note that this formulation is comparable to the one in Equation (1.8) since the logistic as well as the normal distribution is symmetric. The graded response model has an order restriction on the threshold parameters in comparison with the generalized partial credit model. However, the graded response model has an underlying continuous response formulation that will prove to be very useful for estimating and testing parameters. For example, in Chapter 7, the underlying response formulation will be utilized for a more complex situation where measurement characteristics are allowed to vary across nations.

The polytomous models are identified by fixing the scale of the latent ability parameter. This can be done by fixing a threshold parameter and in the case of the generalized partial credit model and the graded response model a discrimination parameter or by fixing the product of discrimination parameters. In Section 4.4, the identification issues are discussed in more detail.

1.2.3 Multidimensional Item Response Models

Some test items require multiple abilities to obtain a correct response. That is, more than one ability is measured by these items. The most common example is a mathematical test item presented as a story that requires both mathematical and verbal abilities to arrive at a correct score. Several assumptions can be made. First, the probability of obtaining a correct response to a test item is nondecreasing when increasing the level of the multiple abilities being measured. This relates to the monotonicity assumption for unidimensional item response models. Second, individual item responses are conditionally independent given the individual's ability values, which is the assumption of local independence. On the basis of these assumptions, the basic form of a multidimensional item response model for binary response data is a direct generalization of the unidimensional item response model. In this generalization, each respondent is described by multiple person parameters rather than a single scalar parameter, where the person parameters represent the multiple abilities that are measured.

This extension to multiple dimensions of the logistic unidimensional two-parameter model has the mathematical representation

$$P(Y_{ik} = 1 \mid \boldsymbol{\theta}_i, \mathbf{a}_k, b_k) = \frac{\exp\left(\sum_q a_{kq}\theta_{iq} - b_k\right)}{1 + \exp\left(\sum_q a_{kq}\theta_{iq} - b_k\right)} = \frac{\exp\left(\mathbf{a}_k^t\boldsymbol{\theta}_i - b_k\right)}{1 + \exp\left(\mathbf{a}_k^t\boldsymbol{\theta}_i - b_k\right)},$$

where respondent i has a vector of ability parameters $\boldsymbol{\theta}_i$ with elements $\theta_{i1}, \ldots, \theta_{iQ}$. The elements of the discrimination matrix for item k, $\mathbf{a}_k$, can be interpreted as the discriminating power of the item. The discriminating level, a_{kq}, reflects the change in the probability of a correct response due to a change in the corresponding ability level θ_{iq}. The dimensionality of the ability parameter can be increased to improve the fit of the model (exploratory) or to support theoretical relationships between the items and the dimensions (confirmatory).

Multidimensional item response models for binary response data were first explored by Lord (1980) and McDonald (1967). Béguin and Glas (2001), and Reckase (1985, 1997), among others, have further explored the utility of multidimensional item response models.

1.3 The Bayesian Approach

In the Bayesian approach, model parameters are random variables and have prior distributions that reflect the uncertainty about the true values of the parameters before observing the data. The item response models discussed for the observed data describe the data-generating process as a function of unknown parameters and are referred to as likelihood models. This is the part of the model that presents the density of the data conditional on the model parameters. Therefore, two modeling stages can be recognized: (1) the specification of a prior and (2) the specification of a likelihood model. After observing the data, the prior information is combined with the information from the data and a posterior distribution is constructed. Bayesian inferences are made conditional on the data, and inferences about parameters can be made directly from their posterior densities.

The Role of Prior Information

Prior distributions of unknown model parameters are specified in such a way that they capture our beliefs about the situation before seeing the data. The Bayesian way of thinking is straightforward and simple. All kinds of information are assessed in probability distributions. Background information or context information is summarized in a prior distribution, and specific information via observed data is modeled in a conditional probability distribution.

Objection to a Bayesian way of statistical inference is often based upon the selection of a prior distribution that is regarded as being arbitrary and subjective (see Gelman, 2008). The specification of a prior is subjective since

it presents the researcher's thought or ideas about the prior information that is available. In this context, the prior that captures the prior beliefs is the only correct prior. The prior choice can be disputable but is not arbitrary because it represents the researcher's thought. In this light, other non-Bayesian statistical methods are arbitrary since they are equally good and there is no formal principle for choosing between them. Prior information can also be based on observed data or relevant new information, or represent the opinion of an expert, which will result in less objection to the subjective prior. It is also possible to specify an objective prior that reflects complete ignorance about possible parameter values. Objective Bayesian methodology is based upon objective priors that can be used automatically and do not need subjective input.

Incorporating prior information may improve the reliability of the statistical inferences. The responses are obtained in a real setting, and sources of information outside the data can be incorporated via a prior model. In such situations where there is little data-based information, prior information can improve the statistical inferences substantially. In high-dimensional problems, priors can impose an additional structure in the high-dimensional parameter spaces. Typically, hierarchical models are suitable for imposing priors that incorporate a structure related to a specific model requirement. By imposing a structure via priors, the computational burden is often reduced.

1.3.1 Bayes' Theorem

Response data can be obtained via some statistical experiment where each event or occurrence has a random or uncertain outcome. Let N observations be denoted as $\mathbf{y} = (y_1, \ldots, y_N)$, and assume that $\mathbf{y}$ is a numerical realization of the random vector $\mathbf{Y} = (Y_1, \ldots, Y_N)$. The random vector $\mathbf{Y}$ has some probability distribution. For simplicity, $\mathbf{Y}$ is a continuous or discrete random vector with probability function $p(\mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}$. This notation is slightly sloppy since a continuous random variable has a probability density function (pdf) and a discrete random variable a probability mass function (pmf). For simplicity, the addition density or mass is often dropped. Formally, probability distributions can be characterized by a probability density function, but the terms distribution and density will be used interchangeably when not leading to confusion.

Assume that response data are used to measure a latent variable $\boldsymbol{\theta}$ that represents person characteristics. The expression $p(\boldsymbol{\theta})$ represents the information that is available a priori without knowledge of the response data. This term $p(\boldsymbol{\theta})$ is called the prior distribution or simply the prior. It will often indicate a population distribution of latent person characteristics that are under study. Then, it provides information about the population from which respondents for whom response data are available were randomly selected.

The term $p(\mathbf{y} \mid \boldsymbol{\theta})$ represents the information about $\boldsymbol{\theta}$ from the observed response data. Considered as a function of the data, this is called the sampling

distribution of the data, and considered as a function of the parameters, it is called the likelihood function. Interest is focused on the distribution of the parameters $\boldsymbol{\theta}$ given the observed data. This conditional distribution of $\boldsymbol{\theta}$ given the response data is

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{y}) \tag{1.9}$$
$$\propto p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{1.10}$$

where $\propto$ denotes proportionality. The term $p(\boldsymbol{\theta} \mid \mathbf{y})$ is the posterior density of the parameter $\boldsymbol{\theta}$ given prior beliefs and sample information. It provides probability beliefs about the parameters from prior and response data information. The denominator in (1.9) is called the marginal density of the data, the marginal likelihood, or the integrated likelihood, and evaluating this expression is often a costly operation in computation time. When it suffices to know the shape of the posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$, the unnormalized density function can be used as in (1.10).

Equation (1.9) represents a mathematical result in probability theory and is known as a statement of Bayes' theorem (Bayes, 1763). The factorization in (1.10) is a product of the likelihood, $l(\mathbf{y};\boldsymbol{\theta})$, and prior since usually $l(\mathbf{y};\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})$. This likelihood function contains all sample information regarding $\boldsymbol{\theta}$. The likelihood principle states that two samples contain the same information about $\boldsymbol{\theta}$ when the likelihoods are proportional (Casella and Berger, 2002). Bayesian inference adheres to the likelihood principle since all inferences are based on the posterior density and the posterior depends on the data only via the likelihood.

The joint posterior density $p(\mathbf{y}, \boldsymbol{\theta})$ can be factorized as

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}) &= p(\boldsymbol{\theta} \mid \mathbf{y})p(\mathbf{y}) \\ &= p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}). \end{aligned}$$

Thus, the joint posterior density can be factorized as the marginal density of the data and the posterior of $\boldsymbol{\theta}$, but also as the prior of $\boldsymbol{\theta}$ and the likelihood of $\boldsymbol{\theta}$ given $\mathbf{y}$. The joint posterior density $p(\mathbf{y}, \boldsymbol{\theta})$ is also known as the unnormalized posterior density function, which leads to the (normalized) posterior of $\boldsymbol{\theta}$ when divided by $p(\mathbf{y})$.

The posterior density of the parameters, $p(\boldsymbol{\theta} \mid \mathbf{y})$, is used for making inferences. Bayesian computational methods make it possible to make inferences without having to rely on asymptotic approximations. Response data are typically nonnormally distributed and together with small amounts of sample information per parameter, particularly at the within-individual level, it is precarious to rely on asymptotic approximations without showing them to be accurate. Fully Bayesian methods provide a way to improve the precision of the parameter estimates. The prior contributes additional information, and the posterior estimate is based on the combined sources of information (likelihood and prior), which leads to greater precision. The influence of prior information on the posterior estimates is illustrated in Section 1.4.1.

Constructing the Posterior

As an illustration, assume that five dichotomous responses $\mathbf{y} = (1, 1, 0, 0, 0)^t$ were observed from a respondent with ability θ. The object is to estimate the posterior density of the ability parameter. Assume that all items are of equal difficulty, say zero. According to the probit version of the Rasch model, let $P(Y_k = 1 \mid \theta) = \Phi(\theta)$ define the probability of a correct response to item k.

It is believed priori that the respondent has a nonzero probability of giving a correct answer and a nonzero probability of giving an incorrect answer. Therefore, let θ be a priori uniformly distributed on the interval $[-3, 3]$ such that $.001 < \Phi(\theta) < .998$.

The likelihood function for θ equals

$$p(\mathbf{y} \mid \theta) = \Phi(\theta)^2 (1 - \Phi(\theta))^3 .$$

Multiplying the likelihood with the prior as in Equation (1.10), the posterior density of θ is

$$p(\theta \mid \mathbf{y}) \propto \Phi(\theta)^2 (1 - \Phi(\theta))^3$$

for $\theta \in [-3, 3]$. The posterior mode, at which the posterior density is maximized, can be computed by taking the first derivative of the logarithm of the posterior, setting the expression equal to zero, and solving the equation for θ. It follows that the posterior mode equals $\theta_m = \Phi^{-1}(2/5) \approx -.25$.

Updating the Posterior

Bayes' theorem can be seen as an updating rule where observed data are used to translate the prior views into posterior beliefs. Assume that the posterior density of θ is based on K item observations. The posterior can be expressed as the product of likelihood times the prior. The response observations are conditionally independent given θ, and it follows that the posterior can be expressed as

$$\begin{aligned} p(\theta \mid y_1, y_2, \ldots, y_K) &\propto p(y_1 \mid \theta)\, p(y_2 \mid \theta) \ldots p(y_K \mid \theta)\, p(\theta) \\ &\propto p(\theta \mid y_1, y_2, \ldots, y_{K-1})\, p(y_K \mid \theta) . \end{aligned}$$

The posterior density given all but the last observation is updated via the likelihood of the last observation.

To illustrate the updating nature of Bayes' theorem, consider 26 responses to the Mini-Mental State Examination (MMSE) for measuring cognitive impairment (the MMSE data will be described in Section 6.6.4). The object is to update the posterior density of a respondent's cognitive impairment, based on previous knowledge, using the subsequent response observation. It is assumed that in the population from which respondents are independently sampled the levels of cognitive impairment are normally distributed, where a high (low)

θ value corresponds to mild (severe) cognitive impairment. A two-parameter item response model defines the probability of a correct response given the level of impairment, and the item parameters are assumed to be known. The complete response pattern of a person consists of six incorrect responses (items 4, 12, 16, 17, 18, and 23).

The updated posterior densities are plotted in Figure 1.4. Without any item observations, the standard normal prior reflects the a priori information about the cognitive impairment (dotted line). The updated posterior densities based on two (with symbol 2) and three (with symbol 3) items are shifted to the right. The person's cognitive impairment is less than expected a priori since the items were answered correctly. A shift in the posterior means can be detected, but the shapes of the posteriors are quite similar in correspondence to the prior density. The fourth item was answered incorrectly. As a result, the updated posterior (with symbol 4) is shifted to the left and the posterior mean is negative. The posterior expectation about the person's cognitive impairment has changed dramatically due to an incorrect answer. Item five is answered correctly, and the updated posterior shifts to the right. It can be seen that when more than five item observations become available, the posterior densities only become tighter, concentrated around the posterior mean.

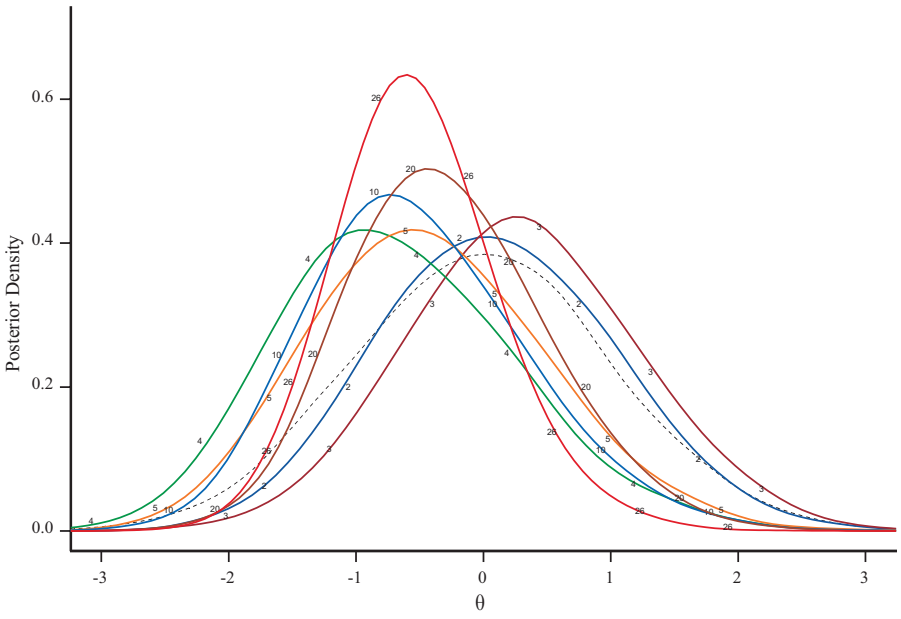

Fig. 1.4. Updated posterior densities of a person's cognitive impairment for 2–26 item observations.

1.3.2 Posterior Inference

In item response modeling, the person and item parameters are often of interest, and the objective of inferences is their posterior distributions. The posterior information is most often summarized by reporting the posterior mean and standard deviation.

Besides the prior density $p(\boldsymbol{\theta})$ for the person parameters $\boldsymbol{\theta}$, let the item characteristics be parameterized by $\boldsymbol{\xi}$ and let $p(\boldsymbol{\xi})$ represent the prior beliefs. The item characteristic parameters have an important role in response modeling, and their prior density will receive special attention in this book. According to Bayes' theorem, the joint posterior density of the parameters of interest can be stated as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{y}) &= p\left(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi}\right) p(\boldsymbol{\theta}, \boldsymbol{\xi}) / p(\mathbf{y}) \\ &= p\left(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi}\right) p(\boldsymbol{\theta}) p(\boldsymbol{\xi}) / p(\mathbf{y}), \end{aligned}$$

where the prior densities are assumed to be independent from each other. Summarizing the complicated high-dimensional joint posterior density is very difficult since the posterior density has analytically intractable forms.

As a first step, when interest is focused on $\boldsymbol{\theta}$, the item parameters need to be marginalized out in the posterior density of interest. In Bayesian inference, the nuisance parameters are eliminated while accounting for their uncertainty simply by integrating the joint distribution over them. It follows that

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}) &= \int p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\theta}) p(\boldsymbol{\xi}) / p(\mathbf{y}) \, \mathrm{d}\boldsymbol{\xi} \\ &= \int p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\xi}. \end{aligned} \tag{1.11}$$

From this point on, the range of integration will often be omitted from the expressions, as it will be specified implicitly by the differentials. Equation (1.11) shows that the marginal posterior of interest is obtained by integrating out the item parameters. In the same way, the marginal posterior of the item parameters is obtained by integrating out the person parameters. More integrals need to be evaluated when the marginal posterior of a single component of, for example, the vector of person parameters is required.

Summarizing the marginal posteriors remains difficult since the mathematical forms are not known. Simulation-based methods will be shown to be capable of generating samples from the marginal posteriors. Subsequently, the samples are used for purposes of statistical inference.

The powerful simulation-based estimation methods (MCMC) will be discussed in Chapter 3. Without diminishing the importance of the estimation methods, attention is first focused on the Bayesian way of thinking and modeling. Until then, WinBUGS (Lunn, Thomas, Best and Spiegelhalter, 2000) is used in the exercises and examples and it will be assumed that samples from the posterior distributions (the common output of simulation-based estimation methods) are available that can be used for making Bayesian inferences.

1.4 A Motivating Example Using WinBUGS

A simple example is given to illustrate the Bayesian modeling approach and the corresponding Bayesian inference. This example is worked out using the program WinBUGS (Lunn et al., 2000). The WinBUGS program is part of the Bayesian inference using Gibbs sampling (BUGS) project and allows one to put together Bayesian models and estimate simultaneously all model parameters, where WinBUGS facilitates the implementation of the simulation-based estimation method. Ntzoufras (2009) gives a thorough introduction to the WinBUGS program and illustrates the many Bayesian modeling possibilities via data examples.

1.4.1 Modeling Examinees' Test Results

In the Netherlands, primary schools administer the Cito Test developed by the National Institute for Educational Measurement (Cito) to get a reliable measurement of what children have learned during eight years of primary education. The scores on the test are used to advise children about the type of secondary education to take.

A relatively small sample of N=200 grade eight students responding to K=5 dichotomously scored mathematics items is considered. For the moment, the nesting of students in schools is ignored, but it will be discussed in Section 6.6.1. It will be assumed that the five math items measure a unidimensional ability in mathematics represented by θ, which is a continuous random variable that assumes values on the real line.

The probability of a correct response by examinee i to item k is modeled by a two-parameter item response model,

$$P\left(Y_{ik} = 1 \mid \theta_i, a_k, b_k\right) = \Phi\left(a_k\theta_i - b_k\right),$$

according to the normal ogive model in Equation (1.4). The response model consists of N ability parameters and K discrimination and K difficulty parameters. The examinees are assumed to be sampled independently from a population, and a normal prior density is specified for the ability parameters with mean zero and variance one. This restriction identifies the two-parameter item response model and also defines a useful scale for interpreting estimated ability values.

Prior densities for the item parameters will be thoroughly discussed in Section 2.2. Here, a common normal prior is assumed for the discrimination and difficulty parameters (e.g., Johnson and Albert, 1999),

$$\begin{aligned} a_k &\sim \mathcal{N}\left(\mu_a, \sigma_a^2\right) I(a_k > 0), \\ b_k &\sim \mathcal{N}\left(\mu_b, \sigma_b^2\right), \end{aligned}$$

for $k = 1, \ldots, K$. The discrimination parameter is restricted to be positive and usually takes on values between 1/2 and 3, and the prior should discourage

smaller or higher values. Difficulty parameters outside the interval $[-4, 4]$ will characterize the item as extremely easy or difficult and will lead to all correct or incorrect responses. The prior mean parameters are set to $\mu_a = 1$ and $\mu_b = 0$, which indicates a moderate level of discrimination and average level of difficulty. Both variance parameters are fixed to one.

WinBUGS

The model is implemented in WinBUGS for a response data matrix of N persons by K items. Each case i represents the responses of examinee i, and each column k represents all responses to item k. In the model description, all data points and parameters need to be specified. Therefore, the description contains a loop over observations (variable name Y), examinees (variable name *theta*), and items (variable names a and b).

Listing 1.1. WinBUGS code: Two-parameter item response model.

```
model{
    for (i in 1:N){
      for (k in 1:K){
         p[i,k] <- phi(a[k]*theta[i]-b[k])
         Y[i,k] ~ dbern(p[i,k])
       }
       theta[i] ~ dnorm(0,1)
   }
   for (k in 1:K) {
      a[k] ~ dnorm(1,1)I(0,)
      b[k] ~ dnorm(0,1)
   }
}
```

The WinBUGS output contains sampled values from each parameter's marginal posterior density. Each marginal posterior density provides complete information about the parameter. For Bayesian inference, the sampled values are usually used to compute summary statistics of posterior densities of parameters of interest. In Table 1.1, the marginal posterior density of each item parameter (discrimination and difficulty) is summarized. The posterior mean provides information on where most of the posterior density is located. The reported posterior mean is the expected value of the item parameter under the marginal posterior density. The posterior standard deviation and quantiles provide information about the spread of the posterior. As measures of spread, the posterior standard deviation and the 2.5% and 97.5% quantiles of each marginal posterior are reported.

The reported posterior means (expected a posteriori) are usually used as point estimates of the parameters. It follows that item five discriminates poorly and item one highly discriminates examinees of different ability. The average estimated discrimination level is .90, which is slightly smaller than the prior mean. The quantiles show that the posterior densities are nonsymmetric and positively skewed (right tails are longer), which follows from the positivity restriction on the discrimination parameter. The mean values are also higher

than the median values. For the difficulty parameter densities, the estimated posterior means are all negative. This means that the items are too easy since each item was answered correctly by more than 50% of the examinees given a zero average population level of ability. The raw data show that the proportions of correct responses of the five items are 56%, 73%, 54%, 71%, and 65%. Most of the students performed well on the test, which makes it more difficult to differentiate examinees. As shown, the items do not differentiate well (four item discriminations are less than one) since the items are too easy.

Table 1.1. Item parameters' posterior density information using WinBUGS.

Item	Mean	SD	2.5%	Median	97.5%
Discrimination Parameter					
1	1.54	.49	.82	1.45	2.75
2	.90	.25	.49	.87	1.47
3	.66	.18	.35	.65	1.05
4	.91	.24	.51	.88	1.43
5	.46	.15	.19	.45	.79
Difficulty Parameter					
1	−.27	.17	−.65	−.26	.04
2	−.79	.15	−1.12	−.78	−.52
3	−.11	.11	−.33	−.12	.09
4	−.73	.15	−1.05	−.72	−.47
5	−.42	.10	−.63	−.42	−.23

The posterior means correspond with the posterior medians, which means that the marginal posterior densities are approximately symmetric. However, the mean prior difficulty level $\mu_b = 0$ does not correspond with the estimated average posterior difficulty of −.46. For items 2, 4, and 5, the 97.5% left-sided posterior density interval does not contain the point zero. That is, the posterior probability that the item difficulty is higher than zero is less than 2.5%, which follows directly from the reported 97.5% quantile. This suggests that there is a discrepancy between the prior information and the sample information concerning the item difficulties. The posterior density is constructed from the prior and sample information, where the prior parameters μ_b and σ_b^2 define the prior weight.

To investigate the influence of this prior on the posterior, two cases will be considered. In the first case, the model is fitted with a prior variance parameter $\sigma_b^2 = .1$. This presents a stronger prior belief (a higher level of confidence), in comparison with $\sigma_b^2 = 1$, in a common item difficulty level of zero. In the second case, the variance parameter is not fixed but modeled via another prior distribution, and an inverse gamma density is used to define a set of

possible values. An inverse gamma prior with its parameters equal to .01 is uninformative or vague about the variance parameter, so that inferences are unaffected by information external to the data (provided that the variation is supported by the data). Subsequently, the variance parameter σ_b^2 becomes a model parameter that needs to be estimated.[3]

In Figure 1.5, the estimated posterior densities of the difficulty parameters are plotted for each prior setting. The stronger belief in the difficulty's prior level where $\sigma_b^2 = .1$ leads to a shift of the posterior density to the right, towards the prior mean. It follows that the prior's variance parameter influences at least the location of the posterior mean. Specifying prior parameters is difficult when not much is known beyond the data. By defining a prior for the variance parameter, instead of fixing its value, the data are used to estimate the prior variance. This approach is advisable when no prior information is available to specify the variance. The location of each posterior mean is constructed by combining sample and prior information, where the level of uncertainty about the prior mean is estimated by the response data. The estimated prior variance equals $\hat{\sigma}_b^2 = .47$ and, as a result, the corresponding posterior densities of the item difficulty parameters (dotted lines) are located approximately in the middle of the posterior densities with fixed prior parameters.

It was shown that the prior parameters influence the posterior analysis, and they require careful attention when making Bayesian inferences about parameters for which not much is known beyond the data. A flexible modeling framework was used that allows specific or noninformative prior settings, which was illustrated by modeling the variance parameter of the prior for the item difficulty parameters. This modeling framework will be explored further to make accurate individual (item) parameter estimates when only a few observations per respondent (item) are observed, to handle different sources of prior information, and to handle complex sampling designs, among other things. The modeling framework needs to be accompanied by a powerful estimation method that supports a realistic and practical way to make Bayesian inferences. Both computational and modeling issues will receive attention throughout the book.

1.5 Computation and Software

For the well-known item response models, various commercial and noncommercial programs are available. It is to the credit of the pioneering work of the researchers involved that today so many popular IRT programs are available. To give a short overview, BILOG-MG (Zimowski, Muraki, Mislevy and Bock, 1996) allows the estimation of IRT parameters for multiple groups and

[3] In WinBUGS, the variance parameter of a normal distribution is parameterized in terms of the inverse variance (precision) such that, in the second case, the precision parameter is modeled by a gamma prior.

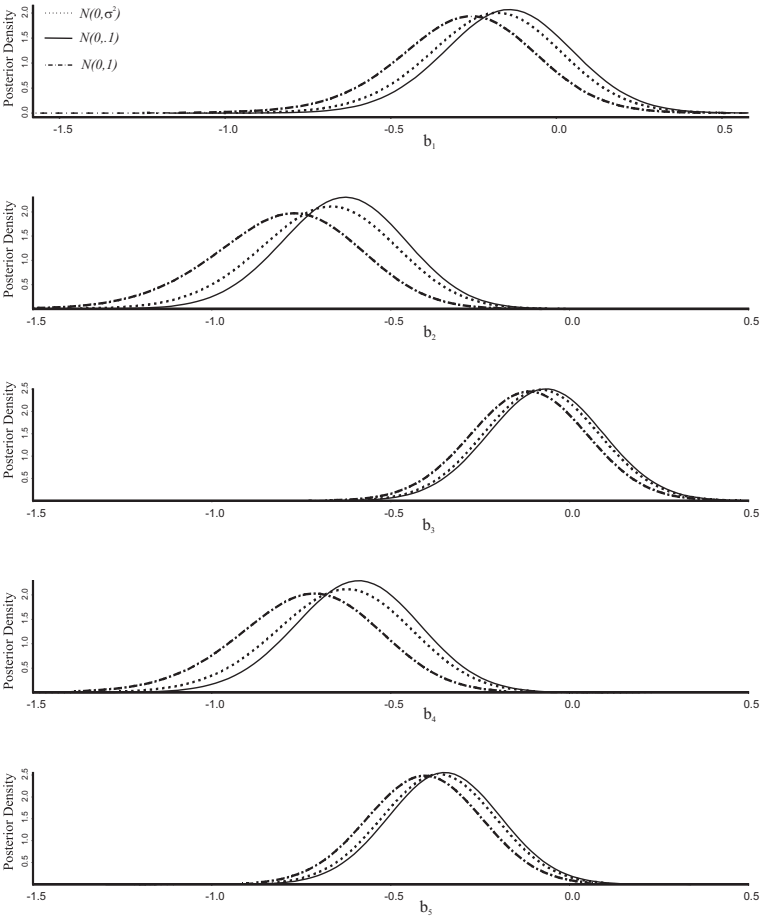

Fig. 1.5. Estimated posterior densities of the difficulty parameters for different prior choices.

enables detection of differential item functioning, among other uses. MULTILOG (Thissen, 1991) can be used specifically to perform a multiple-category IRT analysis for polytomous IRT models. PARSCALE (Muraki and Bock, 1997) is used for IRT scaling, item analysis, and scoring of rating scale data. A popular noncommercial (Dutch) program is OPLM (Verhelst, Glas and Verstralen, 1995); it can handle dichotomously or polychotomously scored items using different one-parameter models.

In the 1990s, the introduction of powerful simulation-based estimation methods made it possible to employ Bayesian methods without the common computational constraints, which also posed new statistical modeling opportunities. Twenty years later, the Bayesian paradigm is available to those with and without programming skills. The WinBUGS program

(www.mrc-bsu.cam.ac.uk/bugs) makes it possible to apply Bayesian methods to the analysis and modeling of data. The WinBUGS site has online help and contains lots of examples. Despite the many advantages of WinBUGS, for many models discussed in this book, the program is often too slow or simply does not work (e.g., Gelman and Hill, 2007). The main advantage is that it is very flexible in constructing models, but it can be slow, can get stuck with large datasets, and cannot handle complex item response models.

The free R (R Development Core Team, 2010) and commercial S+ (TIBCO Software, 2009) statistical programs are very popular in the Bayesian community. The programs contain methods for fitting specific models, but the popular higher programming languages of R and S+ allow one to program any model. Obviously, more knowledge of the estimation algorithm is needed in comparison with WinBUGS. Since the 1990s many R and S+ programs have been made available via the Internet; specifically, so-called R packages have been developed that allow one to construct programs that run within the R software environment. A list of contributed R packages can be found on the Comprehensive R Archive Network (CRAN; http://cran.r-project.org/). Several Bayesian (response) models used in marketing and microeconometrics applications are implemented in the bayesm package of Rossi, Allenby and McCullogh (2005). Bayesian inference for a number of response models using posterior simulation can be performed using the package MCMCpack. Gelman and Hill (2007) developed R programs for hierarchical models, including a two-parameter item response model. Various R packages are regularly updated and extended, and new contributions are frequently made, which makes it impossible to give a complete list of R packages that supports the Bayesian analysis of item response data.

Press (2003, pp. 169–171) listed references to popular Bayesian programs. This includes the Matlab and Minitab programs of Johnson and Albert (1999) for the analyses of ordinal data using Bayesian computational algorithms.

Computer Code Developed for This Book

Some models in this book can be handled by the programs mentioned, and other models require a specific implementation. To be free from the restrictions of other software programs, and to be completely flexible in defining different priors using different computational methods and computing or evaluating various statistics, I have programmed all models and methods in this book. The programs run in the R and S+ environments.

The revolution in Bayesian computational methods led to programs that needed days to come up with a solution. Large datasets, complex models with poorly identified parameters, and poorly implemented methods increased the computation time. In correspondence with Rossi et al. (2005, p. 7), the methods become impractical when more than a few hours of computing time is needed using a common computer. Users usually are not willing to wait that

long, especially when a simplified approach (e.g., by making additional assumptions, ignoring some complicating issues) only takes a few minutes. Furthermore, any statistical analysis requires fitting different models consisting of different priors and summarizing the inferences from different perspectives. Then, after evaluating the outcomes, model expansions with different prior information are considered, which is certainly impractical when each analysis takes more than a few hours.

The developed programs discussed in this book are written in high- and low-level languages to limit the computation time to around two hours. Several programs are written in the R and S+ languages. The programs can be used for the analysis of item response data but also serve as a basis for programming more complex item response models. Changing pieces of code can be very helpful in getting a better understanding of the substance and can be a first step in developing programming skills. Further, R packages and S+ programs are developed that make use of a dynamic link library (dll), which is a shared Microsoft Windows library. In Fortran (Intel Visual Fortran version 11 using IMSL Numerical Library version 6), programs are written that can be called within the R and S+ environment. The tools developed in Fortran are directly accessible, as are their input and output, and they can be manipulated within the statistical programs. To make the more complex models accessible for practical use, a low-level language is needed, and in my experience it will reduce the computation time roughly by a factor of ten.

Despite the increased CPU time and the increased size of available memory, the computational elements are important to make Bayesian inferences possible in a reasonable amount of time. Computation plays an important part in Bayesian statistical modeling, and to stress the importance of the computational methodology, the implemented algorithms are also described in this book. Those who just want to apply the models can use the software, but it also aims to serve those who want to implement and/or learn to develop and implement algorithms by themselves.

The programs and the data for the examples in the book are available on the World Wide Web at `www.jean-paulfox.com`, which contains more supporting material.

1.6 Exercises

WinBUGS and Listing 1.1 can be used to obtain the sampled values from the marginal posterior densities for making posterior inferences. The following exercises are based on output from WinBUGS. When fitting an item response model in WinBUGS, run one chain of 10,000 MCMC iterations and use the last 5,000 iterations.

1.1. In the example in Section 1.4, samples are obtained from each ability posterior density $p(\theta_i \mid \mathbf{y}_i)$.

(a) Graph the posterior density of the ability parameter of respondent i. Argue that the plotted posterior is not necessarily a symmetric density although a symmetric prior was assumed.
(b) Explain the summary statistics of θ_i reported by WinBUGS using the posterior density plot of Exercise 1.1(a).
(c) Graph the posterior density of a respondent's ability parameter that has all items correct and one that has all items incorrect. Given that a standard normal prior for the ability parameters was assumed, explain the direction of the skewness of the plotted densities.
(d) For a respondent who scores perfect, will the skewness of the ability posterior density increase or decrease when more items are administered?

1.2. The ability posterior density of examinee i is summarized.
(a) Argue that the posterior mean is often considered to be a good point estimate of the ability parameter. Note that the posterior mean equals the expected posterior ability and can be expressed as

$$E\left(\theta_i \mid \mathbf{y}_i\right) = \int \theta_i p\left(\theta_i \mid \mathbf{y}_i\right) \mathrm{d}\theta_i.$$

(b) Explain when the posterior mode might be considered as a point estimate. Note that the posterior mode equals the posterior ability point θ_{MAP} (maximum a posteriori) at which the posterior density is maximized,

$$\theta_{MAP} = \max_{\theta_i} p\left(\theta_i \mid \mathbf{y}_i\right).$$

(c) Given the sampled values, show how the posterior mean can be estimated and that the computation of the posterior mode is more complex.
(d) Argue that the posterior mean and variance can be used for adequately summarizing a symmetric posterior density but that various central points such as the mean, mode, and median, together with a region of high posterior probability are needed to summarize a nonsymmetric density.

1.3. (continuation of Exercise 1.1) Consider the computed posterior means as estimates of the ability parameters.
(a) Graph the density of the estimated abilities, and explain that this is the estimated empirical population density or sample density.
(b) Explain that the empirical population density is expected to be positively skewed where the right tail of the density is longer. (Note that the estimated item difficulty parameters are all negative.)
(c) Compute the sample skewness of the empirical population density with

$$\frac{\sqrt{N}\sum_{i=1}^{N}\left(\theta_i - \bar{\theta}\right)^3}{\left(\sum_{i=1}^{N}\left(\theta_i - \bar{\theta}\right)^2\right)^{3/2}},$$

where $\bar{\theta}$ is the estimated mean ability. (Listing 1.2 provides code to compute the sample skewness,)

Listing 1.2. WinBUGS code: Computing the sample skewness.

```
for(i in 1:N){
    numerator[i] <- (1/N)*pow(theta[i] - mean(theta[]),3)
    }
skewness <- sum(numerator[])/(pow(sd(theta[]),3))
```

(d) Does a skewed empirical population density indicate a model violation since a normal population prior is assumed?

1.4. (continuation of Exercise 1.1) Each model parameter has a (posterior) density function, which makes it possible to compute (posterior) probability statements.
(a) Compute the prior probability that the ability of examinee $i = 1$ is below the population average; that is,

$$P(\theta_1 < 0) = \int_{-\infty}^{0} \phi(x; \mu = 0, \sigma = 1)\, \mathrm{d}x.$$

(b) Compute the posterior probability that the ability of examinee $i = 1$ is below the population average; that is,

$$P(\theta_1 < 0 \mid \mathbf{y}) = \int_{-\infty}^{0} p(\theta_1 \mid \mathbf{y})\, \mathrm{d}\theta_1.$$

Use the WinBUGS code of Listing 1.3 or the sampled values from the posterior density to compute the posterior probability since the analytical form of the posterior density is unknown.

Listing 1.3. WinBUGS code: Computing the posterior probability of the event $\theta_1 < 0$.

```
counting <- max(theta[1],0)
probability <- equals(counting,0)
```

(c) In the same way, compute the prior and posterior probabilities that item three appears to be more difficult than item one.

1.5. Define priors for discrimination and difficulty parameters when additional information is available.
(a) Define an item difficulty prior that reflects a known order of items by difficulty. Explain how this influences the estimated item characteristic curves.
(b) Define an item discrimination prior that reflects a known order of items by discrimination. Explain how this influences the estimated item characteristic curves.
(c) Define a prior for the item parameters that reflects an ordering of items by difficulty and discrimination.
(d) Define a prior for the item parameters such that it is expected a priori that the more difficult an item is, the better it will discriminate.

2

Bayesian Hierarchical Response Modeling

In the first chapter, an introduction to Bayesian item response modeling was given. The Bayesian methodology requires careful specification of priors since item response models contain many parameters, often of the same type. A hierarchical modeling approach is introduced that supports the pooling of information to improve the precision of the parameter estimates. The Bayesian approach for handling response modeling issues is given, and specific Bayesian elements related to response modeling problems will be emphasized. It will be shown that the Bayesian paradigm engenders new ways of dealing with measurement error, limited information about many individuals, clustered response data, and different sources of information.

2.1 Pooling Strength

In test situations, interest is focused on the within-individual and between-individual response heterogeneities. The within-individual response heterogeneity provides information about test characteristics, individual response behavior, and the individual level of ability, among other things. The between-individual response heterogeneity provides information about test characteristics, the relationship between ability and individual background information, and clustering effects of respondents, among other things. The within-individual response differences are informative about the average characteristic levels across items, whereas the between-individual response differences are informative about each single item characteristic.

It is important to score respondents on a common scale, and a response model is needed that is capable of generating individual-level parameter estimates and associated uncertainties. At the same time, a response model is needed that is capable of producing item-level estimates and a characterization of the corresponding uncertainties. The focus on within-individual and between-individual differences corresponds with the interest in making inferences at disaggregate and aggregate levels. Bayesian hierarchical response

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_2,

models will prove to be very useful for making inferences at different hierarchical levels and make it possible to construct posterior distributions of individual-level parameters given that the data are sparse at the individual level.

Although response data are traditionally collected at the individual level, there is a tendency, specifically in large-scale survey studies, to collect (more) background information at various aggregated levels. Obtaining data at different levels allows inferences to be made at different levels. For example, response data from students across schools make it possible to compare students and schools with respect to the performances of the students. Background information can be used for making correct comparisons and also for explaining differences in performance at the student and school levels. The fact that more data become available at different levels creates modeling challenges and provides opportunities for making inferences that exploit the heterogeneity. The main challenge is to obtain accurate individual-level parameter estimates given very limited individual-level information but a large amount of survey respondents.

A common assumption is to regard the individual-level data as independent conditional on individual-level parameters. In that case, the marginal posterior of N respondents' abilities in Equation (1.11) can be written as

$$p(\theta_1, \ldots, \theta_N \mid \mathbf{y}) \propto \int \prod_i p\left(\mathbf{y}_i \mid \theta_i, \boldsymbol{\xi}\right) p\left(\theta_1, \ldots, \theta_N \mid \boldsymbol{\theta}_P\right) p\left(\boldsymbol{\xi}\right) \, \mathrm{d}\boldsymbol{\xi} \quad (2.1)$$

$$\propto \prod_i \int p\left(\mathbf{y}_i \mid \theta_i, \boldsymbol{\xi}\right) p\left(\theta_i \mid \boldsymbol{\theta}_P\right) p\left(\boldsymbol{\xi}\right) \, \mathrm{d}\boldsymbol{\xi},$$

where the proportionality sign is used since the normalizing constant is left out of the equation.

The joint prior on the abilities is simplified by assuming that the abilities are independent conditional on the hyperparameters $\boldsymbol{\theta}_P$ and identically distributed. The distinction between a parameter and a hyperparameter is that the sampling distribution of the (response) data is expressed directly conditional on the former (Lindley and Smith, 1972). The parameter's prior density is parameterized by the hyperparameters.

The first term on the right-hand side is the conditional likelihood. It follows that inferences for each respondent's ability θ_i can be made independently of all other respondents. Furthermore, the individual-level inferences are based on the sample and prior information. Therefore, the posterior mean is constructed from a combination of the prior mean and a likelihood-based estimate. For example, when the common population parameters $\boldsymbol{\theta}_P$ provide detailed information about the value of θ_i, the posterior mean will be shrunk more towards the prior mean. Note that such a shift towards the prior mean (of item difficulty locations) was shown in the example of Section 1.4.

The amount of shrinkage is determined by the form of the prior and the values of the hyperparameters. When the amount of within-individual infor-

mation is relatively small, the posterior for the θ_i will reflect a high level of shrinkage induced by the prior. The opposite is also true. The posterior for the θ_i will reflect a low level of shrinkage induced by the prior when the amount of within-individual information is high. The level of shrinkage induced by the prior can be inferred from the data by constructing a second-stage prior on the hyperparameters $\boldsymbol{\theta}_P$. This prior on the hyperparameters is sometimes called a hyperprior (Berger, 1985). The sample information and the prior for the hyperparameters will be used to make inferences about the $\boldsymbol{\theta}_P$. Then, the level of shrinkage is determined by the information in the data (Exercise 3.3).

In a strictly hierarchical modeling approach, it is assumed that the responses are nested in individuals, individuals nested in groups, and so on. Typically, response data are clustered in a cross-classified way where respondent effects and item effects are present. To separate the effects of individuals and items, two cross-cutting hierarchies need to be specified. Therefore, assume that the item-level response data are independent conditional on the item-level parameters $\boldsymbol{\xi}_k$. This independence assumption leads to a within-item and between-item structure and defines the second hierarchy of the response data. Analogous to the specification of the joint posterior of abilities, the joint posterior of the item parameters can be written as

$$\begin{aligned} p(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K \mid \mathbf{y}) &\propto \int \prod_k p\left(\mathbf{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\xi}_k\right) p\left(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K \mid \boldsymbol{\xi}_P\right) p\left(\boldsymbol{\theta}\right) \, \mathrm{d}\boldsymbol{\theta} \quad (2.2) \\ &\propto \prod_k \int p\left(\mathbf{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\xi}_k\right) p\left(\boldsymbol{\xi}_k \mid \boldsymbol{\xi}_P\right) p\left(\boldsymbol{\theta}\right) \, \mathrm{d}\boldsymbol{\theta}, \end{aligned}$$

where the prior for the item parameters is structured by assuming independence conditional on the hyperparameters $\boldsymbol{\xi}_P$. Subsequently, prior and sample information are used to estimate the item effects, where shrinkage effects can be inferred from the data by adding a second layer for the hyperparameters $\boldsymbol{\xi}_P$ (see the example in Section 1.4).

2.2 From Beliefs to Prior Distributions

The Bayesian approach to response modeling starts with the specification of prior distributions. In general, there are two types of prior information. First, when there is prior information about the values of the model parameters from, for example, related datasets, this information can be used to construct a case-specific prior. This type will be discussed in subsequent chapters when additional data are available and case-specific information can be used to construct priors.

Second, the prior information can come from mathematical properties. Prior distributions can be classified into conjugate or nonconjugate priors. A conjugate prior has the property that the posterior has the same algebraic form as the prior. This has the advantage that the posterior has a known

analytical form, which simplifies the statistical analysis. A nonconjugate prior leads to a posterior that often has a complicated functional form, which makes the statistical analysis mathematically more challenging. A conjugate prior can easily reflect a likely range of parameter values, but remember that the shape of the prior also has an impact on the posterior analysis. Lord (1986) stressed the practical advantages of priors since, for example, the priors restrict the parameter values to a plausible range. Although a range of priors can be considered for item response models, only a limited number based on well-known distributions are used in practice.

The hierarchical or multistage prior will prove to be very useful (e.g., Berger, 1985, Section 3.6; Lindley and Smith, 1972). The hierarchical prior consists of a first-stage prior for many parameters of the same type that are assumed to be independent given hyperparameters. The hyperparameters have their own prior at a second stage. The first-stage cluster-specific parameters (e.g., item and person parameters) can be related with an aggregated data source at the cluster level. The similarity between the first-stage parameters and their possible relationship with aggregated prior information motivates the hierarchical prior based on combining information to improve the precision of each first-stage parameter estimate. The hierarchical prior improves the estimation of the first-stage parameters by pooling information (borrowing strength) over clusters and by accounting for uncertainty in the hyperparameter estimates. Typically, there are many individuals but relatively little response data on each individual. Pooling information over individuals exploits the assumed similarity between the individual parameters to improve the individual parameter estimates. The hierarchical prior can be extended to more than two stages, and such extensions will be discussed in subsequent chapters.

A Hierarchical Prior for Item Parameters

In most cases, there is not much information about the values of the item parameters, and the response data are the source of information to distinguish the item parameters from each other. Without a priori knowledge to distinguish the item parameters, it is reasonable to assume a common distribution for them. In that case, the item parameters have a common population distribution and it is not possible to order or group the parameters. Typically, the parameters are said to be exchangeable in their joint distribution, which means that the joint probability of the item parameters is invariant with respect to permutations of the indices. That is, it is assumed that the prior information about the item parameters is exchangeable.

The exchangeability assumption is closely related to the concept of independently and identically distributed but not the same. Independently and identically distributed implies exchangeability, and exchangeable item parameters do have identical marginal distributions, but they are not necessarily independent. The assumption of exchangeable item parameters is equivalent

to the assumption of conditionally independent and identically distributed item parameters given hyperparameters and a prior density on the hyperparameters.

An intuitive assumption of an item characteristic curve is that the higher a respondent's ability level the more likely it is that the respondent scores well on the item. This so-called monotonicity assumption implies that $P(Y_{ik} = 1 \mid \theta_i)$ is nondecreasing in θ_i, which is satisfied when the discrimination parameter is restricted to be positive. For example, in Bilog-MG (Zimowski et al., 1996), a lognormal prior distribution can be specified to restrict a discrimination parameter to be positive. A normal prior is often used for each difficulty parameter (e.g., Albert, 1992; Patz and Junker, 1999a). Exchangeability is usually assumed such that the priors have the same hyperparameters in addition to the common form.

The assessment of the hyperparameters can be challenging since they can have a substantial effect on the item parameter estimates. The greater the discrepancy between the sample-based information and the prior information, the larger the amount of shrinkage towards the prior mean when keeping other factors constant. A small prior variance leads to an informative prior and greater shrinkage. A large prior variance implies a noninformative prior and almost no shrinkage. The effective use of a prior for item parameters depends on the hyperparameter specifications, and that requires insight into the test characteristics and respondents. Mislevy (1986) remarked that incorrectly specifying the prior mean can result in biased item difficulty estimates. Below, a hierarchical prior is defined such that the hyperparameters are estimated by the data, and the amount of shrinkage is inferred from the data. Then, the appropriateness of the mean and variance of the prior is arranged using an estimation procedure.

Albert (1992) and Patz and Junker (1999a), among others, suggested independent prior distributions for the item parameters with fixed hyperparameter values: a (log)normal prior for the discrimination parameters and a normal prior for the difficulty parameters. Johnson and Albert (1999) defined a hierarchical prior for the discrimination parameters: a normal density prior at stage 1, and at stage 2 an inverse gamma prior for the variance (with hyperparameters equal to one) and a uniform prior for the mean. Bradlow, Wainer and Wang (1999) and Kim, Cohen, Baker, Subkoviak and Leonard (1994) also defined independent hierarchical priors for the item parameters. Tsutakawa and Lin (1986) proposed a bivariate prior for the item parameters induced by the probability of correct responses to the items at different ability levels. This requires specifying the degree of belief about the probability of a correct response to each item at two ability levels. Although their prior provides a way to specify dependence among parameters within an item, it is often difficult to specify prior beliefs about success probabilities objectively.

A more straightforward realistic multivariate normal prior allows for within-item characteristic dependencies; that is,

$$(a_k, b_k)^t \sim \mathcal{N}(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)\, I_{\mathcal{A}_k}(a_k), \tag{2.3}$$

where $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$. The truncated multivariate normal density in Equation (2.3) is an exchangeable prior for item parameters $\boldsymbol{\xi}_k$ $(k = 1, \ldots, K)$. The hyperparameters are modeled at a higher level since there is usually little known about the mean item discrimination, mean item difficulty, and variances. An inverse Wishart density, denoted as $\mathcal{IW}$, with scale matrix $\boldsymbol{\Sigma}_0$ and degrees of freedom $\nu \geq 2$, is commonly used to specify $\boldsymbol{\Sigma}_\xi$.[1] This is a conjugated prior for the covariance matrix. Then, a normal prior is used to specify the prior mean given the prior variance matrix $\boldsymbol{\Sigma}_\xi$. The joint prior density for $(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ equals

$$\boldsymbol{\Sigma}_\xi \sim \mathcal{IW}(\nu, \boldsymbol{\Sigma}_0), \tag{2.4}$$

$$\boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\xi / K_0). \tag{2.5}$$

This joint prior density is known as a normal inverse Wishart density, where K_0 denotes the number of prior measurements (e.g., Gelman, Carlin, Stern and Rubin, 1995). A proper noninformative prior is specified with $\boldsymbol{\mu}_0 = (1, 0)^t$, $\nu = 2$, a scale matrix $\boldsymbol{\Sigma}_0$ that is a minimally informative prior guess of $\boldsymbol{\Sigma}_\xi$, and K_0 a small number.

The hierarchical prior accounts for within-item dependencies and for uncertainty of the prior's parameters. The mean and variance of the normal prior for the item parameters are modeled at a higher prior level and need to be estimated from the data. As a result, the hierarchical prior gives rise to shrinkage estimates of the item parameters, where the amount of shrinkage is inferred from the data. An extreme and infinite item parameter estimate caused by the fact that the item is answered correctly or incorrectly by all respondents is avoided due to shrinkage towards the mean of the prior.

When observing responses on a continuous scale that are assumed to be normally distributed, the hierarchical prior presented is a conjugate prior (e.g., Mellenbergh, 1994b). In Section 4.2, for different reasons, the discrete observed response data will be augmented with normally distributed continuous data. The hierarchical normal prior will be shown to be a conjugate prior with respect to the normal likelihood of the augmented data, which will simplify the procedure for making posterior inferences. A lognormal version of the normal prior presented in Equation (2.3) will be discussed in Exercise 2.3 (see also Exercise 4.3).

The following theoretical motivation provides more arguments for modeling a covariance structure between item parameters. Let θ_l and θ_u be the average ability levels of respondents sampled from low- and high-ability groups with corresponding success probabilities $P_k(\theta_l)$ and $P_k(\theta_u)$ for item k, respectively. Consider the difference between success probabilities, $P_k(\theta_u) - P_k(\theta_l)$,

[1] A $q \times q$ random positive-definite symmetric matrix $\boldsymbol{\Sigma}_\xi$ is distributed according to an $\mathcal{IW}$ distribution with ν degrees of freedom and scale matrix $\boldsymbol{\Sigma}_0$ if its probability density function is proportional to $|\boldsymbol{\Sigma}_\xi|^{-(\nu+q+1)/2} \exp\left(-\operatorname{tr}\left(\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_\xi^{-1}\right)/2\right)$.

as an estimate of discrimination of item k. The higher this difference, the better the item discriminates between respondents with low and high abilities. Consider the mean success probability $(P_k(\theta_u) + P_k(\theta_l))/2$ as an estimate of the item difficulty. A linear relationship between the item parameter estimates is defined as

$$\begin{aligned} & \left(P_k(\theta_u) + P_k(\theta_l)\right)/2 = \rho\left(P_k(\theta_u) - P_k(\theta_l)\right) \\ \iff & P_k(\theta_u) + P_k(\theta_l) = 2\rho\left(P_k(\theta_u) - P_k(\theta_l)\right) \\ \iff & \frac{P_k(\theta_u)}{P_k(\theta_l)} = \frac{\rho + 1/2}{\rho - 1/2} \end{aligned} \tag{2.6}$$

for $k = 1, \ldots, K$, and for any constant $\rho > 1/2$. A consistency in the ratio of the mean success probabilities across the K items induces a covariance structure between the item parameters. In Equation (2.6), items of decreasing difficulty will discriminate better between the low- and high-ability respondents for $\rho > 1/2$. If the success probabilities on the left-hand side of Equation (2.6) are replaced by the failure probabilities, items of increasing difficulty will discriminate better between the low- and high-ability respondents for $\rho > 1/2$. The relationships show that a consistency in the ratio of group-specific response probabilities across items can correspond with a common within-item dependency structure.

In the three-parameter model, the guessing parameter, c_k, is bounded above by one and below by zero since it represents the probability that a respondent correctly guessed the answer to item k. A convenient prior is the beta density with parameters α and β that reflects $\alpha - 1$ prior successes and $\beta - 1$ prior failures (Swaminathan and Gifford, 1986; Zimowski et al., 1996). Most often the guessing parameters are assumed to be a priori independently and identically beta distributed (Exercise 2.4),

$$p\left(c_k \mid \alpha, \beta\right) = \frac{\Gamma\left(\alpha + \beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} c_k^{\alpha-1}(1 - c_k)^{\beta-1}. \tag{2.7}$$

The normalizing constant contains the gamma function, which is defined as $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x$ for $\alpha > 0$ (Exercise 3.8). A hierarchical prior for the guessing parameters is constructed by modeling the parameters of the beta prior. Usually a uniform hyperprior is specified for the beta parameters.

Both priors in Equations (2.3) and (2.7) are nonconjugate for the observed-data likelihood.[2] In Chapter 4, it will be shown that a nonconjugate Bayesian analysis is more complex and requires a different method for estimating the item parameters. An exchangeable hierarchical normal prior for the three item parameters (a_k, b_k, c_k) is possible via a reparameterization of the guessing parameter. A straightforward way is to use an inverse normal transformation

[2] In Section 4.2, it is shown that, for a specific data augmentation scheme, the posterior of c_k given augmented data is also a beta density, which makes the beta prior in Equation (2.7) a conjugate prior.

function, $\tilde{c}_k = \Phi^{-1}(c_k)$, which ensures that $\tilde{c}_k$ is defined on the whole real line (see Exercise 4.9).

The prior distribution for the threshold parameters in, for example, the graded response model is usually chosen to be noninformative. Albert and Chib (1993) defined a uniform prior distribution for parameter $\kappa_{k,c}$ but truncated to the region $\{\kappa_{kc} \in \mathcal{R}, \kappa_{k,c-1} < \kappa_{k,c} \leq \kappa_{k,c+1}\}$ to take account of the order constraints. Johnson and Albert (1999) argued that a uniform prior for the thresholds assigns equal weight to all grades, although some grades may be known to be rare. They proposed a more sophisticated prior where the lower threshold equals the upper threshold when they correspond to an unobserved grade. Other prior distributions for the threshold parameters will be discussed in subsequent chapters.

A Hierarchical Prior for Person Parameters

A prior for person parameters assumes that the respondents represent a sample from a known population. In the case of a simple random sample, a subset of individuals (a sample) are chosen from a larger set (a population) and each individual is chosen randomly, where each individual has the same probability of being chosen at any stage during the sampling process. The respondents are assumed to be sampled independently from a large population; that is, the person parameters are independently distributed from a normal population distribution,

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \tag{2.8}$$

for $i = 1, \ldots, N$, where the mean and variance parameters are unknown and are modeled at a higher level. A normal inverse gamma prior is the conjugate prior for the normal distribution with unknown mean and variance. Therefore, a joint hyperprior is specified as

$$\sigma_\theta^2 \sim \mathcal{IG}(g_1, g_2), \tag{2.9}$$

$$\mu_\theta \mid \sigma_\theta^2 \sim \mathcal{N}(\mu_0, \sigma_\theta^2/n_0), \tag{2.10}$$

where g_1 and g_2 are the parameters of the inverse gamma density denoted as $\mathcal{IG}$ and n_0 presents the number of prior measurements.[3] Other prior population distributions for more complex sampling designs will be thoroughly discussed in later chapters.

2.2.1 Improper Priors

There have been attempts to construct noninformative priors that contain no (or minimal) information about the parameters. The noninformative prior

[3] Parameter σ is inverse gamma ($\mathcal{IG}$) distributed with shape parameter g_1 and scale parameter g_2 when its pdf is proportional to $\sigma^{-(g_1+1)} \exp(-g_2/\sigma)$. The inverse gamma distribution with ν degrees of freedom corresponds to an inverse chi-square distribution when $g_1 = \nu/2$ and $g_2 = 1/2$.

does not favor possible values of the parameter over others. A prior for the difficulty parameter that is completely determined or dominated by the likelihood is one that does not change much in the region where the likelihood is appreciable and does not assume large values outside this region. A prior with these properties is referred to as a locally uniform prior (Box and Tiao, 1973, p. 23). An exchangeable locally uniform prior for the difficulty parameters leads to improper independent priors $p(b_k)$ equal to a constant. When the difficulty parameters have a locally uniform prior, meaning that they bear no strong relationship to one another, they can be considered fixed effects parameters. In that case, interest is focused on the estimation of the difficulty parameters rather than the properties of their population distribution.

For example, a noninformative prior for the item difficulty parameter gives equal weight to all possible values. That is, $p\left(b_k\right) = c$, where c is a constant greater than zero. As a result, $\int p\left(b_k\right) \mathrm{d}b_k = \infty$, and this noninformative prior is an improper prior since it does not integrate to one. The value of c is unimportant, and typically the noninformative prior density is $p\left(b_k\right) = 1$.

Improper priors may lead to improper posteriors, which means that the posterior is not a valid distribution and posterior moments have no meaning (Exercise 6.8). Ghosh, Ghosh, Chen and Agresti (2000) showed that for the one-parameter model the specification of improper priors for person and item parameters leads to an improper joint posterior. Improper priors are often conjugate and may lead to a straightforward simulation-based estimation method. Such simulation-based methods make use of conditional distributions of the model parameters that are available in closed form, but the marginal posterior distributions are not. The conditional distributions can all be well defined and can be simulated from, yet the joint posterior can be improper (Robert and Casella, 1999, pp. 328–332). Demonstrating the property of the posterior is often impossible. The best way to avoid improper posteriors is to use proper priors. Noninformative proper priors can also be used to specify less or no prior information, and they avoid the risk of getting improper posteriors.

2.2.2 A Hierarchical Bayes Response Model

Summing up, the posterior density of interest is constructed from a response model for the observed data and a hierarchical prior density. In a hierarchical prior modeling approach, the prior parameters are modeled explicitly and have hyperprior densities at the second stage of the prior.

Suppose posterior inferences are to be made about the item and person parameters that require integration over the density functions of the hyperparameters. The hyperparameters are denoted by $\boldsymbol{\theta}_P = \left(\mu_\theta, \sigma_\theta^2\right)$ and $\boldsymbol{\xi}_P = (\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$. The posterior density of interest can be expressed as

$$
\begin{aligned}
p(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \mathbf{y}) &\propto \int\int p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})\, p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \boldsymbol{\theta}_P, \boldsymbol{\xi}_P)\, p(\boldsymbol{\theta}_P, \boldsymbol{\xi}_P)\, \mathrm{d}\boldsymbol{\xi}_P \mathrm{d}\boldsymbol{\theta}_P \\
&\propto \int\int p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})\, p(\boldsymbol{\theta} \mid \boldsymbol{\theta}_P)\, p(\boldsymbol{\xi} \mid \boldsymbol{\xi}_P)\, p(\boldsymbol{\theta}_P)\, p(\boldsymbol{\xi}_P)\, \mathrm{d}\boldsymbol{\xi}_P \mathrm{d}\boldsymbol{\theta}_P \\
&\propto \int\int \prod_{i,k} \left[p(y_{ik} \mid \theta_i, \xi_k)\, p(\theta_i \mid \boldsymbol{\theta}_P)\, p(\xi_k \mid \boldsymbol{\xi}_P) \right] \cdot \\
&\quad p(\boldsymbol{\theta}_P)\, p(\boldsymbol{\xi}_P)\, \mathrm{d}\boldsymbol{\xi}_P \mathrm{d}\boldsymbol{\theta}_P. \qquad (2.11)
\end{aligned}
$$

The hierarchical prior model can be recognized in the first equation, where the (hyper)parameters of the prior also have a prior density. The item and person parameters are assumed to be independent from each other, and the corresponding hyperparameters are also assumed to be independent from each other. This leads to the factorization in the second equation.

Typically, the observations are assumed to be conditionally and independently distributed given item and person parameters. That is, the observations are assumed to be clustered in a cross-classified way. Furthermore, the person parameters as well as the item parameters are assumed to be independent from one another. As a result, the joint posterior of the parameters of interest can be expressed as a product of identically distributed observations given person and item parameters, where the person and the item parameters are identically distributed given common hyperparameters; see Equation (2.11).

The last factorization, Equation (2.11), illustrates the hierarchical modeling approach. The observations are modeled conditionally independent at the first stage given item and person parameters, $p(y_{ik} \mid \theta_i, \xi_k)$. This is the likelihood part of the model which describes the distribution of the data given first-stage parameters. At the second stage, priors are specified for the first-stage parameters. The first-stage priors consist of a prior describing the between-individual heterogeneity, $p(\theta_i \mid \boldsymbol{\theta}_P)$, and a prior describing the between-item heterogeneity, $p(\xi_k \mid \boldsymbol{\xi}_P)$. At the third stage, hyperpriors are defined for the parameters of the first-stage priors.

Typically, the variability between individuals is modeled via a conditionally independent prior by conditioning on second-stage parameters. This allows making inferences independently of other respondents, and the conditional independence assumption simplifies the joint prior for the numerous person parameters. In the same way, the prior that describes the between-item variability in item characteristics assumes independent item characteristics given second-stage parameters. The second-stage parameters or hyperparameters control the priors for the lower-level parameters. Then, a second-stage prior for the hyperparameters, $p(\boldsymbol{\theta}_P)$ and $p(\boldsymbol{\xi}_P)$, is defined. The combination of a likelihood and a hierarchical prior that consists of a first-stage conditional independent prior and a second-stage prior for the hyperparameters is called a hierarchical Bayes model.

Inferences about the parameters of interest, the first-stage parameters, are based on information from the data and prior information. The contribution

of the first-stage prior to the posterior depends on the values of the prior parameters. For example, the posterior means of the first-stage parameters will show an amount of shrinkage towards the prior mean depending on the values of the hyperparameters. More shrinkage will occur when specifying a more informative first-stage prior. Assessment of the hyperparameters is difficult, and it is desirable to let the level of control of the first-stage prior be driven by information in the data. Therefore, the hyperparameters are modeled at the second stage of the prior. As a result, the priors' contribution to the posterior is arranged by information in the data taking the hyperparameters' uncertainty into account.

Posterior Computation

The integration problem quickly expands when computing marginal posterior means for all parameters and when extending the prior and/or likelihood model. For example, the computation of the ability posterior density requires integration over the item parameter, item population, and ability population densities. This leads to the computation of

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int\int\int p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})\, p(\boldsymbol{\theta} \mid \boldsymbol{\theta}_P)\, p(\boldsymbol{\theta}_P) \\ p(\boldsymbol{\xi} \mid \boldsymbol{\xi}_P)\, p(\boldsymbol{\xi}_P)\, \mathrm{d}\boldsymbol{\theta}_P\, \mathrm{d}\boldsymbol{\xi}\, \mathrm{d}\boldsymbol{\xi}_P. \tag{2.12}$$

When assuming normal priors, the integrations might be performed via a Gauss Hermite. Although other quadrature (numerical integration) methods are available, a quadrature approach is limited to a certain range of integration problems.

Obtaining a satisfactory approximation of the integration problem in (2.12) via numerical integration is a complex task, and the available numerical integration methods cannot handle integrals in dimensions greater than four. It will be shown that in large-scale survey research, the computation of properties of posteriors of interest is often further complicated due to the use of complex parameter spaces (e.g., constrained parameter sets), complex sampling models with intractable likelihoods (e.g., due to the presence of missing data), an extremely large dataset, complex prior distributions, or a complex inferential procedure.

In the next chapter, simulation-based Bayesian estimation methods are discussed that are capable of estimating all model parameters simultaneously. This powerful estimation method supports the hierarchical modeling framework in a natural way. In subsequent chapters, it will be shown that all kinds of model extensions can also be handled by the computational methods discussed.

2.3 Further Reading

Hierarchical response modeling makes up part of the popular hierarchical modeling framework, which has a tremendous amount of literature. The hierarchical models are used in many areas of statistics. For a more complete overview with applications in the educational and social sciences, see, for example, Goldstein (2003), Longford (1993), Raudenbush and Bryk (2002), Skrondal and Rabe-Hesketh (2004), and Snijders and Bosker (1999).

Recently, the Bayesian formulation (Berger, 1985; Lindley and Smith, 1972) has received much attention. Congdon (2001), Gelman and Hill (2007), and Rossi et al. (2005) show various examples of Bayesian hierarchical modeling using WinBUGS and R. The Bayesian modeling approach accounts for uncertainty in the variance parameters, which is particularly important when the hierarchical variances are difficult to estimate or to distinguish from zero (see Carlin and Louis, 1996; Gelman et al., 1995). The Bayesian hierarchical modeling approach for response data will be thoroughly discussed in subsequent chapters.

The Bayesian hierarchical item response modeling framework has been advocated by Mislevy (1986), Novick, Lewis and Jackson (1973), Swaminathan and Gifford (1982, 1985), Tsutakawa and Lin (1986), and Tsutakawa and Soltys (1988). Swaminathan and Gifford (1982) defined a hierarchical Rasch model with, at the second stage, normal priors for the ability and difficulty parameters. At the third stage, parameters of the prior for the ability parameters were fixed to identify the model. The mean and variance parameters of the exchangeable prior for the difficulty parameters were assumed to be uniformly and inverse chi-square distributed, respectively. For the hierarchical two-parameter model, a chi-square prior was additionally specified for the discrimination parameters with fixed parameters (Swaminathan and Gifford, 1985, 1986). Tsutakawa and Lin (1986) described a hierarchical prior for the difficulty parameters and a standard normal prior for the ability parameters.

In the 1990s, the hierarchical item response modeling approach was picked up by Kim et al. (1994) and Bradlow et al. (1999). After the millennium, the hierarchical item response modeling approach became more popular. This approach will be further pursued in Chapter 4, where several examples are given to illustrate its usefulness.

2.4 Exercises

The exercises can be made using WinBUGS and are focused on hierarchical item response modeling. Different response models are specified in WinBUGS, and the output is used for making posterior inferences. The computational aspects will be discussed in Chapter 3. When it is required, run one chain of 10,000 MCMC iterations and use the last 5,000 iterations.

2.1. Johnson and Albert (1999) consider mathematics placement test data of freshmen at Bowling Green State University. The mathematics placement test is designed to assess the math skill levels of the entering students and recommend an appropriate first college math class. The test form B consists of 35 (dichotomously scored) multiple-choice items.
(a) Adjust the code in Listing 1.1 to fit a two-parameter logistic response model. Obtain item parameter estimates given the fixed hyperparameter values.
(b) Explain that a less informative prior is defined when increasing the prior variance.
(c) Estimate the item parameters for different item prior variances.
(d) Observe and explain the effects of shrinkage by comparing the item parameter estimates.
(e) Evaluate the estimated posterior standard deviations of the item parameters. Do they change due to different hyperparameter settings?

2.2. (continuation of Exercise 2.1) In this problem, attention is focused on assessing the parameters of the difficulty and discrimination prior.
(a) Explain the hierarchical item prior specified in Listing 2.1.

Listing 2.1. WinBUGS code: Independent hierarchical prior for item parameters.

```
for(k in 1:K){
    a[k] ~ dnorm(mu[1],prec[1])I(0,)
    b[k] ~ dnorm(mu[2],prec[2])
 }

  mu[1] ~ dnorm(1,1.0E-02)
  mu[2] ~ dnorm(0,1.0E-02)

  prec[1] ~ dgamma(1,1)
  sigma[1] <- 1/prec[1]

  prec[2] ~ dgamma(1,1)
  sigma[2] <- 1/prec[2]
```

(b) Fit the two-parameter logistic response model with the independent hierarchical prior for the item parameters. Investigate the influence of the second-stage prior settings. (Use appropriate starting values for the hyperparameters.)
(c) Plot the estimated posterior means of the discrimination parameters against the differences between the estimated posterior means of Exercises

2.2(b) and 2.1(a), and explain the graph. Do the same for the difficulty parameters.

2.3. (continuation of Exercise 2.1) In this problem, the within-item characteristic dependencies are investigated.
(a) Given the results from Exercise 2.2(b), plot the posterior means of the difficulty parameters against the posterior means of the discrimination parameters and see whether there is a trend visible.
(b) A hierarchical prior for the item parameters is specified in Listing 2.2. Explain that the discrimination parameters are positively restricted by a log transformation.

Listing 2.2. WinBUGS code: Hierarchical prior for item parameters.

```
for (k in 1:K) {
        item[k,1:2] ~ dmnorm(mu[1:2], prec[1:2,1:2])
        a[k] <- exp(item[k,1])
        b[k] <- item[k,2]
   }

    prec[1:2,1:2] ~ dwish(S[1:2,1:2],4)
    Sigma[1:2,1:2] <- inverse(prec[1:2,1:2])
```

(c) Use the code in Listing 2.2 to define a two-parameter logistic response model with the hierarchical prior for the item parameters, and fit the model. (Use appropriate starting values for the hyperparameters.)
(d) Evaluate the estimated posterior means of the within-item covariance parameters, and explain the results.

2.4. (continuation of Exercise 2.1) In this problem, the three-parameter logistic response model is used to study guessing behavior of students taking the mathematics placement test.
(a) Adjust the code of Listing 1.1 to define a three-parameter response model (Equation (1.5)) with a beta prior for the guessing parameters (Equation (2.7)).
(b) The parameters α and β of the beta prior specify the mean and variance of the density via

$$E(c_k) = \frac{\alpha}{\alpha+\beta},$$

$$\mathrm{Var}(c_k) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

and the beta density is unimodal when both parameters are greater than one. Specify the beta parameters such that a priori the mean guessing probability is .25 (random guessing) with a standard deviation greater than .1. Fit the three-parameter model, and evaluate the estimated posterior means of the guessing parameters. Do they vary around the prior mean?
(c) Assume, at the second prior stage, that the parameters of the beta prior are uniformly distributed on the interval $[2, 100]$, and fit the three-parameter model. Does the beta prior dominate the final results?

3

Basic Elements of Bayesian Statistics

A review of Bayesian estimation and testing methods is given that is not a thorough overview but concentrates on some specific elements. First, simulation-based methods for parameter estimation, like the Gibbs sampling and the Metropolis-Hastings algorithms, from the general class of Markov chain Monte Carlo algorithms, are discussed. Second, the Bayesian approach to model selection and hypothesis testing is presented. The techniques and methods described in this chapter are needed to completely exploit the Bayesian machinery for item response modeling.

3.1 Bayesian Computational Methods

The approach to the computational problems mentioned in Section 2.2.2 is based on computer simulations that exploit the probabilistic properties of the integrands and can handle the dimensionality of the problem. Numerical integration methods are often based on analytical properties of the integrand where less gain is to be expected, and they often rely on specific knowledge of the distribution of interest. Furthermore, simulation-based methods are appealing for two reasons. First, integration problems often involve probability distributions in the integrand, which insinuates the use of simulation methods. Second, the simulation methods are generally straightforward to implement since they are often based on a few principles of simulation and the structure of the problem.

The introduction of powerful simulation methods made Bayesian modeling possible in various research fields covering a wide range of applications. The posterior simulation methods make the posterior distributions accessible; that is, the algorithms for posterior simulation can be used to obtain approximates of posterior moments.

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_3,

3.1.1 Markov Chain Monte Carlo Methods

Simulating values directly from the posterior of interest (direct sampling) is often not possible. Intractable marginal posteriors and the dimensionality of the problem lead to difficulties in obtaining directly simulated values from the posteriors. In general, more sophisticated methods are needed. A class of simulation methods known as Markov chain Monte Carlo (MCMC) build sequences that converge in distribution to the posterior (target) distribution. Then, sample averages are computed to estimate posterior expectations. Essentially, MCMC is Monte Carlo integration using Markov chains. In Monte Carlo integration, samples are drawn to perform the integration (Exercise 3.4). In MCMC, the samples are drawn via a constructed Markov chain that has the target distribution as its stationary distribution. The number of MCMC methods is growing rapidly, and this overview is not meant to be exhaustive. It is an overview of methods that have proven to be useful in the analysis of item response data. For a more complete overview, see Chen, Shao and Ibrahim (2000), Gilks, Richardson and Spiegelhalter (1995), and Robert and Casella (1999), among others.

Gibbs Sampling

The most popular MCMC method is Gibbs sampling, its name originating from a class of probability distributions for modeling spatial interactions and spatial stochastic processes (e.g., Geman and Geman, 1984). It starts with the partitioning or blocking of parameters or random vectors of interest in subvectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q)$. The joint posterior density of the random vector equals $p(\boldsymbol{\theta} \mid \mathbf{y})$, and this is also the target density. A transition process from $\boldsymbol{\theta}^{(m)}$ to $\boldsymbol{\theta}^{(m+1)}$ is defined by making draws at iteration $m+1$ from the conditional pdf of each subvector,

$$
\begin{aligned}
\boldsymbol{\theta}_1^{(m+1)} &\sim p\Big(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(m)}, \ldots, \boldsymbol{\theta}_Q^{(m)}, \mathbf{y}\Big) \\
\boldsymbol{\theta}_2^{(m+1)} &\sim p\Big(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(m+1)}, \boldsymbol{\theta}_3^{(m)}, \ldots, \boldsymbol{\theta}_Q^{(m)}, \mathbf{y}\Big) \\
&\vdots \\
\boldsymbol{\theta}_Q^{(m+1)} &\sim p\Big(\boldsymbol{\theta}_Q \mid \boldsymbol{\theta}_1^{(m+1)}, \ldots, \boldsymbol{\theta}_{Q-1}^{(m+1)}, \mathbf{y}\Big).
\end{aligned}
$$

The Gibbs sampler manages the transition process, and the form of the conditional densities and the choice of blocking characterizes each sampler. Under some regularity conditions, the MCMC chain has a stationary density equal to $p(\boldsymbol{\theta} \mid \mathbf{y})$. This means that $\boldsymbol{\theta}^{(m)}$ converges to $\boldsymbol{\theta}$ in distribution for $m \to \infty$ (see, e.g., Tierney, 1994).

The regularity conditions can be summarized as follows. The MCMC chain needs to satisfy three assumptions. First, the chain is irreducible; that is, it can reach every nonempty set with positive probability. Each state in the sample

space can be reached from any other state by repeatedly sampling from the conditionals as described. Second, the chain is aperiodic. It is not allowed to be periodic such that it can move between states in regular periodic movements. Third, the chain is positive recurrent. A chain is positive recurrent when an initial value is distributed according to the target distribution and then all subsequent sampled values are distributed according to it (see, e.g., Roberts, 1995).

Metropolis-Hastings

The Metropolis-Hastings (M-H) algorithm was developed by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and generalized by Hastings (1970). The M-H algorithm generalizes the Gibbs sampler since it offers a solution to the problem of sampling from a conditional distribution, from which it is difficult to sample directly. In a two-step procedure, the surface of the posterior of interest is explored via a Markov chain. In the first step, a candidate is drawn from a proposal density that may be chosen to approximate the desired posterior density. In the second step, this draw is accepted or rejected based on a specified acceptance criterion. That is, an acceptance probability is constructed from the posterior density ratio to evaluate the candidate with the last accepted draw. Simply stated, a candidate with high (low) posterior density is likely to be accepted (rejected). The acceptance probability ensures that the algorithm generates samples from the target density. The proposal density is allowed to depend on the current state $\boldsymbol{\theta}^{(m)}$.

Technically, at iteration m, a candidate, $\boldsymbol{\theta}_q^*$, is drawn from a proposal density, $q\big(\boldsymbol{\theta}_q \mid \boldsymbol{\theta}_q^{(m)}\big)$, and the transition from $\boldsymbol{\theta}_q^{(m)}$ to $\boldsymbol{\theta}_q^{(m+1)} = \boldsymbol{\theta}_q^*$ is made if

$$u^{(m)} \leq \frac{p\big(\boldsymbol{\theta}_q^* \mid \mathbf{y}\big)/q\big(\boldsymbol{\theta}_q^* \mid \boldsymbol{\theta}_q^{(m)}\big)}{p\big(\boldsymbol{\theta}_q^{(m)} \mid \mathbf{y}\big)/q\big(\boldsymbol{\theta}_q^{(m)} \mid \boldsymbol{\theta}_q^*\big)}, \tag{3.1}$$

where $u^{(m)}$ is the mth observation from random variable U, which is uniformly distributed on the interval $[0, 1]$. If the proposed value is not accepted, the chain remains at its current state. The right-hand side of (3.1) is known as the acceptance ratio. The algorithm is very simple, and the shape of the proposal density is not restricted. However, there is a strong relationship between the proposal distribution and the target distribution with respect to the rate of convergence to the target distribution. Two special cases can be considered. First, when the proposal density is symmetric, $q\big(\boldsymbol{\theta}_q^* \mid \boldsymbol{\theta}_q^{(m)}\big) = q\big(\boldsymbol{\theta}_q^{(m)} \mid \boldsymbol{\theta}_q^*\big)$, the terms cancel out in the ratio defined in (3.1) and the algorithm of Metropolis et al. (1953) is obtained. Second, when the proposal density is not dependent on former values, $q\big(\boldsymbol{\theta}_q \mid \boldsymbol{\theta}_q^{(m)}\big) = q\big(\boldsymbol{\theta}_q\big)$, the Metropolis independent chain (Tierney, 1994) is obtained.

It is efficient to have an acceptance ratio close to one, but the proposed values should also cover the entire range of probable values under the target distribution. That is, proposals are to be drawn from the entire region

where $p(\boldsymbol{\theta} \mid \mathbf{y})$ is appreciable. In most applications, it is efficient to have a proposal distribution with slightly heavier tails than the target distribution. However, when the proposal distribution is too diffuse, many proposed values are rejected and convergence to the target distribution is very slow. When the proposal distribution has less heavy tails, the proposed values can be restricted to a subregion of the entire sample space where $p(\boldsymbol{\theta} \mid \mathbf{y})$ is appreciable, leading not only to slow convergence but even biased results. This fine-tuning of the proposal distribution is very important since it highly affects the convergence of the algorithm. In high-dimensional problems, it is often helpful to explore the effects of different proposals and their parameter choices. This is useful for establishing fast convergence but also for rapid mixing of the chain through the parameter space.

The M-H algorithm can be used within a Gibbs sampler, also denoted as M-H within Gibbs. This retains the idea of sequential sampling but M-H steps are used on some variables rather than attempting to sample from the exact conditional distribution, which might not be feasible. This way, a hybrid algorithm is defined where the sampled values converge in distribution to the target.

The M-H algorithm is useful when a proposal distribution can be found from which it is easy to simulate that leads to accepted values and an acceptable convergence rate and does not complicate computation of the acceptance ratio. Note that the target density is only present in the algorithm via the acceptance ratio; therefore, it suffices to evaluate $p(.)$ in the acceptance ratio (3.1) up to a normalizing constant.

Issues in MCMC

The iterative nature of an MCMC algorithm presents two problems. First, it must be decided when the algorithm has reached convergence after a finite number of iterations. Second, the MCMC samples drawn are dependent in the sense that values close to one another in the sequence are more alike than values that are far apart. The latter leads to problems in estimating the error variance of an MCMC estimate.

Assessing convergence can be done via a single chain or multiple chains. The subject of whether inferences are to be made from one long chain or multiple chains has received much attention (e.g, Brooks and Gelman, 1998; Cowles and Carlin, 1996; Gelman, 1995; Geyer, 1992; Raftery, 1995). A single chain can be more precise since fewer burn-in iterations are discarded and multiple runs that are too short cannot be used for inference. The use of parallel chains reduces the variability and dependency on the initial values. With parallel chains it is often easier to establish convergence by comparing quantities of different chains. When the cost of running multiple chains is not that expensive using multiple computers or parallel processors, it is the most simple and most reliable way to go.

Single Chain Analysis

The number of iterations before a chain converges to the stationary distribution (burn-in period) depends on several factors, such as the starting values or starting distribution, amount of missing data, identification rules, implementation strategies, and rate of convergence. Various methods have been proposed to establish the burn-in time. Computing theoretical upper bounds on burn-in time is difficult for relatively simple problems (e.g., Roberts and Tweedie, 1999; Rosenthal, 1995). The usefulness of these methods is rather limited in the case of more complex models due to the computational burden that is involved. Therefore, most users apply the more easy to use convergence diagnostics to sampler output. An overview of MCMC diagnostics can be found in Cowles and Carlin (1996).

The most popular way is to plot the iterates of the parameters from the simulation runs and monitor trends. The plot is used to detect abnormal or nonstationary behavior of the chain and investigate the convergence of the empirical average and the mixing of the chain by looking for trends. Plots of successive draws are so-called trace plots or time-series plots, and they may be misleading when the chain is slowly mixing. The autocorrelations between successive iterates can be monitored to investigate their relationships (Equation 3.3). A slow convergence to stationarity often corresponds to highly correlated samples. Note that trace and autocorrelation plots can only provide indications of convergence. After establishing convergence with respect to some model parameters, it is still possible that the algorithm has not converged with respect to some components or functions of model parameters that were not examined. It is also possible that the algorithm has an inadequate opportunity to visit some part of the parameter space due to multiple modes and oddly shaped posteriors.

The BOA (Smith, 2010) and CODA (Best, Cowles and Vines, 2010) software for diagnosing convergence of the distribution of the iterates to the stationary distribution include most common methods (Gelman and Rubin, 1992; Geweke, 1992; Heidelberg and Welch, 1983; Raftery and Lewis, 1992).

The next step is establishing convergence of posterior summaries like the marginal mean and variance. Let $\theta^{(m)}$, $m = 1, \ldots, M$, denote the output from one simulation run. This sample can be used to estimate the posterior mean. That is,

$$\hat{\theta}_M = \sum_{m=1}^{M} \theta^{(m)} \Big/ M \tag{3.2}$$

approximates the posterior mean given dependent samples from the posterior. The central limit theorem ensures that $\hat{\theta}_M$ converges to the (existing) expected value of θ under the target distribution if $M \to \infty$ (Tierney, 1994). Furthermore, the same holds for any real-valued function of θ. The population density can be estimated via a histogram of the sampled values. Although most summaries that are appropriate for an independent random sample are

appropriate for a dependent sample, the variance of the sample average in Equation (3.2) is not the sample variance

$$\frac{1}{M-1}\sum_m \left(\theta^{(m)} - \hat{\theta}_M\right)^2$$

when the sampled values are highly correlated. A variety of methods have been proposed to estimate the variance of an MCMC estimate. Geyer (1992) proposed several time-series methods for estimating the variance. Moreover, Chen et al. (2000) reviewed several approaches. Estimating the variance from a single chain is also possible by resampling from the sampled values to obtain an approximately independent subsample. This procedure is not very efficient, and when M iterations are made it is better to summarize the results based on the M iterates since the average of M iterates is more precise than the average of a related subsample (e.g., MacEachern and Berliner, 1994).

The sample autocorrelation at lag h is defined as

$$r_h = \frac{\sum_{m=h+1}^{M}\left(\theta^{(m)} - \hat{\theta}_M\right)\left(\theta^{(m-h)} - \hat{\theta}_M\right)}{\sum_{m=h+1}^{M}\left(\theta^{(m)} - \hat{\theta}_M\right)^2}. \tag{3.3}$$

The correlations between adjacent iterates of an MCMC sequence can be monitored using the sample autocorrelation estimates of different orders, where the iterates from the burn-in period are ignored.

Multiple Chain Analysis

A popular quantitative convergence diagnostic is the method of Gelman and Rubin (1992). In that case, R chains of length M are simulated and iterate m from run r is denoted as $\theta^{(r,m)}$. The idea is to establish convergence by comparing the between-run variation, σ_b, with the (average) within-run variation, σ_w, where

$$\hat{\sigma}_b = \frac{M}{R-1}\sum_r \left(\hat{\theta}_{rM} - \hat{\theta}_M\right)^2,$$

$$\hat{\sigma}_w = \frac{1}{R(M-1)}\sum_r\sum_m \left(\theta^{(r,m)} - \hat{\theta}_{rM}\right)^2.$$

Here, $\hat{\theta}_{rM}$ denotes the sample average of M iterates from run r and $\hat{\theta}_M$ the sample average across the R runs. An estimate of the variance of the sample average is obtained via a weighted average of σ_w and σ_b and equals

$$\hat{\sigma} = \frac{M-1}{M}\hat{\sigma}_w + \frac{\hat{\sigma}_b}{M}.$$

This estimate is an unbiased variance estimate of the posterior mean of θ when the sampled values are obtained from the target distribution.

Gelman and Rubin (1992) considered the variance ratio term $\hat{\sigma}/\sigma_w$ (scale reduction factor) as an MCMC convergence diagnostic. For a large value of the scale reduction factor and overdispersed starting values relative to the target distribution, it is concluded that the between-run variation is much larger than the within-run variation, and the between-run variation can be decreased or the within-run variation can be increased by further simulations. When the scale reduction factor is close to one, it is concluded that the simulated values are close to being distributed according to the target distribution. At that point, the within-run variance dominates the between-run variance, and it is argued that the chains are no longer influenced by their starting values and have traversed all of the target distribution.

The procedure has some limitations. First, the procedure relies on a normal approximation to the posterior, which might be inadequate. Second, it can be difficult to find starting values or a starting distribution that is indeed overdispersed with respect to the target distribution. Knowledge of the target distribution is needed to verify that the starting values are overdispersed. Third, running multiple chains is very inefficient when from each run a substantial number of iterations are discarded. For example, if one compares 10 single runs of 100 iterations with a single long run of 1,000 iterations, then disregarding the first 50 draws eliminates 500 draws from the multiple chains and only 50 draws of the single long chain. Moreover, the last 900 simulated values of the single long run are probably closer to the target distribution than any of the values from the multiple chains.

A beneficial effect of running multiple chains is that it enables the computation of a variance estimate. The variance term of any statistic computed from output of a single run can be estimated when this single run is replicated R times. The approach is essentially univariate, but a multivariate extension has been proposed by Brooks and Gelman (1998).

3.2 Bayesian Hypothesis Testing

Bayesian inference involves specifying a hypothesis and collecting evidence that supports or does not support the statistical hypothesis. The amount of evidence can be used to specify the degree of belief in a hypothesis in probabilistic terms. With enough evidence, the probability of supporting the hypothesis can become very high or low. Hypotheses with a high degree of belief in probabilistic terms are accepted as true, and those with low belief are rejected as false.

The problem of comparing alternative hypotheses has received a lot of attention and remains a subject of much discussion. A conflict in the classical approach is that a frequentist's one-sided hypothesis test uses a probability (p-value) computed on all possible observations for a fixed parameter value, although the hypothesis is about a parameter value being restricted to a certain interval. Furthermore, the probability that the observations belong to

a certain class is calculated using the null distribution. That is, attention is focused not only on the observations that are made but also on others that might have been made. Jeffreys (1961) remarked that in this sense the use of a p-value implies that a hypothesis can be rejected because of a lack of predicted observations that have not occurred. Press (2003, pp. 220–224) gives a complete overview of problems with frequentist methods for testing. These peculiarities are avoided in a Bayesian approach, but other issues arise. The probabilities of various hypotheses are computed given the data. However, an exact hypothesis has zero probability since a continuous prior assigns zero probability to any exact value. Furthermore, prior choices can highly influence the estimation of posterior probabilities of the hypotheses and the associated decision.

In this section, a short overview of Bayesian hypothesis testing is given. Several situations will be looked at: one-sided and two-sided hypotheses, point null or precise hypotheses, and the Bayes factor. A substantial amount of literature can be considered for a more thorough treatment of hypothesis testing from a Bayesian point of view (see, e.g., Berger and Delampady, 1987; Berger and Selke, 1987; Jeffreys, 1961; Lee, 2004; Lindley, 1965; Zellner, 1971).

Let H_0 represent a hypothesis, $\boldsymbol{\theta} \in \Theta_0$, called a null hypothesis with prior probability π_0 and H_1, $\boldsymbol{\theta} \in \Theta_1$, the alternative hypothesis with prior probability $\pi_1 = 1 - \pi_0$ such that $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The Bayesian framework allows researchers to place probabilities on competing hypotheses.

Let $p(\mathbf{y} \mid H_0)$ denote the posterior density of $\mathbf{y}$ under the null hypothesis. The posterior density $p(\mathbf{y})$ is the marginal density of the data under all mutually exclusive hypotheses. $P(H_0 \mid \mathbf{y})$ is called the posterior probability of H_0 given the data. This probability will be large when it is likely that the observations are made when the null hypothesis under consideration is true. The prior probabilities on the hypotheses are updated via Bayes' theorem given the observed data, which leads to the posterior probabilities of the hypotheses. For example, the posterior probability of the null hypothesis can be expressed as

$$\begin{aligned} P(H_0 \mid \mathbf{y}) &= \frac{\pi_0 p(\mathbf{y} \mid H_0)}{\pi_0 p(\mathbf{y} \mid H_0) + \pi_1 p(\mathbf{y} \mid H_1)} \\ &= \frac{\pi_0 \int p(\mathbf{y} \mid \boldsymbol{\theta}, H_0) p(\boldsymbol{\theta} \mid H_0) d\boldsymbol{\theta}}{\pi_0 p(\mathbf{y} \mid H_0) + \pi_1 p(\mathbf{y} \mid H_1)} \\ &= \frac{\pi_0 \int_{\boldsymbol{\theta} \in \Theta_0} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}{\pi_0 \int_{\boldsymbol{\theta} \in \Theta_0} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} + \pi_1 \int_{\boldsymbol{\theta} \in \Theta_1} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}, \end{aligned} \tag{3.4}$$

where it is assumed that the conditional prior $p(\boldsymbol{\theta} \mid H_0)$ and $p(\boldsymbol{\theta} \mid H_1)$ equal $p(\boldsymbol{\theta})$ restricted to $\boldsymbol{\theta} \in \Theta_0$ and $\boldsymbol{\theta} \in \Theta_1$, respectively.

For a one-sided hypothesis test, Equation (3.4) can be reduced to the posterior probability of the null hypothesis being true. Therefore, define $g(\boldsymbol{\theta})$ as the unnormalized prior,

$$p(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}\in\Theta_0} g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} = \frac{g(\boldsymbol{\theta})}{P\big(\boldsymbol{\theta}\in\Theta_0\big)}.$$

The prior probability of the null hypothesis equals $P(H_0) = \pi_0 = P(\boldsymbol{\theta} \in \Theta_0)$. As a result,

$$\begin{aligned} P\big(H_0 \mid \mathbf{y}\big) &= \frac{\int_{\boldsymbol{\theta}\in\Theta_0} p(\mathbf{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}\in\Theta_0} p(\mathbf{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} + \int_{\boldsymbol{\theta}\in\Theta_1} p(\mathbf{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \\ &= \frac{\int_{\boldsymbol{\theta}\in\Theta_0} p(\boldsymbol{\theta} \mid \mathbf{y}) p(\mathbf{y})\,\mathrm{d}\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}\in\Theta} p(\mathbf{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \\ &= \int_{\boldsymbol{\theta}\in\Theta_0} p(\boldsymbol{\theta} \mid \mathbf{y})\,\mathrm{d}\boldsymbol{\theta}. \end{aligned}$$

The null hypothesis is rejected when this posterior probability is less than the significance level. A one-sided hypothesis can be tested by computing its posterior probability using the posterior distribution of $\boldsymbol{\theta}$. In the same way, a Type I error (reject when H_0 is true) and a Type II error (do not reject when H_0 is false) can be calculated given the posterior distribution of $\boldsymbol{\theta}$.

Multiple hypothesis testing can be done in a similar manner by assigning prior probabilities to each hypothesis and calculating the corresponding posterior probabilities (see, e.g., Zellner, 1997, pp. 364–382).

A point null hypothesis can be stated as $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus H_1 $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. In practice, the hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is never exactly the case due to, for example, minor biases in the experiment. Therefore, a precise hypothesis is often better represented as $H_0 : \boldsymbol{\theta} \in \Theta_0$ where $\Theta_0 = (\boldsymbol{\theta}_0 - \epsilon, \boldsymbol{\theta}_0 + \epsilon)$ versus $H_1{:}\boldsymbol{\theta} \notin \Theta_0$ where ϵ is close to zero. This null hypothesis is approximated by the point null hypothesis when the posterior probabilities are close.

A continuous prior density cannot be used since it assigns zero prior (and posterior) probability to $\boldsymbol{\theta}_0$. Therefore, a positive prior probability π_0 is given to $\boldsymbol{\theta}_0$ and a continuous prior $\pi_1 p(\boldsymbol{\theta})$ is defined for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\pi_1 = 1 - \pi_0$. Subsequently, the posterior probability of the null hypothesis equals

$$\begin{aligned} P(H_0 \mid \mathbf{y}) &= \frac{\pi_0 p(\mathbf{y} \mid \boldsymbol{\theta}_0)}{\pi_0 p(\mathbf{y} \mid \boldsymbol{\theta}_0) + \pi_1 \int_{\boldsymbol{\theta}\neq\boldsymbol{\theta}_0} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \\ &= \frac{\pi_0 p(\mathbf{y} \mid \boldsymbol{\theta}_0)}{\pi_0 p(\mathbf{y} \mid \boldsymbol{\theta}_0) + (1 - \pi_0) p_1(\mathbf{y})} \qquad (3.5) \\ &= \frac{\pi_0 p(\mathbf{y} \mid \boldsymbol{\theta}_0)}{p(\mathbf{y})}. \qquad (3.6) \end{aligned}$$

The Bayes factor (see, e.g., Berger, 1985; Jeffreys, 1961; Kass and Raftery, 1995) is defined as the ratio of the posterior probabilities of the null and alternative hypotheses over the ratio of prior probabilities of the null and

alternative hypotheses. The Bayes factor in favor of the null can be constructed from prior and posterior odds on H_0 against H_1,

$$\frac{P(H_0 \mid \mathbf{y})}{P(H_1 \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid H_0)}{p(\mathbf{y} \mid H_1)} \frac{P(H_0)}{P(H_1)},$$

and the Bayes factor equals

$$BF = \frac{P\left(H_0 \mid \mathbf{y}\right)}{P\left(H_1 \mid \mathbf{y}\right)} \Big/ \frac{P(H_0)}{P(H_1)} = \frac{p\left(\mathbf{y} \mid H_0\right)}{p\left(\mathbf{y} \mid H_1\right)}.$$

In this respect, Equation (3.5) can be presented in such a way that it contains the Bayes factor. Then, the posterior probability of the null hypothesis can be found from its prior probability and the Bayes factor,

$$\begin{aligned} P(H_0 \mid \mathbf{y}) &= \frac{\pi_0 p(\mathbf{y} \mid \boldsymbol{\theta}_0)}{\pi_0 p(\mathbf{y} \mid \boldsymbol{\theta}_0) + (1-\pi_0) p_1(\mathbf{y})} \\ &= \left(1 + \frac{1-\pi_0}{\pi_0} \frac{p_1(\mathbf{y})}{p(\mathbf{y} \mid \boldsymbol{\theta}_0)}\right)^{-1} \\ &= \left(1 + \frac{1-\pi_0}{\pi_0} BF^{-1}\right)^{-1}. \end{aligned}$$

The Bayes factor summarizes the evidence provided by the data. A value of $BF = 2$ means that the null hypothesis is supported by the data two times as much as the alternative hypothesis. Values of the Bayes factor are often evaluated on the $\log_{10}$ scale (Jeffreys, 1961), and the evidence against the null is called decisive if $\log_{10}(BF)$ is less than $1/2$.

The Bayes factor is usually influenced by prior information that is neither vague nor diffuse, but hypothesis testing via a Bayes factor is only possible using proper priors (integrate to unity). Modifications of the Bayes factor have been proposed to accommodate it for improper priors. Most of them are based on training data. Examples are the partial Bayes factor, intrinsic Bayes factor of Berger and Pericchi (1996), and fractional Bayes factor of O'Hagan (1995). Furthermore, a sensitivity analysis is needed to investigate influences of prior choices on the outcomes of the Bayes factor. Sinharay and Stern (2002) showed that different prior distributions for the model parameters lead to different Bayes factors and a sensitivity analysis is needed before conclusions can be drawn.

3.2.1 Computing the Bayes Factor

The computation of the Bayes factor requires numerical methods for evaluating the integrals. Typically, when an implemented MCMC algorithm is available, methods for computing the Bayes factor are based on the available posterior draws. Different identities exist that express the Bayes factor as an expectation of quantities such that available draws from the posteriors can be used. Here, a few techniques will be described. Kass and Raftery (1995) and Rossi et al. (2005), among others, reviewed and compared different methods.

Importance Sampling

Importance sampling is often used to estimate the value of integrals using approximate samples from complex high-dimensional distributions. The importance sampling method is especially useful in the computation of normalizing constants; for example, in the computation of marginal likelihoods, the Bayes factor, and likelihood inference. For a complete overview, see, for example, Geweke (2005), Robert and Casella (1999), and references therein.

Under a hypothesis, say H_0, the marginal likelihood of the data can be expressed as

$$\begin{aligned} p\left(\mathbf{y} \mid H_0\right) &= \int_{\mathcal{R}_\theta} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{\int_{\mathcal{R}_\theta} p(\mathbf{y} \mid \boldsymbol{\theta}) w(\boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\mathcal{R}_\theta} w(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}, \end{aligned} \tag{3.7}$$

where $w(\boldsymbol{\theta}) = p(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ are known as importance ratios or importance weights. The denominator in (3.7) appears when $w(\boldsymbol{\theta})$ is only known up to a normalizing constant. Here, $q(.)$ is the importance sampling density function from which samples can be drawn. The integral is evaluated using samples $\boldsymbol{\theta}^{(m)}$ from the importance sampling density:

$$\hat{p}\left(\mathbf{y} \mid H_0\right) = M^{-1} \sum_m p\left(\mathbf{y} \mid \boldsymbol{\theta}^{(m)}\right) w\left(\boldsymbol{\theta}^{(m)}\right). \tag{3.8}$$

When the importance sampling function is known up to a constant, the estimate on the right-hand side of (3.8) is divided by $\sum_m w(\boldsymbol{\theta}^{(m)})/M$ to take account of the normalizing constant. The choice of the importance sampling function highly influences the accuracy of the outcome. Poor results are obtained when the importance ratios are small with high probability and large with low probability. This occurs when the integrand has wide tails compared with the importance sampling density. To compute the Bayes factor, the marginal likelihood needs to be estimated under both hypotheses, each with a separate importance density.

Newton and Raftery (1994) proposed a specific importance sampling function to evaluate the marginal likelihood of the data. They suggested $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y})$ since samples $\boldsymbol{\theta}^{(m)}$ from this posterior are often easily obtained via an MCMC sampler. As a result,

$$\hat{p}(\mathbf{y}) = \left[\frac{1}{M} \sum_m \frac{1}{p\left(\mathbf{y} \mid \boldsymbol{\theta}^{(m)}\right)} \right]^{-1} \tag{3.9}$$

is considered to be an estimate of the marginal likelihood of the data. This harmonic mean of likelihood values is easily computed given sampled values but can be unstable since the inverse likelihood does not have a finite variance.

Using Identities and MCMC Output

Gelfand and Dey (1994) used an identity for estimating the marginal likelihood that leads to

$$\begin{aligned} p(\mathbf{y})^{-1} &= \int_{\mathcal{R}_\theta} \frac{f(\boldsymbol{\theta})}{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})} p(\boldsymbol{\theta} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta} \\ &= \lim_{M \to \infty} M^{-1} \sum_m \frac{f\left(\boldsymbol{\theta}^{(m)}\right)}{p\left(\mathbf{y} \mid \boldsymbol{\theta}^{(m)}\right) p\left(\boldsymbol{\theta}^{(m)}\right)}, \end{aligned} \tag{3.10}$$

where $f(\boldsymbol{\theta})$ is a proper function having support contained in $\mathcal{R}_\theta$. In the case where $f(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, the right-hand-side term in (3.10) resembles the right-hand-side term in (3.9). A stable estimate of the marginal likelihood can be obtained via a stability function $f(.)$. Geweke (2005) showed how to construct a probability density function $f(.)$, a truncated multivariate normal density, from the generated samples $\boldsymbol{\theta}^{(m)}$.

The bridge sampling estimator of Meng and Wong (1996) is based on draws from the posterior and an importance density function $q(\boldsymbol{\theta})$. They provide an identity that contains an arbitrary function $f(\boldsymbol{\theta})$ that can be used to estimate the marginal likelihood,

$$\hat{p}\left(\mathbf{y}\right) = \frac{M_q^{-1} \sum f\left(\boldsymbol{\theta}^{(m_q)}\right) p\left(\boldsymbol{\theta}^{(m_q)} \mid \mathbf{y}\right)}{M^{-1} \sum f\left(\boldsymbol{\theta}^{(m)}\right) q\left(\boldsymbol{\theta}^{(m)} \mid \mathbf{y}\right)},$$

where $\boldsymbol{\theta}^{(m_q)}$ $m_q = 1, \ldots, M_q$ are drawn from the importance density $q(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^{(m)}$ are drawn from the posterior density. The method of Gelfand and Dey (1994) and importance sampling are special cases of bridge sampling where the function $f(.)$ equals the importance density function or the unnormalized posterior, respectively.

For nested hypotheses, a useful identity exists that simplifies the computation of the Bayes factor. Let $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ be the null hypothesis H_0 and $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ the alternative hypothesis H_1. It follows that

$$p\left(\mathbf{y} \mid \boldsymbol{\xi}, H_0\right) = p\left(\mathbf{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\xi}, H_1\right)$$

due to the nested nature of the hypotheses. Furthermore, suppose that

$$p\left(\boldsymbol{\xi} \mid H_0\right) = p\left(\boldsymbol{\xi} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0, H_1\right). \tag{3.11}$$

Then, it can be proven (see Exercise 6.3) that the Bayes factor equals

$$\begin{aligned} BF &= \frac{p\left(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid \mathbf{y}, H_1\right)}{p\left(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid H_1\right)} \\ &= \lim_{M \to \infty} M^{-1} \sum_m \frac{p\left(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid \boldsymbol{\xi}^{(m)}, \mathbf{y}, H_1\right)}{p\left(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid H_1\right)}, \end{aligned} \tag{3.12}$$

where $\boldsymbol{\xi}^{(m)}$ are sampled from $p(\boldsymbol{\xi} \mid \mathbf{y}, H_1)$. Equation (3.12) is referred to as the Savage-Dickey density ratio. It is a widely used tool for estimating the Bayes factor, but the priors under both hypotheses need to be conditionally linked as in Equation (3.11). Verdinelli and Wasserman (1995) proposed a generalization when the priors do not satisfy Equation (3.11).

Bayes Factor for Item Response Models

Response data provide information about the characteristics of the items and the respondents in terms of abilities that are measured. Selecting more respondents will improve the item parameter estimates but will automatically increase the dimensionality of the ability parameter space. In the same way, adding more items will improve the precision of the ability estimates but will increase the dimensionality of the item parameter space. In both ways, sampling more respondents and/or increasing the item set will enlarge the dimensionality of the parameter space. Typically, the complexity of the problem of computing the marginal likelihood is getting more complex when observing more data. The unnormalized posterior density function that needs to be integrated increases in dimension when observing more response data.

The discrete nature of the response data leads to a nonlinear likelihood model. The joint and marginal posterior distributions of the item and person parameters have unknown analytical properties in a high-dimensional parameter space. This highly complicates the computation of a marginal likelihood.

An asymptotic method for computing the Bayes factor is based on the assumption that the unnormalized posterior converges to a normal distribution when increasing the number of observations. Then, a Taylor expansion of the unnormalized posterior around the posterior mode is considered to be an asymptotic approximation of the marginal likelihood. The Bayesian information criterion (BIC; Schwarz, 1978) is based on an asymptotic approximation to the marginal likelihood. Subsequently, twice the log of the Bayes factor is approximately equal to the difference in the BIC. That is, the ΔBIC approximates $2 \log BF$ (see, e.g., Raftery, 1995) and is defined as

$$\Delta BIC = 2 \log \left(\frac{p(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, H_0)}{p(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, H_1)} \right) - (p_2 - p_1) \log n, \tag{3.13}$$

where p_1 and p_2 are the number of parameters under H_0 and H_1, respectively, n the number of observations, and $\hat{\boldsymbol{\theta}}$ the value that maximizes $p(\mathbf{y} \mid \boldsymbol{\theta})$.

However, the asymptotic approximation does not hold for item response models where the number of model parameters increases with the number of observations. Typically, the asymptotic approximation only holds when increasing the number of observations while holding the dimension of the parameter space constant. In response modeling, the BIC is known to be extremely imprecise and is not recommended for estimating a Bayes factor.

Another problem is that the within-individual and between-respondent modeling structure leads to a hierarchical model that complicates the specification of the number of free parameters. For example, a tight prior constrains the model and a flat prior corresponds to a more flexible model. The implied prior structure is often needed to handle the high-dimensional parameter space but often it is not clear in what way the prior reduces the number of free parameters.

Accurate computation of the Bayes factor for complex nonlinear response models with an interest in respondent-level and item-level parameters is very difficult. The posterior distributions that comprehend information at different levels are high-dimensional objects and are not members of a known class of distributions. This complicates the computation of the marginal likelihood, which requires integrating out the parameters in the joint posterior density. Below, for the class of response models where the Bayes factor is difficult to compute or where improper priors are used, alternative ways of testing hypotheses are explored.

3.2.2 HPD Region Testing

Box and Tiao (1973) developed a procedure for Bayesian testing of significance that allows the use of diffuse priors. In this situation, interest is focused on a specific null hypothesis. Typical examples are testing the equality of population means and testing whether a regression coefficient equals zero. In both cases, the prior distribution under the null hypothesis that $\theta = \theta_0$ is flat such that it is assumed that values near θ_0 are as likely as $\theta = \theta_0$. The procedure is based on a Bayesian confidence interval for a significance level α given the (unimodal) posterior density $p(\theta \mid \mathbf{y})$. Then, the null hypothesis is rejected when θ_0 falls outside this interval. The procedure is based on the fact that the density function for θ can be used to investigate whether the region where $\theta = \theta_0$ has high posterior density. If the region where $\theta = \theta_0$ has low posterior density, then this value for θ is not likely and the null hypothesis is rejected. This way of testing hypotheses is also known as highest posterior density (HPD) interval testing or, in a multidimensional setting, HPD region testing.

The interval method of testing hypotheses is based on the construction of a Bayesian confidence interval. In the literature, a Bayesian confidence interval (e.g., Box and Tiao, 1973; Lindley, 1965) is usually referred to as a credible set or a credible interval. A credible set for a scalar parameter θ (continuous valued) presents the set of values such that θ lies within this set with probability $1-\alpha$. The posterior density is used to quantify the probability that θ lies in a credible set. That is, a $100(1-\alpha)\%$ credible interval is a subset $\mathcal{C} \subset \Theta$ for θ such that

$$P(\mathcal{C} \mid \mathbf{y}) = \int_{\mathcal{C}} p(\theta \mid \mathbf{y}) \, \mathrm{d}\theta = 1 - \alpha. \tag{3.14}$$

An important defect of such a credible set is that it does not specify whether values of θ within the set are more probable than values outside the set. It is preferable to choose a credible set of values with the highest posterior density. Therefore, an extension of the credible set is the highest posterior density (HPD) interval that covers at least a probability of $1 - \alpha$ and contains the most likely values of θ,

$$P(\mathcal{C} \mid \mathbf{y}) \geq 1 - \alpha, \tag{3.15}$$

and for $\theta_1 \in \mathcal{C}$, $\theta_2 \notin \mathcal{C}$ it follows that $p(\theta_1 \mid \mathbf{y}) \geq p(\theta_2 \mid \mathbf{y})$. The posterior density of every point inside the HPD interval is greater than that for every point outside the interval.

For a multidimensional parameter $\boldsymbol{\theta}$, the credible set is defined as a credible region rather than a credible interval. In the same way, the HPD region has the property that it has the smallest possible volume in the parameter space of $\boldsymbol{\theta}$. For a symmetric unimodal posterior distribution, the HPD region equals an equal-tail credible set that is defined by taking the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution as the credible region with probability $1 - \alpha$. Note that a specific point $\boldsymbol{\theta}_0$ is inside the HPD region if and only if

$$P\big(p(\boldsymbol{\theta} \mid \mathbf{y}) > p(\boldsymbol{\theta}_0 \mid \mathbf{y}) \mid \mathbf{y}\big) \leq 1 - \alpha, \tag{3.16}$$

where $p(\boldsymbol{\theta} \mid \mathbf{y})$ is treated as a random variable. The expression (3.16) becomes especially useful when it is possible to evaluate the probability content of the distribution directly. For the HPD region method of hypothesis testing, Box and Tiao (1973) specified an event directly or in terms of some monotonic function in such a way that the probability that this event belongs to a particular (HPD) region is easily computed.

Berger and Delampady (1987) argued that concentrating only on intervals is not correct since it ignores the special nature of θ_0. A specific point outside an interval often has a likelihood that is not much smaller than the average likelihood of θ within the region. Then, there is not much support for rejecting the null hypothesis. Credible regions often correspond to diffuse prior distributions and are not appropriate in the case of a single special value. The Bayes factor is used to judge whether a specific point is supported by the data, specifically when H_0 is precise. However, the main advantage of credible or HPD regions is that they display the actual gap between θ and θ_0. Besides the Bayes factor, HPD regions remain important since they indicate the magnitude of the possible discrepancy. Furthermore, HPD regions may prove worthwhile in situations where prior knowledge is vague, where the Bayes factor is difficult to compute, or where the Bayes factor is highly influenced by prior choices.

3.2.3 Bayesian Model Choice

The problem of model choice differs from that of testing since it holds more inferential goals and more demanding computational tasks than mere testing.

The Bayes factor can still be used for model selection problems but depends on prior densities for the parameters. Other criteria such as the BIC and deviance information criterion (DIC; Spiegelhalter, Best, Carlin and van der Linde, 2002) do not have to make a reference to prior densities for model parameters. These selection methods become particularly interesting when computation of the Bayes factor is (almost) impossible. Both methods are based on a measure of fit and some penalty function based on the number of free parameters for the complexity of the model. A bias–variance trade-off exists between these two quantities since a more complex model often leads to a better fit but a less complex model involves more accurate estimation.

In a setting where there are two competing models denoted as $\mathcal{M}_1$ and $\mathcal{M}_2$ that have respective parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the object is to select the best model for the observed data $\mathbf{y}$. When prior probabilities π_i $(i = 1, 2)$ are specified for both models, the posterior probability of model $\mathcal{M}_i$ equals

$$P(\mathcal{M}_i \mid \mathbf{y}) = \frac{\pi_i \int_{\Theta_i} p(\mathbf{y} \mid \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i) \, \mathrm{d}\boldsymbol{\theta}_i}{\sum_{j=1,2} \pi_j \int_{\Theta_j} p(\mathbf{y} \mid \boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j) \, \mathrm{d}\boldsymbol{\theta}_j},$$

according to Bayes' theorem. Subsequently, the model with the highest posterior density is selected. This procedure is easily extended to more models. The Bayes factor involves the computation of possibly high-dimensional and intractable integrals. Furthermore, the Bayes factor for model selection depends on prior choices, and the sensitivity to them should be investigated.

The BIC defines the change in the conditional density of the data of model $\mathcal{M}_2$ to that of model $\mathcal{M}_1$ plus a penalty term for the extra parameters in model $\mathcal{M}_2$. This difference in size between models can be seen by considering model $\mathcal{M}_2$ to be the saturated or full model and $\mathcal{M}_1$ the reduced model. Subsequently, if the BIC in (3.13) is negative, model $\mathcal{M}_1$ is preferred by the data, and if the BIC is positive, model $\mathcal{M}_2$ fits the data better.

The BIC is a rough but straightforward penalized likelihood ratio method that is easy to compute and enables the direct comparison of nonnested models. However, difficulties may arise in specifying the difference in the number of parameters between models and/or the number of observations. For example, the effective number of parameters in a hierarchical model is often difficult to calculate. Although the nominal number of parameters follows directly from the likelihood, the prior distribution imposes additional restrictions on the parameter space and reduces its effective dimension. In a random effects model, the effective number of parameters depends strongly on the higher-level variance parameters. When the variance of the random effects approaches zero, all random effects are equal and the model reduces to a simple linear model with one mean parameter. But when the variance goes to infinity, the number of free parameters approaches the number of random effects.

Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) for model comparison when the number of parameters is not clearly defined. The DIC is defined as the sum of a deviance measure and a penalty

term for the effective number of parameters based on a measure of model complexity described below. The deviance defined as $D(\boldsymbol{\theta}) = -2\log p(\mathbf{y} \mid \boldsymbol{\theta}) + 2\log p(\mathbf{y})$ is not a suitable discriminating measure between models since it will always prefer the higher-dimensional models. Therefore, a penalizing term based on the estimated number of effective parameters is added to correct for the bias of the deviance towards higher-dimensional models. This term estimates the number of effective model parameters and equals

$$\begin{aligned} p_D &= E\big(-2\log p(\mathbf{y} \mid \boldsymbol{\theta})\big) + 2\log p\big(\mathbf{y} \mid \hat{\boldsymbol{\theta}}\big) \\ &= \overline{D(\boldsymbol{\theta})} - D\big(\hat{\boldsymbol{\theta}}\big), \end{aligned}$$

where the expectation is taken with respect to the pdf of $\boldsymbol{\theta}$ given the data and $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$. Here, $\overline{D(\boldsymbol{\theta})}$ is the posterior mean deviance and $D\big(\hat{\boldsymbol{\theta}}\big)$ the estimated deviance given a posterior estimate of $\boldsymbol{\theta}$. Subsequently, the DIC is defined as

$$\begin{aligned} \text{DIC} &= \overline{D(\boldsymbol{\theta})} + \Big(\overline{D(\boldsymbol{\theta})} - D\big(\hat{\boldsymbol{\theta}}\big)\Big) \\ &= \overline{D(\boldsymbol{\theta})} + p_D \\ &= D\big(\hat{\boldsymbol{\theta}}\big) + 2p_D. \end{aligned}$$

Given a closed form of $D(\boldsymbol{\theta})$, the posterior mean of the deviance can be estimated given simulated values of $D(\boldsymbol{\theta})$ using, for example, MCMC. The term $D\big(\hat{\boldsymbol{\theta}}\big)$ is approximated by plugging in an estimate for $\boldsymbol{\theta}$. Note that for model comparisons it can be assumed that the standardizing factor $p(\mathbf{y})$ equals one such that $D(\boldsymbol{\theta}) = -2\log p(\mathbf{y} \mid \boldsymbol{\theta})$. Finally, the best model is associated with the smallest DIC value.

3.3 Discussion and Further Reading

Bayesian modeling is focused on a likelihood function and priors. They are the building blocks, and statistical inferences are based on the product that defines the posterior of the unknown model parameters. Those interested in a complete introduction are referred to Berger (1985), Bernardo and Smith (1994), Box and Tiao (1973), Gelman et al. (1995), Press (2003), or Zellner (1971).

The computation of high-dimensional posterior densities, and summarizing information from them, is possible mainly through the use of MCMC simulation algorithms. MCMC methods are particularly suited for the hierarchical response models, including cross-classified hierarchies. The hierarchical models are built on conditional distributions and therefore give rise to the construction of an MCMC sampler (e.g., Gelman and Hill, 2007; Leonard and Hsu, 1999; Rossi et al., 2005). For example, the problem of sampling individual-level parameters is simplified by conditioning on the higher-level parameters.

The iterative nature of MCMC methods makes it possible to estimate simultaneously the parameters of complex hierarchical models for cross-classified response data.

More complex models can be constructed by adding layers to the hierarchy, which can be handled by MCMC methods since they just lead to additional sampling steps. A variety of posterior simulation methods have been developed since the 1990s, and a more complete overview can be found in Geweke (2005) and Robert and Casella (1999). In a shortcut approximation of a fully Bayesian analysis of hierarchical models, higher-level parameters are estimated using the data and then plugged into the lower levels. An overview of this empirical Bayes estimation can be found in Carlin and Louis (1996) and Morris (1983).

There is extensive literature about the Bayesian approach to hypothesis testing. An introduction is given by Berger (1985), Jeffreys (1961), and Press (2003). An introduction from a decision-making point of view can be found in DeGroot (1970) and Zellner (1971). In Chapter 5, more attention will be given to model choice and assessment, and prior and posterior predictive methods are also considered.

3.4 Exercises

The exercises provide insight and background information for various statistical results that will be used further on.

3.1. A Bayesian analysis of the normal mean is considered. Let the observations Y_i $(i = 1, \ldots, n)$ be independently normally distributed with mean θ and variance σ^2.
(a) Show that the likelihood function is given by

$$p\left(y_1, \ldots, y_n \mid \theta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(S + n(\hat{\theta} - \theta)^2\right)\right),$$

where $\hat{\theta} = \sum y_i/n$ and $S = \sum_i \left(y_i - \hat{\theta}\right)^2$.
(b) Assume a standard noninformative prior $p\left(\theta, \sigma\right) \propto \sigma^{-1}$, and show that the posterior density $p(\theta \mid \mathbf{y}, \sigma^2)$ is a normal density with mean $\hat{\theta}$ and variance σ^2/n.

3.2. (continuation of Exercise 3.1) The posterior density of θ is constructed from normally distributed sample information $\mathbf{Y}$ and a normal prior with mean θ_0 and variance σ_0^2 (see, e.g., Box and Tiao, 1973; Zellner, 1971).
(a) Derive the posterior density of θ given $\mathbf{y}, \sigma^2$ from

$$\begin{aligned} p(\theta \mid \mathbf{y}, \sigma^2) &\propto p(\mathbf{y} \mid \theta, \sigma^2) p(\theta) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right) \exp\left(-\frac{1}{2\sigma_0^2}(\theta - \theta_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} n(\theta - \hat{\theta})^2 - \frac{1}{2\sigma_0^2}(\theta - \theta_0)^2\right) \end{aligned} \tag{3.17}$$

by factorizing $(\sigma_0^2 + \sigma^2/n)/(2\sigma_0^2\sigma^2/n)$ in (3.17).
(b) Show that $\mathbf{Y}$ and θ are distributed as

$$\begin{pmatrix} \mathbf{Y} \\ \theta \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{1}_n\theta_0 \\ \theta_0 \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{I}_n + \sigma_0^2\mathbf{J}_n & \sigma_0^2\mathbf{1}_n \\ \sigma_0^2\mathbf{1}_n^t & \sigma_0^2 \end{bmatrix} \right)$$

where $\mathbf{1}_n$ is a vector of ones of length n, $\mathbf{I}_n$ the unity matrix of dimension n, and $\mathbf{J}_n$ a matrix of ones of dimension n.
(c) Show via (a) and (b) that the posterior expectation of θ equals

$$E(\theta \mid \mathbf{y}) = \frac{\hat{\theta}n/\sigma^2 + \theta_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2} \tag{3.18}$$

and

$$E(\theta \mid \mathbf{y}) = \theta_0 + \sigma_0^2\mathbf{1}_n^t(\sigma^2\mathbf{I}_n + \sigma_0^2\mathbf{J}_n)^{-1}(\mathbf{y} - \theta_0), \tag{3.19}$$

respectively, and that (3.18) equals (3.19) using the identity

$$\left(\sigma^2\mathbf{I}_n + \sigma_0^2\mathbf{J}_n\right)^{-1} = \sigma^{-2}\left(\mathbf{I}_n - \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2}\mathbf{J}_n\right). \tag{3.20}$$

3.3. The posterior expectation in Equation (3.18) is referred to as the Bayes estimate of θ, where $\hat{\theta}$ is the least squares estimate (Lindley and Smith, 1972).
(a) Show that the Bayes estimate is biased in contrast to the least squares estimate.
(b) Argue that the Bayes estimate is a weighted average, and explain the influence of the variances and the number of observations. (Notice that the Bayes estimate is particularly useful when the least squares estimate is unreliable.)
(c) Explain that the amount of shrinkage can be inferred from the data when defining priors for the variance parameters.

3.4. Let the posterior density of a linear regression parameter β depend on observations $\mathbf{y}$ and explanatory observations $\mathbf{x}$. Assume that samples can be drawn from the posterior $p(\beta \mid \mathbf{y}, \mathbf{x})$.
(a) Explanatory values $\mathbf{x}$ are observed with error, and the measurement error model is defined by $X_i \sim \mathcal{N}(\mu, 1)$. Explain that the marginal posterior of β can be expressed as

$$p(\beta \mid \mathbf{y}) = \int p(\beta \mid \mathbf{y}, \mathbf{x})\, p(\mathbf{x} \mid \mu)\, \mathrm{d}\mathbf{x}.$$

(b) Monte Carlo integration allows the calculation of integrals using random draws. Show how independent samples $\mathbf{x}^{(m)}$ can be used to estimate the posterior expected value $E(\beta \mid \mathbf{y})$.
(c) Assume $\mu \sim \mathcal{N}(0, 1)$, and show that draws $\mu^{(m)} \sim p(\mu)$ and $\mathbf{X}^{(m)} \mid \mu^{(m)} \sim p(\mathbf{x} \mid \mu)$ can be used to estimate the posterior expected value of β.
(d) In a Gibbs sampling scheme, draws are made from the full conditionals. Argue that, depending on the prior choices, a Gibbs sampling algorithm requires fewer iterations to obtain an accurate estimate.

3.5. Assume that realizations of a normally distributed random variable are never observed due to a censoring mechanism. Let $Z_i \sim \mathcal{N}(\theta, 1)$ denote the (unobserved) underlying random variable and let Y_i denote the censored random variable.
(a) Assume that $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ if $Z_i \leq 0$, and show that

$$P(Y_i = 1) = P(Z_i > 0) = \Phi(\theta).$$

(b) Show that the likelihood of the observed data can be expressed as

$$\begin{aligned} p(\mathbf{y} \mid \theta) &= \prod_i \Phi(\theta)^{y_i} (1 - \Phi(\theta))^{1-y_i} \\ &= \prod_i \int p(z_i, y_i \mid \theta) \, \mathrm{d}z_i. \end{aligned}$$

(c) Explain that it is much easier to work with the complete data $(\mathbf{y}, \mathbf{z})$ for making inferences concerning θ. (This motivates a Gibbs sampling algorithm that includes the simulation of latent data $\mathbf{z}$; see Chapter 4.)

3.6. Assume n observations $(y_1, \ldots, y_n)$ from a normally distributed random variable, $Y \sim \mathcal{N}(\theta, 1)$.
(a) Show that the posterior density of θ is normal with mean $\sum_i y_i/n = \overline{y}$ and variance $1/n$ given an improper prior $p(\theta) \propto 1$.
(b) Let U be uniformly distributed on the interval $(0, 1)$, $U \sim \mathcal{U}_{(0,1)}$. Making use of the probability integral transformation,[1] show that a draw θ^* from the posterior density can be obtained via

$$\theta^* = \overline{y} + \frac{1}{\sqrt{n}} \Phi^{-1}(U),$$

where $\Phi^{-1}(.)$ is the inverse cumulative normal distribution function.
(c) Given a standard normal prior, show how to draw θ^* from the posterior density.

3.7. (continuation of Exercise 3.6) Assume that θ is uniformly distributed on the interval (θ_0, θ_1).
(a) Let $f(\theta) = \sqrt{n}(\theta - \overline{y})$. Show that the posterior probability of $\theta \leq \theta^*$ equals

$$P(\theta \leq \theta^* \mid \mathbf{y}) = \frac{\Phi(f(\theta^*)) - \Phi(f(\theta_0))}{\Phi(f(\theta_1)) - \Phi(f(\theta_0))}.$$

(b) Let U be uniformly distributed on $(0, 1)$. Show that a posterior draw θ^* can be obtained via

[1] Let Y have a continuous cumulative distribution function $F(Y)$, and define random variable U as $U = F(Y)$. Then U is uniformly distributed on $(0, 1)$ such that $P(U \leq u) = u$ for $0 < u < 1$ (Casella and Berger, 2002).

$$\theta^* = \overline{y} + \frac{1}{\sqrt{n}}\Phi^{-1}\Big(\big(\Phi(f(\theta_1)) - \Phi(f(\theta_0))\big)U + \Phi(f(\theta_0))\Big),$$

which is known as the inverse sampling method (Ripley, 1987).
(c) Assume that the observations are truncated to the region $\mathcal{R}_y = \{y_i : y_0 < y_i < y_1, i = 1, \ldots, n\}$. Then, the conditional density of y_i is given by

$$p\,(y_i \mid \theta) = \frac{\phi\,(y_i; \theta)}{\int_{y_0}^{y_1} \phi\,(y; \theta)\,\mathrm{d}y}.$$

Show that the parameter of interest θ is also present in the normalizing constant of $p(y_i \mid \theta)$, which hinders an inverse sampling method for simulating values θ^*.
(d) Show that the cumulative probability of $Y_i \leq y_i$ given θ equals

$$P(Y_i \leq y_i \mid \theta) = \frac{\Phi(y_i - \theta) - \Phi(y_0 - \theta)}{\Phi(y_1 - \theta) - \Phi(y_0 - \theta)}.$$

(e) A nontruncated random variable Z_i $(i = 1, \ldots, n)$ can be defined via an inverse normal transformation,

$$Z_i = \theta + \Phi^{-1}\big(\tilde{U}_i\big),$$

where $\tilde{U}_i = P(Y_i \leq y_i \mid \theta)$. Show how the augmented Z_i can be used to draw a value from the posterior density of θ.

3.8. Bayesian inferences and decision making under uncertainty can be done via posterior probabilities of the hypotheses. There is uncertainty about θ where under H_0 it is assumed that $\theta \leq \theta_0$ and under H_1 it is assumed that $\theta > \theta_0$. A loss function $L(a, \theta)$ can be defined for each action a,

$$L(a, \theta) = \begin{cases} \theta_0 - \theta & \text{accept } H_1,\ H_0 \text{ true} \\ \theta - \theta_0 & \text{accept } H_0,\ H_1 \text{ true} \\ 0 & \text{otherwise,} \end{cases}$$

where the loss of an incorrect decision depends on the severity of the mistake.
(a) Show that the posterior expected loss with respect to the posterior density $p(\theta \mid \mathbf{y})$ equals

$$E(L(a, \theta) \mid \mathbf{y}) = E(\theta_0 - \theta \mid \theta \leq \theta_0) + E(\theta - \theta_0 \mid \theta > \theta_0).$$

(b) Let Y denote the number of successes from n Bernoulli trials such that $Y \sim \mathcal{B}(n, \theta)$, and assume a beta prior for θ, denoted as $\mathcal{B}e(\alpha, \beta)$. Show that the marginal posterior density of θ is $\mathcal{B}e(\alpha', \beta')$ with $\alpha' = y + \alpha$ and $\beta' = n - y + \beta$.
(c) Derive that the posterior expected loss equals

$$E(L(a, \theta) \mid \mathbf{y}) = \theta_0\big(2P(\theta_0 \mid \mathbf{y}) - 1\big) - \int_0^{\theta_0} \theta^{\alpha'}(1 - \theta)^{\beta' - 1}\,\mathrm{d}\theta / B(\alpha', \beta') + \int_{\theta_0}^{1} \theta^{\alpha'}(1 - \theta)^{\beta' - 1}\,\mathrm{d}\theta / B(\alpha', \beta'),$$

where $B(\alpha', \beta')$ is the normalizing constant and $P(\theta_0 \mid \mathbf{y})$ the cumulative posterior distribution function; that is,

$$B(\alpha', \beta') = \int_0^1 \theta^{\alpha'-1}(1-\theta)^{\beta'-1}\,\mathrm{d}\theta.$$
$$P(\theta_0 \mid \mathbf{y}) = \int_0^{\theta_0} \theta^{\alpha'-1}(1-\theta)^{\beta'-1}\,\mathrm{d}\theta / B(\alpha', \beta').$$

3.9. (continuation of Exercise 3.1) Interest is focused on defining an HPD interval for the variance parameter.
(a) Show that the joint posterior density of (θ, σ^2) given the noninformative prior can be expressed as

$$p\left(\theta, \sigma^2 \mid \mathbf{y}\right) \propto \left(\sigma^2\right)^{-(n/2+1)} \exp\left(-\frac{1}{2\sigma^2}\left(S + n(\hat{\theta}-\theta)^2\right)\right).$$

(b) Verify that the marginal posterior density function of σ^2 can be expressed as

$$\begin{aligned} p\left(\sigma^2 \mid \mathbf{y}\right) &= \frac{p\left(\theta, \sigma^2 \mid \mathbf{y}\right)}{p\left(\theta \mid \sigma^2, \mathbf{y}\right)} \\ &\propto \left(\sigma^2\right)^{-(n+1)/2} \exp\left(\frac{-S}{2\sigma^2}\right). \end{aligned}$$

Show that σ^2 is inverse chi-square distributed with $\nu = n-1$ degrees of freedom and scale parameter νs^2, where $s^2 = S/\nu$.
(c) Derive the limits of a $(1-\alpha)$ HPD interval for σ^2.
(d) An HPD interval is not invariant under a nonlinear transformation of the parameters. Show that the endpoints of the HPD interval of Exercise 3.9(c) are not proportional to the endpoints of an HPD interval of σ.
(e) Assume that an MCMC sample $\sigma^{2(m)}$ $(m = 1, \ldots, M)$ is obtained from the (unimodal) marginal posterior density of $p(\sigma^2 \mid \mathbf{y})$. Construct an estimator of the $(1-\alpha)$ credible interval based on the ordered MCMC sample.
(f) Show how to compute a $(1-\alpha)$ HPD interval using the order statistics estimator. Note that an HPD interval is that credible interval with the smallest width. (Chen and Shao, 1999, proved convergence results of the order statistics estimator for unimodal posteriors and gave a conjecture for multimodal posteriors.)

4

Estimation of Bayesian Item Response Models

The general form of a Bayesian item response model consists of a probability model for the responses, prior distributions for the model parameters, and possibly prior distributions for the hyperparameters. An overview of Bayesian procedures for simultaneous estimation is given in which MCMC estimation methods are emphasized. Interest is focused on simultaneous estimation of marginal posterior densities of item and person parameters.

4.1 Marginal Estimation and Integrals

The estimation of parameters is often characterized as the problem of estimating structural parameters in the presence of incidental or nuisance parameters (e.g., Anderson, 1980; Ghosh, 1995). Some of the model parameters are of interest and others are not especially interesting and can be treated as nuisance parameters. The distinction between structural parameters and nuisance parameters is made with respect to the parameters that are to be estimated. The object is to obtain the marginal posterior density of the structural model parameters by integrating the joint posterior density of all parameters over the density of the nuisance parameters. Once the marginal posterior density of the parameters of interest is calculated, posterior probability statements, intervals, the posterior mean, and the posterior mode can be derived relative to the marginal posterior density. The marginal posterior density of the structural parameters summarizes all posterior information and is regarded as a basis for Bayesian inference regarding the structural parameters.

A different approach proposed by Swaminathan and Gifford (1982, 1985, 1986), avoids the integration problem. The maximizers of the joint posterior distribution of nuisance and structural parameters are considered to be the Bayesian estimates. A Newton-Raphson procedure is used to obtain the estimates that maximize the joint posterior. The joint posterior estimates are considered to be inferior to the marginal posterior estimates. First, the joint posterior estimates have to be based on informative priors that impose limits

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_4,

on the range of values. Otherwise, the estimates may drift out of bounds. Second, O'Hagan (1976) provided numerical evidence that marginal estimates are more accurate than the nonmarginal estimates, especially when the nuisance parameters are poorly determined in the joint posterior. Third, Mislevy (1986) argued that marginal estimates are also preferred on the basis of asymptotic behavior. In this case, the marginal posterior mode estimates correspond with the Bayes modal estimates that are asymptotically normally distributed under some regularity conditions. However, when the number of nuisance parameters increases to infinity and the number of structural parameters is held constant, the regularity conditions are not always satisfied (Neyman and Scott, 1948). As a result, asymptotic normality of the joint posterior mode estimator does not hold, but it may hold for the marginal posterior mode estimator depending on the prior distributions specified. Kim et al. (1994) performed a simulation study to compare the marginal posterior estimates with the joint posterior estimates using simulated data generated under different conditions. They found smaller root mean square differences and bias for the marginal posterior item parameter estimates, especially in the case of small samples and short tests.

In the literature, several procedures have been proposed to handle the numerical integration problem for obtaining marginal likelihood or marginal posterior estimates. Bock and Lieberman (1970) estimated the item parameters on the assumption that the person parameters are normally distributed and used Gauss-Hermite quadrature to perform the integration, where the person parameters are treated as nuisance parameters. The main computational problem was the size of the information matrix that needs to be inverted in the Newton-Raphson iterations to obtain the marginal estimates. Bock and Aitkin (1981), Mislevy (1984), and Thissen (1982) explored an approach based on the EM algorithm (Dempster, Laird and Rubin, 1977). In this approach, they assumed independent items, independent respondents, and independence between items and persons. Then, an EM solution can be obtained item by item. That is, there is one log-likelihood per item, involving only the parameters for that item. Interest is focused on estimating the item parameters, and the person parameters are treated as nuisance parameters. The distribution of person parameters is approximated by a discrete distribution covering a finite set of levels that can be recognized as the set of quadrature points. The term typically maximized that leads to marginal maximum likelihood item parameter estimates equals

$$\max l(\mathbf{y};\boldsymbol{\xi}) = \int l(\mathbf{y};\boldsymbol{\xi},\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \tag{4.1}$$

where the left-hand side is the marginal likelihood function. The derivative of the marginal log-likelihood that needs to be maximized is approximated using sufficient statistics. The sufficient statistics, defined at each individual latent variable level defining so-called pseudo-data, are (1) the number of respondents expected to have a certain latent variable level and (2) the number of respondents with this level expected to correspond correctly. They

are updated in the E-step of the EM algorithm given estimates for the item parameters. In the M-step, item parameter estimates are obtained given the updated pseudo-data. Both steps are repeated until a convergence criterion is satisfied. Note here that the assumption of normally distributed person parameters refers to a population model or to subjective uncertainty about the person parameters and was not supposed to define a prior distribution.

Mislevy (1986) implemented this EM procedure of Bock and Aitkin (1981) for the binary one-, two-, and three-parameter models including a hierarchical prior model for the item and person parameters (see also Lord, 1986). Posterior mode estimates (MAP, maximum a posteriori) were obtained instead of marginal maximum likelihood estimates via the Bock and Aitkin procedure.

The posterior modes (MAP) are usually used as estimates for the parameters of interest since they are easier to compute, although they are not invariant with respect to marginalization. The mean of the posterior (EAP, expected a posteriori) is invariant with respect to marginalization but is more difficult to compute when no closed-form solution is available. However, joint Bayes modal estimates are obtained for item and population (hyper)parameters since only the person parameters are integrated out in the joint posterior and the marginalized posterior still depends on the hyperparameters. Subsequently, the person parameters can be estimated by plugging in the marginal posterior mode estimates of the item parameters in the relevant conditional posterior distribution (Bock and Aitkin, 1981; Mislevy, 1986). In this approach, the uncertainty in the item and population parameter estimates is ignored.

For large samples, the estimated item parameters are approximately normally distributed with the inverse of the Fisher information matrix as the covariance matrix. Since the item parameters are usually unknown, the Fisher information matrix is difficult to compute and the posterior variances are approximated by computing the observed information matrix, defined as the negative second derivative of the log-likelihood function evaluated at the estimated parameter values (e.g., Mislevy, 1986; Tsutakawa, 1984). A more general overview of IRT estimation procedures, including specific EM implementations, can be found for example in Baker and Kim (2004) and van der Linden and Hambleton (1997).

According to Chapter 2, the marginal posterior of interest is proportional to

$$p(\boldsymbol{\xi} \mid \mathbf{y}) \propto \int\int\int p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})\, p(\boldsymbol{\theta} \mid \boldsymbol{\theta}_P)\, p(\boldsymbol{\xi} \mid \boldsymbol{\xi}_P) \\ p(\boldsymbol{\xi}_P)\, p(\boldsymbol{\theta}_P)\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\theta}_P\, \mathrm{d}\boldsymbol{\xi}_P. \tag{4.2}$$

In a hierarchical Bayes approach, inferences about the item parameters are to be made from their marginal posterior but the computation of the marginal posterior requires high-dimensional numerical integration. Typically, the integrand in Equation (4.2) resembles the (unnormalized) joint posterior that contains all information about the unknown parameters. Subsequently, the

marginal posterior of a single parameter is obtained by integrating over the distribution of the other parameters. Posterior summaries of the marginal density such as the posterior mean and the posterior standard deviation also require evaluating integrals.

In an empirical Bayes approach, the parameters in the highest level of the hierarchy are estimated using the data (Morris, 1983). That is, the computation of marginal population parameter estimates requires integration over person and item parameter distribution. It seems appealing to marginalize over person and item parameters to obtain the marginal posterior mode estimates of the hyperparameters, say $(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\xi}}_P)$. An empirical Bayes estimate of the item parameters can be obtained by computing marginal posterior estimates of the population parameters and the modal item parameter estimates in the marginalized conditional density given the population parameter estimates, $p(\boldsymbol{\xi} \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\xi}}_P)$.

The empirical Bayes estimates may not differ much from the hierarchical Bayes estimates when the number of respondents is large, leading to nearly symmetric posterior densities and stable parameter estimates. Empirical Bayes procedures are usually developed to approximate the hierarchical Bayes estimates that are obtained via high-dimensional integration. However, Berger (1985) noted that an empirical Bayes procedure can differ substantially from a hierarchical Bayes procedure since the hyperparameter estimation error is ignored. The empirical Bayes procedure fails to indicate how to incorporate the hyperparameter estimation error. The errors are automatically incorporated in the hierarchical Bayes analysis.

To perform a hierarchical Bayes approach, (potentially) high-dimensional integrals need to be evaluated. Therefore, computational methods are needed that support the computation of complicated marginal posterior densities and posterior summaries. In Section 4.2, it will be shown that MCMC methods enable the computation of high-dimensional integrals and facilitate the superior hierarchical Bayes modeling approach.

In conclusion, several remarks can be made concerning advantages of hierarchical Bayes estimates over marginal maximum likelihood estimates.

- In the situation where information is available from previous administrations or reasonable prior distributions can be constructed given the current data, Bayesian estimates will be superior to marginal maximum likelihood estimates. The Bayesian approach enables the use of additional information for estimating the parameters.
- The Bayesian procedure uses prior and sample information to estimate a parameter, and the estimate will have a smaller standard error than the standard error corresponding to a marginal maximum likelihood estimate. This is only an advantage when reasonable prior information is available.
- A Bayes estimate is based on a combination of prior and sample information that results in an estimate that relies on the response pattern but also on individual characteristics and/or group-specific information. For exam-

ple, a respondent's ability estimate might shrink towards a group mean in an upward (downward) direction when the respondent is a member of a high (low) ability group. Mislevy (1986) extended an item response model with prior population models such that observations provide information about the person parameters but also contribute information about the population to which they belong. Knowledge of the populations is used to improve the estimation of the person parameters.

- The use of prior information has the advantage that the parameter estimates (item and person) are restricted to a plausible range of values and avoid estimates that drift out of range. Person and item parameter estimates can be obtained for perfect and imperfect response patterns.
- A Bayesian estimation procedure is more appropriate for moderate and smaller sample sizes since it does not rely on large-sample asymptotic results like the marginal maximum likelihood procedure. For very large sample sizes, Bayesian and marginal maximum likelihood estimates tend to be similar, the posterior becomes dominated by the likelihood, and the influence of the prior becomes negligible.
- Most statistical inferences follow easily from the estimated posterior distribution of the structural parameters. For instance, one can obtain parameter estimates and related credible sets. This is in contrast with the marginal maximum likelihood procedure, where obtaining estimates and related confidence intervals are two different (computational) problems.

4.2 MCMC Estimation

In the 1990s, several MCMC implementations were developed for logistic and normal ogive item response models. The developed simulation-based algorithms can be grouped using two MCMC sampling methods: Metropolis-Hastings (M-H) and Gibbs sampling. The Gibbs sampling method was used by Albert (1992) for the two-parameter normal ogive model (see also Albert and Chib, 1993). The Gibbs sampling implementation of Albert (1992) is characterized by the fact that all full conditionals can be obtained in closed form and it is possible to directly sample from them given augmented data. This approach has been extended in several directions, and several expansions will be considered in subsequent chapters.

Combining Gibbs sampling with M-H sampling leads to an M-H within Gibbs algorithm that turns out to be a powerful technique for obtaining samples from the target distribution (see Chib and Greenberg, 1995, for a general description). Patz and Junker (1999a, 1999b) proposed an M-H within Gibbs sampler for several logistic response models. Their M-H within Gibbs scheme is extended by allowing within-item dependencies and treating the hyperparameters as unknown model parameters.

In each M-H step, the so-called candidate-generating density equals a normal proposal density with its mean at the current state. Then, the acceptance

ratio is defined by a posterior probability ratio since the proposal density is symmetric. For the two-parameter logistic response model (Equation (1.3)) a hierarchical prior for the item parameters is defined as

$$\tilde{\boldsymbol{\xi}}_k = (\log a_k, b_k)^t \sim \mathcal{N}(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi), \tag{4.3}$$

with the second-stage prior for the hyperparameters according to Equations (2.4) and (2.5). An exchangeable hierarchical prior is defined for θ_i according to Equations (2.8)–(2.10). This leads to the following M-H within Gibbs scheme.

MCMC Scheme 1 (Two-Parameter Logistic Model)

1. *Sample* $\theta_i^* \sim \mathcal{N}(\theta_i^{(m)}, \varphi_\theta)$ *and* $U_i \sim \mathcal{U}_{[0,1]}$, *and set* $\theta_i^{(m+1)} = \theta_i^*$ *when*

$$u_i \leq \frac{p\left(\mathbf{y}_k \mid \theta_i^*, \boldsymbol{\xi}^{(m)}\right) p\left(\theta_i^* \mid \mu_\theta^{(m)}, \sigma_\theta^{2(m)}\right)}{p\left(\mathbf{y}_k \mid \theta_i^{(m)}, \boldsymbol{\xi}^{(m)}\right) p\left(\theta_i^{(m)} \mid \mu_\theta^{(m)}, \sigma_\theta^{2(m)}\right)}$$

   and otherwise set $\theta_i^{(m+1)} = \theta_i^{(m)}$ *for* $i = 1, \ldots, n$.

2. *Sample* $\tilde{\boldsymbol{\xi}}_k^* \sim \mathcal{N}\left(\tilde{\boldsymbol{\xi}}_k^{(m)}, \varphi_\xi\right)$ *and* $U_k \sim \mathcal{U}_{[0,1]}$, *and set* $\boldsymbol{\xi}_k^{(m+1)} = \boldsymbol{\xi}_k^*$ *when*

$$u_k \leq \frac{p\left(\mathbf{y}_i \mid \boldsymbol{\xi}_k^*, \boldsymbol{\theta}^{(m+1)}\right) p\left(\tilde{\boldsymbol{\xi}}_k^* \mid \boldsymbol{\mu}_\xi^{(m)}, \boldsymbol{\Sigma}_\xi^{(m)}\right)}{p\left(\mathbf{y}_i \mid \boldsymbol{\xi}_k^{(m)}, \boldsymbol{\theta}^{(m+1)}\right) p\left(\tilde{\boldsymbol{\xi}}_k^{(m)} \mid \boldsymbol{\mu}_\xi^{(m)}, \boldsymbol{\Sigma}_\xi^{(m)}\right)}$$

   and otherwise set $\boldsymbol{\xi}_k^{(m+1)} = \boldsymbol{\xi}_k^{(m)}$ *for* $k = 1, \ldots, K$.

3. *Sample hyperparameter values* $(\mu_\theta, \sigma_\theta^2, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ *from their full conditionals*

   a) *Sample* $\mu_\theta^{(m+1)}, \sigma_\theta^{2(m+1)}$ *from the conditional densities*

$$\mu_\theta \mid \sigma_\theta^{2(m)}, \boldsymbol{\theta}^{(m+1)} \sim \mathcal{N}\left(\frac{n_0}{n+n_0}\mu_0 + \frac{n}{n+n_0}\overline{\theta}, \frac{\sigma_\theta^2}{n+n_0}\right),$$
$$\sigma_\theta^2 \mid \boldsymbol{\theta}^{(m+1)} \sim \mathcal{IG}\left(g_1 + \frac{n}{2}, \sigma_n^2\right),$$

   where $\sigma_n^2 = g_2 + (n-1)s^2/2 + \frac{nn_0}{2(n+n_0)}(\overline{\theta} - \mu_0)^2$ *(see Exercise 4.2 for details).*

   b) *Sample* $\boldsymbol{\mu}_\xi^{(m+1)}, \boldsymbol{\Sigma}_\xi^{(m+1)}$ *from the conditional densities*

$$\boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi^{(m)}, \tilde{\boldsymbol{\xi}}^{(m+1)} \sim \mathcal{N}\left(\frac{K_0}{K_0+K}\boldsymbol{\mu}_0 + \frac{K}{K_0+K}\overline{\boldsymbol{\xi}}, \frac{\boldsymbol{\Sigma}_\xi}{K+K_0}\right),$$
$$\boldsymbol{\Sigma}_\xi \mid \tilde{\boldsymbol{\xi}}^{(m+1)} \sim \mathcal{IW}(K+\nu, \boldsymbol{\Sigma}^*),$$

   where $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0 + K\mathbf{S} + \frac{KK_0}{K+K_0}\left(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_0\right)\left(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_0\right)^t$ *(see Exercise 4.3 for details).*

The advantage of an M-H algorithm is the fact that an arbitrary conditional proposal density can lead to simulating values from the posterior target density. But at the same time, the drawback of an M-H algorithm is that the choice of the proposal density greatly affects the rate of convergence. If the proposal leads to a high acceptance rate, the Markov chain takes small steps, resulting in slow convergence. If it leads to a low acceptance rate, the Markov chain converges slowly because most of the proposed candidates are rejected. Establishing good convergence in a more complex M-H scheme, including several proposals, can be quite difficult. In some cases, an acceptable convergence rate can be obtained by adjusting the variance of the proposal density (Gelman, Meng and Stern, 1996).

MCMC scheme 1 for estimating the parameters of the two-parameter logistic model is easily extended to the three-parameter logistic model. Since the posterior density of the guessing parameter is less peaked, it is difficult to sample values from its conditional density, via a proposal density, around the posterior mean (Patz and Junker, 1999b). That is, a lot of the sampled values are located in the tails of the posterior, resulting in an unstable estimate of the guessing parameter. Moreover, the convergence of the M-H algorithm becomes more complicated since the variance of the proposal density of the guessing parameter also has to be identified.

4.3 Exploiting Data Augmentation Techniques

A Gibbs sampling implementation has two advantages over an M-H implementation. First, a Gibbs sampler avoids the specification of a proposal distribution and possibly related convergence problems; that is, an M-H method often needs tuning to obtain a reliable and efficient algorithm. Second, a Gibbs sampler, in contrast to an M-H algorithm, is more easily implemented since it consists of sampling from well-known distributions. The use of a Gibbs sampler implies limitations on the choice of the posterior distribution since it requires direct sampling from it. This can be averted by data augmentation with a completion construction also known as the auxiliary variable method.

In the method of auxiliary variables, realizations from a complicated distribution can be obtained by augmenting the variables of interest by one or more additional variables such that the full conditionals are tractable and easy to simulate from. The construction of sampling algorithms via the introduction of (unobserved) augmented data received much attention since it resulted in both simple and fast algorithms (e.g., Higdon, 1998; Meng and van Dyk, 1999; Neal, 1997). Tanner and Wong (1987) introduced augmented variables to improve the speed of convergence and showed that their data augmentation scheme made a sampling procedure feasible and simple. It will be shown that these developments also proved to be useful for constructing efficient MCMC schemes for item response models.

The data augmentation algorithm is based on the introduction of so-called augmented data, denoted as Z, that are linked to the observed data via a many-to-one mapping such that $f(Z) = Y$, so infinitely many values of Z will give the same value of Y. Equivalent inferences about a model parameter θ are to be made when using a model for the observed data, $p(y \mid \theta)$, or when using a model for the augmented data, $p(z \mid \theta)$, since

$$p(y \mid \theta) = \int_{f(z)=y} p(z \mid \theta) \, \mathrm{d}z. \tag{4.4}$$

A proper augmentation scheme obeys the restriction that the distribution of the observed data is implied by the distribution of the augmented data. The augmented data are introduced for computational purposes and should not alter the analysis model. Appropriate augmented data are useful mainly when sampling from both $p(z \mid y, \theta)$ and $p(\theta \mid z)$ becomes easier or feasible and sampling directly from $p(\theta \mid y)$ is difficult. It is possible that via data augmentation the Markov chains mix more quickly and thus reduce the computation time (van Dyk and Meng, 2001). Further on, the many-to-one mapping that specifies the range of integration will often be omitted from the expression, as it will be specified implicitly by the data augmentation scheme.

An augmented variable will also be referred to as a latent or auxiliary variable. Latent responses are also referred to as augmented data when the observed data are seen as indicators of underlying continuous responses.

4.3.1 Latent Variables and Latent Responses

The auxiliary or latent variable approach has several important advantages. First, the approach is very flexible and can handle almost all sorts of discrete responses. Typically, the likelihood of the observed response data has a complex structure but the likelihood of the augmented (latent) data has a known distribution with convenient mathematical properties. Second, conjugate priors, where the posterior has the same algebraic form as the prior, can be more easily defined for the likelihood of the latent response data, which has a known distributional form, than for the likelihood of the observed data. Third, the augmented variable approach facilitates easy formulation of a Gibbs sampling algorithm based on data augmentation. It will turn out that by augmenting with a latent continuous variable, conditional distributions can be defined based on augmented data, from which samples are easily drawn. Fourth, the conditional posterior given augmented data has a known distributional form such that conditional probability statements can be directly evaluated for making posterior inferences. The likelihood of the augmented response data is much more easily evaluated than the likelihood of the observed data and can be used to compare models. Fifth, item response models are not identified such that a different set of parameter values leads to the same distribution of the response data. This identification problem is often better understood in

a latent variable formulation. Then, the introduced latent response variable becomes a dependent variable and its scale needs to be fixed to identify the model. This viewpoint provides a direct interpretation of the identification problem. A proper (conjugate) prior can be defined that identifies the posterior although the likelihood is not identified. It will be shown that the scale of the latent responses can also be fixed in each MCMC iteration without imposing exact restrictions on the model parameters.

The likelihood $p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})$ specifies the distribution of the observed responses given levels of item and person parameters. Below, it will be shown that the likelihood of the augmented response data will be consistent with the conditional likelihood for the observed response data. Subsequently, various latent variable formulations will be given to handle probit and logistic response models for binary and polytomous response data.

4.3.2 Binary Data Augmentation

Torgerson (1958) showed that the normal ogive item characteristic curve that relates the level of ability to the probability of a correct response (see Figure 1.2) can also be presented by the level of ability and a normal deviate Z_{ik} that corresponds to the probability of a correct response. In Figure 4.1, levels of ability, θ_i, are plotted against deviate values $Z_{ik} = a_k\theta_i - b_k$ (right vertical axis). When the deviate values are considered to be normal deviate values, the levels of ability plotted against the corresponding success probabilities (left vertical axis) represent the normal ogive ICC.

The two-parameter logistic model can be written as $P(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k) = \Psi(a_k\theta_i - b_k)$, where $\Psi(.)$ is the standard logistic cumulative distribution function with mean zero and variance $\pi^2/3$. This model can be written in terms of the log odd or the logit of the probabilities:

$$\log\left(\frac{P(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k)}{1 - P(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k)}\right) = a_k\theta_i - b_k.$$

It follows that the logistic model defines a linear relationship between the logits and the parameters; that is, the model is linear in terms of the logits. When the deviate values in Figure 4.1 are considered to be logistic deviate or logit values, the levels of ability plotted against the corresponding success probabilities represent the logistic ICC. In Figure 4.1, a logistic scale factor of $d = 1.7$ is used such that a close agreement is obtained between the logistic and normal ogive ICCs (see Exercise 4.6).

Several parameterizations are discussed in Torgerson (1958) to describe the straight-line ICC of Figure 4.1. One of the parameterizations stems from Tucker (1952), who described the straight line in Figure 4.1 as $Z_{ik} = a_k\theta_i - b_k$, where a_k and b_k correspond to the slope and the intercept, respectively, and they are sometimes referred to as Tucker's item parameters (e.g., Torgerson, 1958). This parameterization is also pursued here.

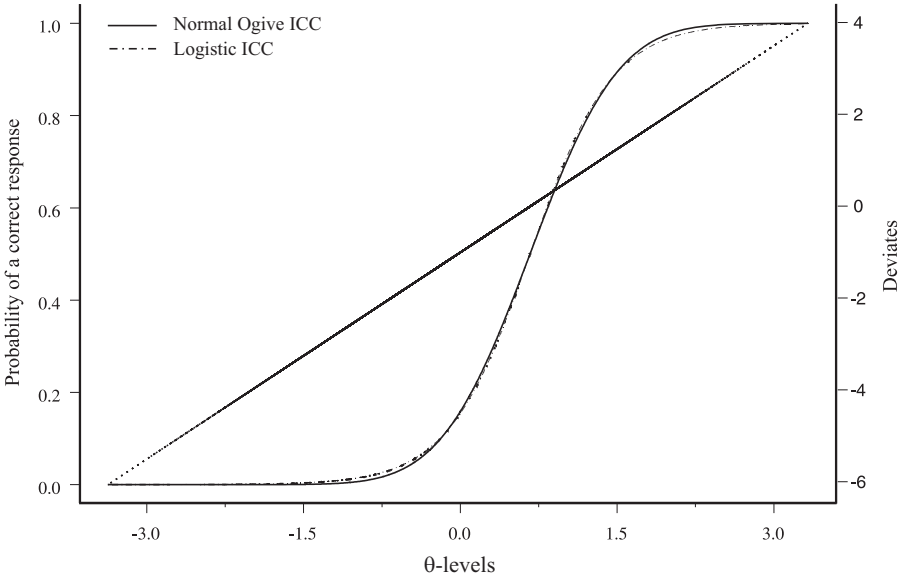

Fig. 4.1. Normal ogive and logistic ICCs with respect to the deviate values and the corresponding success probabilities.

Let the latent variable Z_{ik} determine the performance of respondent i on item k. The respondent answers the item correctly if $Z_{ik} > 0$ ($Y_{ik} = 1$) and incorrectly if $Z_{ik} \leq 0$ ($Y_{ik} = 0$). Then, the response probability that person i answers item k correctly is defined by

$$\begin{aligned} P\left(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k\right) &= P(Z_{ik} > 0 \mid \theta_i, \boldsymbol{\xi}_k) \\ &= \int_0^\infty \phi(z; a_k\theta_i - b_k)\,\mathrm{d}z = \Phi(a_k\theta_i - b_k) \qquad (4.5) \\ &= \int_0^\infty \psi(z; d(a_k\theta_i - b_k))\,\mathrm{d}z = \Psi(d(a_k\theta_i - b_k)), \quad (4.6) \end{aligned}$$

where $\phi(\cdot)$ and $\psi(\cdot)$, as defined in Chapter 1, denote the normal and logistic density functions, respectively, and d the logistic scale factor.

Torgerson (1958, pp. 386–388) and Lord and Novick (1968, pp. 358–394) defined three conditions leading to a two-parameter item response model. The first two conditions state that the regression of this latent variable Z_{ik} on the latent ability parameter θ_i is linear and with the same error variance for all θ_i. The third condition states that a normal ogive model is defined when the conditional distribution of Z_{ik} given θ_i is normal (Equation (4.5)) and a logistic model is defined when the conditional distribution is logistic (Equation (4.6)). It can be concluded that for the normal ogive model the deviates are normally distributed, corresponding to a linear relationship between the ability levels and the normal deviates, and for the logistic model the deviates

are logistically distributed, corresponding to a linear relationship between the ability levels and the logistic deviates or logits.

Albert (1992) (see also Albert and Chib, 1993) constructed an MCMC algorithm using the latent variable Z_{ik} (normal deviate) as an auxiliary variable for estimating the two-parameter normal ogive model. A data augmentation scheme for different logistic item response models was developed by Maris and Maris (2002) that used the logistically distributed latent variable Z_{ik}. The augmented data are defined in such a way that each full conditional becomes an indicator function with bounds specified by the other parameter values. As a result, the sampling of the parameters is easy; however, the sampled values are highly correlated due to this incorporated dependency structure. The samples cannot be drawn freely from the target distribution but are restricted to a subspace specified by the other parameter values.

Here, an augmented data scheme will be defined that can handle unidimensional logistic as well as normal ogive response models and will lead to a simple and fast M-H within Gibbs algorithm. Let $\mathcal{L}(0,1)$ and $\mathcal{N}(0,1)$ denote the standard logistic and the standard normal density functions, respectively.[1] Two different approaches can be followed for augmenting data. The first approach is based on defining a deviate as $Z_{ik} = a_k\theta_i - b_k$. Now, augmented data are conditionally distributed as

$$Z_{ik} \mid Y_{ik}, \theta_i, \boldsymbol{\xi}_k \sim \begin{cases} \mathcal{L}(d(a_k\theta_i - b_k), 1) \\ \mathcal{N}(a_k\theta_i - b_k, 1), \end{cases} \tag{4.7}$$

where Y_{ik} is the indicator that Z_{ik} is positive and $d = 1.7$ (see Exercise 4.6 for more details about the logistic scale factor d). This approach for the normal deviate corresponds with the procedure of Albert (1992). In the second approach, augmented data are standard normally or standard logistically distributed but truncated in such a way that the distribution of the augmented data implies the distribution of the observed data. In this case, the augmented data are conditionally distributed as

$$\tilde{Z}_{ik} \mid Y_{ik}, \theta_i, \boldsymbol{\xi}_k \sim \begin{cases} \mathcal{L}(0,1) \\ \mathcal{N}(0,1), \end{cases} \tag{4.8}$$

where Y_{ik} is the indicator that assumes a value one if $\tilde{Z}_{ik} > d(b_k - a_k\theta_i)$ and zero otherwise ($d = 1.7$ and $d = 1$ for the two-parameter logistic and normal ogive response models, respectively). The auxiliary variables Z_{ik} and $\tilde{Z}_{ik}$ as defined in Equations (4.7) and (4.8) are related to each other; that is, $Z_{ik} = \tilde{Z}_{ik} + d(a_k\theta_i - b_k)$ such that Z_{ik} is normally (logistically) distributed with mean $d(a_k\theta_i - b_k)$ and variance one ($\pi^2/3$) when $\tilde{Z}_{ik}$ is standard normally (logistically) distributed.

Given the defined augmentation schemes, say Z_{ik} is normally distributed (Equation (4.7)) and $\tilde{Z}_{ik}$ logistically distributed (Equation (4.8)), it follows

[1] It follows that $Z \sim \mathcal{L}(0,1)$ corresponds with $p(z) = \psi(z;0,1)$ and $Z \sim \mathcal{N}(0,1)$ corresponds with $p(z) = \phi(z;0,1)$.

that the distribution of the observed data is implied by the distribution of the augmented data,

$$\begin{aligned} P\left(Y_{ik}=1 \mid \theta_i, \boldsymbol{\xi}\right) &= P\left(\tilde{Z}_{ik} > d(b_k - a_k\theta_i) \mid \theta_i, \boldsymbol{\xi}_k\right) \\ &= \int_{d(b_k-a_k\theta_i)}^{\infty} \psi\left(x\right)\, \mathrm{d}x = \Psi(d(a_k\theta_i - b_k)) \qquad (4.9) \\ &= P\left(Z_{ik} > 0 \mid \theta_i, \boldsymbol{\xi}_k\right) \\ &= \int_{0}^{\infty} \phi\left(x; a_k\theta_i - b_k\right)\, \mathrm{d}x = \Phi\left(a_k\theta_i - b_k\right). \qquad (4.10) \end{aligned}$$

It can be seen from Equation (4.9) that the definition of the conditional logistically distributed auxiliary variable $\tilde{Z}_{ik}$ leads to the probability of a correct response under the two-parameter logistic model. From Equation (4.10) it follows that the definition of the conditional normally distributed auxiliary variable Z_{ik} leads to the probability of a correct response under the two-parameter normal ogive model. As a result, the choice of the measurement model is defined in this data augmentation step. The augmentation step defines a probit or logit analysis.

The parameters of the complete-data likelihood can be separated from the distribution of the latent response data. As an example, the conditional posterior density of the augmented data, $\tilde{\mathbf{Z}}$, can be written as

$$\begin{aligned} p\left(\tilde{\mathbf{z}} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) \propto \prod_{i,k} \Big[& I_{(0,\infty)}(\tilde{z}_{ik} + d(a_k\theta_i - b_k))^{y_{ik}} \cdot \\ & I_{(-\infty,0)}(\tilde{z}_{ik} + d(a_k\theta_i - b_k))^{1-y_{ik}} \Big] p(\tilde{z}_{ik}), \qquad (4.11) \end{aligned}$$

where $p(\tilde{z}_{ik})$ is the standard logistic or normal density function. Now, the conditional posterior density of the ability parameter can be constructed from the complete-data likelihood and its prior (see also Exercise 4.4(e)). The restriction on the $\tilde{z}_{ik}$ can be expressed as a restriction on the θ_i. Therefore, define the set $\mathcal{A}_{ik} = \{\theta_i \in \mathcal{R}; \theta_i > (db_k - \tilde{z}_{ik})/da_k\}$ with the complement set $\mathcal{A}_{ik}^c$. It follows that the full conditional posterior density of θ_i can be expressed as

$$\begin{aligned} p\left(\theta_i \mid \mathbf{y}, \tilde{\mathbf{z}}, \boldsymbol{\xi}, \boldsymbol{\theta}_P\right) \propto \prod_{k} \Big[& I_{\mathcal{A}_{ik}}(\theta_i) I(Y_{ik}=1) + I_{\mathcal{A}_{ik}^c}(\theta_i) \\ & I(Y_{ik}=0) \Big] p(\theta_i \mid \mu_\theta, \sigma_\theta^2) \\ \propto \prod_{k \mid y_{ik}=1} I_{\mathcal{A}_{ik}}(\theta_i) & \prod_{k \mid y_{ik}=0} I_{\mathcal{A}_{ik}^c}(\theta_i) p(\theta_i \mid \mu_\theta, \sigma_\theta^2) \\ \propto I_{\cup_{k \mid y_{ik}=1} \mathcal{A}_{ik}}(\theta_i) & I_{\cap_{k \mid y_{ik}=0} \mathcal{A}_{ik}^c}(\theta_i) p(\theta_i \mid \mu_\theta, \sigma_\theta^2). \qquad (4.12) \end{aligned}$$

The set indicator function of the union (intersection) of sets is the maximum (minimum) function of the indicator functions. Therefore, define

$$\Delta_l = \max_{k|y_{ik}=1} (db_k - \tilde{z}_{ik})/da_k, \tag{4.13}$$

$$\Delta_u = \min_{k|y_{ik}=0} (db_k - \tilde{z}_{ik})/da_k, \tag{4.14}$$

and it follows that Equation (4.12) can be written as

$$p\left(\theta_i \mid \mathbf{y}, \tilde{\mathbf{z}}, \boldsymbol{\xi}, \mu_\theta, \sigma_\theta^2\right) \propto I_{(\Delta_l, \Delta_u)}(\theta_i) p(\theta_i \mid \mu_\theta, \sigma_\theta^2). \tag{4.15}$$

The conditional posterior can be recognized as a truncated prior that is conditionally independent of the distribution of the augmented data. The following MCMC scheme can be defined for estimating the parameters of logistic and normal ogive models.

MCMC Scheme 2 (Two-Parameter Model)

1. *Sample augmented data* $\tilde{\mathbf{z}}^{(m+1)}$ *according to Equation* (4.8) *for a logistic or normal ogive model.*
2. *Let the prior for* θ_i *be defined by Equation* (2.8)*. According to* (4.15)*, sample* $\theta_i^{(m+1)}$ *from*

$$\theta_i \mid \mathbf{y}, \tilde{\mathbf{z}}^{(m+1)}, \boldsymbol{\xi}^{(m)}, \mu_\theta^{(m)}, \sigma_\theta^{2(m)} \sim \mathcal{N}\left(\mu_\theta, \sigma_\theta^2\right) I_{\mathcal{R}_\theta}(\theta_i), \tag{4.16}$$

   where $\mathcal{R}_\theta = \{\theta \in \mathcal{R}, \Delta_l < \theta < \Delta_u\}$ *with* (Δ_l, Δ_u) *defined in Equations* (4.13) *and* (4.14)*, respectively.*
3. *Let the prior for* $\boldsymbol{\xi}_k$ *be defined by Equation* (2.3) *with known values for* $\boldsymbol{\mu}_\xi$ *and* $\boldsymbol{\Sigma}_\xi$*. Let* $\mathbf{x} = \left(d\boldsymbol{\theta}^{(m+1)}, -d\mathbf{1}_n\right)$*, and sample* $\boldsymbol{\xi}_k^{(m+1)}$ *from the conditional density*

$$\boldsymbol{\xi}_k \mid \mathbf{z}_k^{(m+1)}, \boldsymbol{\theta}^{(m+1)}, \boldsymbol{\mu}_\xi^{(m)}, \boldsymbol{\Sigma}_\xi^{(m)} \sim \mathcal{N}\left(\boldsymbol{\mu}_\xi^*, \boldsymbol{\Omega}_\xi\right) I_{\mathcal{A}_k}(a_k), \tag{4.17}$$

   where

$$\boldsymbol{\Omega}_\xi^{-1} = \varphi^{-1}\left(\mathbf{x}^t\mathbf{x}\right) + \boldsymbol{\Sigma}_\xi^{-1}, \tag{4.18}$$

$$\boldsymbol{\mu}_\xi^* = \boldsymbol{\Omega}_\xi\left(\mathbf{x}^t\mathbf{z}_k + \boldsymbol{\mu}_\xi\boldsymbol{\Sigma}_\xi^{-1}\right), \tag{4.19}$$

   with $Z_{ik} = \tilde{Z}_{ik} + d\left(a_k\theta_i - b_k\right)$ *and* $\varphi = 1$ *or* $\varphi = \pi^2/3$ *if* $\tilde{\mathbf{Z}}_k$ *is normally or logistically distributed, respectively. The density in Equation* (4.17) *is used to generate candidates evaluated in an M-H step when* $\tilde{\mathbf{Z}}_k$ *is logistically distributed.*
4. *Sample values of prior parameters* $\mu_\theta^{(m+1)}$*,* $\sigma_\theta^{2(m+1)}$ *according to step 3 in MCMC scheme 1.*

In step 3 of scheme 2, the standard logistic density function is approximated by a normal density function with variance $\pi^2/3$ and an M-H step is used to correct any deficiencies in the approximation since the tail of the logistic density function is somewhat longer. Candidate values drawn from the

density in Equation (4.17) are almost always accepted since both densities (normal and logistic) are closely comparable (see, for example, Figure 4.1).

A sampling step for the hyperparameters $(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ is complicated due to the positivity restriction on the discrimination parameters (see Exercise 4.4). The lognormal prior for the item parameters in Equation (4.3) is a nonconjugate prior to the augmented data likelihood. Then, an M-H step is required to draw samples from their full conditional, but the hyperparameters can be sampled according to step 3 of MCMC scheme 1.

MCMC scheme 2 can be extended to sample the parameters of the three-parameter model. This is done by introducing another auxiliary variable, denoted as S_{ik}, which equals one when person i knows the correct answer to item k and equals zero otherwise (Béguin and Glas, 2001). Note that, according to Equation (1.6) or (1.7), a correct response can be obtained when person i knows and gives a correct response but also when person i does not know the correct response but gives it by guessing correctly. In this light, the conditional probability that respondent i knows the correct answer to item k given a correct response equals

$$P\left(S_{ik}=1 \mid Y_{ik}=1,\theta_i,\boldsymbol{\xi}_k\right)=\frac{\Phi\left(a_k\theta_i-b_k\right)}{\Phi\left(a_k\theta_i-b_k\right)+c_k\left(1-\Phi\left(a_k\theta_i-b_k\right)\right)}. \quad (4.20)$$

In the same way, the conditional probability that respondent i does not know the correct answer to item k given a correct response equals

$$P\left(S_{ik}=0 \mid Y_{ik}=1,\theta_i,\boldsymbol{\xi}_k\right)=\frac{c_k\left(1-\Phi\left(a_k\theta_i-b_k\right)\right)}{\Phi\left(a_k\theta_i-b_k\right)+c_k\left(1-\Phi\left(a_k\theta_i-b_k\right)\right)}. \quad (4.21)$$

The value of S_{ik} equals zero with probability one when an incorrect response is given by respondent i to item k. Note that the normal ogive response probabilities can be replaced by logistic response probabilities in Equations (4.20) and (4.21). The following conditional sampling scheme can be adopted. Set S_{ik} equal to zero when $Y_{ik}=0$. When $Y_{ik}=1$, sample S_{ik} from a Bernoulli distribution with a success probability defined by Equation (4.20). Per item k, the number of correct responses via guessing is binomially distributed with parameters $n_s=\sum_{i|S_{ik}=0} I\left(Y_{ik}=1\right)$ and success probability c_k. This likelihood combined with the conjugate beta prior in Equation (2.7) leads to a beta posterior density,

$$c_k \mid \mathbf{s},\mathbf{y} \sim \mathcal{B}e\left(\alpha+n_s,\beta+\tilde{s}_k-n_s\right), \quad (4.22)$$

where $\tilde{s}_k$ denotes the number of persons who do not know the correct answer to item k and guess the response. A comparable implementation is given by Sahu (2002), who showed a better performance of the Gibbs sampler over the M-H algorithm in terms of the effective sample size. The sampling of augmented data $\mathbf{S}$ and parameters $\mathbf{c}$ is easily integrated in MCMC scheme 2, but note that augmented data, item, and person parameters are to be sampled conditionally on the values of $\mathbf{S}$.

4.3.3 TIMMS 2007: Dutch Sixth-Graders' Math Achievement

The Netherlands was one of the participating countries in TIMSS 2007 (Trends in International Mathematics and Science Study), a large-scale international assessment of math and science achievement conducted on a 4-year cycle that started in 1995. The TIMMS 2007 data can be found at `http://timss.bc.edu/timss2007/idb_ug.html` (January 2010).

The math achievement of Dutch pupils in group 6 (first class of the upper level in primary education, equivalent to the fourth school year after kindergarten) was assessed. For illustration purposes, eight math items of the content domain number stored in booklets 1 and 14 were considered.

A two-parameter response model was used for analyzing the item response data. A standard normal prior was defined for the ability parameters for identification purposes (see Section 4.4.2). Independence between the discrimination and difficulty parameters and a common normal distribution for both sets of parameters was assumed such that

$$a_k \sim \mathcal{N}\left(\mu_a, \sigma_a^2\right), \tag{4.23}$$

$$b_k \sim \mathcal{N}\left(\mu_b, \sigma_b^2\right). \tag{4.24}$$

The hyperparameters μ_a and μ_b were normally distributed with means one and zero, respectively, and a large variance. The hyperparameters σ_a^2 and σ_b^2 both had an inverse gamma prior with a small value for the shape and scale parameters.

Normally distributed augmented data were sampled to fit a normal ogive response model. A popular MCMC scheme was used that corresponds closely to Albert's (1992) scheme:

1. Sample augmented data according to Equation (4.7).
2. Sample ability parameters as specified in Exercise 4.5(a).
3. Sample the item parameters (see Exercise 4.5(d)).
4. Sample the hyperparameters (see Exercise 4.5(e)).

Starting values were generated from the priors with $\mu_a = 1$ and $\mu_b = 0$. The MCMC algorithm was run for 10,000 iterations, and the first 2,000 iterations were considered as a burn-in. Imputed data were used for the missing observations. The imputation procedure is described in Exercise 4.10(e).

In Table 4.1, the item parameter estimates and the hyperparameter estimates are given. It can be seen that the items discriminate quite well and there is not much variation in discrimination values. The variation in difficulty values is much higher (around 1.1). Items one, three, and four appear to be much more difficult than items five, seven, and eight. This follows from the fact that items one, three, and four were answered correctly by only 15% to 20% of the pupils and items five, seven, and eight were answered correctly by 70% to 90% of the pupils.

Table 4.1. TIMMS 2007: Item parameter estimates of eight math items of the content domain number.

	Discrimination		Difficulty	
Item	Mean	SD	Mean	SD
1	.864	.132	1.072	.114
2	.616	.101	.499	.075
3	.920	.141	1.029	.117
4	.875	.143	1.377	.140
5	.544	.098	−.615	.074
6	.760	.114	−.003	.074
7	.924	.142	−.942	.108
8	.759	.128	−1.433	.130
μ_a	.786	.102		
μ_b	.119	.378		
σ_a^2	.063	.046		
σ_b^2	1.113	.732		

The MCMC output was used to compute the posterior probability of answering the difficult item 4 incorrectly and the easy item 8 correctly. This follows from

$$P\left(Y_{i4}=0, Y_{i8}=1 \mid \mathbf{y}\right) = \int P\left(Y_{i4}=0 \mid \theta_i, \mathbf{y}\right) P\left(Y_{i8}=1 \mid \theta_i, \mathbf{y}\right) p\left(\theta_i \mid \mathbf{y}\right) \mathrm{d}\theta_i$$
$$\approx M^{-1} \sum_m \left(1 - \Phi\left(a_4^{(m)} \theta_i^{(m)} - b_4^{(m)}\right)\right) \Phi\left(a_8^{(m)} \theta_i^{(m)} - b_8^{(m)}\right)$$

for M MCMC samples $\left(\mathbf{a}^{(m)}, \theta^{(m)}, \mathbf{b}^{(m)}\right)$ from the joint posterior density. High- and low-ability pupils have a low posterior probability of up to .40. As expected, average-ability pupils have a high posterior probability of up to .80. Answering item 4 correctly and item 8 incorrectly is very unlikely for all pupils. This is reflected by the computed posterior probabilities, which vary from .001 to .008 for pupils with different levels of ability.

It was investigated whether the boys performed better on the eight items than the girls. Therefore, a fixed factor was defined with values equal to one for the boys and equal to zero for the girls. This fixed factor was used to define a mean difference in ability between boys and girls. That is, the prior for the ability parameters had a fixed mean of zero for the girls and an unknown mean, μ_θ, for the boys.[2] The estimated posterior population mean of the boys' abilities equalled .282, with a posterior standard deviation of .115, and it was concluded that the boys significantly outperformed the girls.

[2] The boys and girls respond to a common set of (anchor) items, and the scale of the girls (reference) group is identified, which suffices for identification (see Section 7.3).

4.3.4 Ordinal Data Augmentation

Albert and Chib (1993) defined a Gibbs sampler for the univariate ordinal probit response model. They made the transition to polytomous scored items by defining the polytomous response Y_{ik} as an indicator of Z_{ik} falling into one of the response categories defined by threshold parameters $\boldsymbol{\kappa}$. In this case, an auxiliary variable Z_{ik} is defined as

$$Z_{ik} \mid Y_{ik} = c, \theta_i, \boldsymbol{\kappa}_k, a_k \sim \mathcal{N}\left(a_k\theta_i, 1\right) I\left(\kappa_{k,c-1} < Z_{ik} \leq \kappa_{k,c}\right). \tag{4.25}$$

The ordering of the response categories is displayed as

$$\kappa_{k,0} < \kappa_{k,1} \leq \kappa_{k,2} \leq \ldots < \kappa_{k,C_k}, \tag{4.26}$$

where there are C_k categories, the number of categories per item may differ, and $\kappa_{k,0} = -\infty$ and $\kappa_{k,C_k} = \infty$. The probability that an individual with ability θ_i obtains a grade $c \geq 1$ or gives a response falling into category c on item k is defined by (see also Equation (1.8))

$$\begin{aligned} P\left(Y_{ik} = c \mid \theta_i, a_k, \boldsymbol{\kappa}_k\right) &= \Phi\left(a_k\theta_i - \kappa_{k,c-1}\right) - \Phi\left(a_k\theta_i - \kappa_{k,c}\right) \\ &= \Phi\left(\kappa_{k,c} - a_k\theta_i\right) - \Phi\left(\kappa_{k,c-1} - a_k\theta_i\right) \\ &= P\left(\kappa_{k,c-1} < Z_{ik} \leq \kappa_{k,c} \mid \theta_i, a_k, \boldsymbol{\kappa}_k\right) \\ &= \int_{\kappa_{k,c-1}}^{\kappa_{k,c}} \phi(z; a_k\theta_i)\mathrm{d}z. \end{aligned} \tag{4.27}$$

This shows that (4.25) qualifies as a data augmentation scheme.

Simulating ability and discrimination parameter values is done in a similar way as for the two-parameter normal ogive model. There has been some discussion in the literature about the best way to sample the threshold parameter values (e.g., Chen et al., 2000; Cowles, 1996; Fox, 2005b; Lee and Zhu, 2000; Shi and Lee, 1998; Song and Lee, 2001). One way is to to derive the full conditional distribution using a conjugate prior that takes the order constraint in (4.26) into account. Define a uniformly distributed variable U_{ik} over $[0, 1]$ such that

$$U_{ik} \leq P\left(Z_{ik} \leq \kappa_{k,c} \mid \mathbf{y}, \boldsymbol{\kappa}_k, \theta_i, a_k\right) I\left(i \in \mathcal{A}_1\right), \tag{4.28}$$

$$U_{ik} > P\left(Z_{ik} > \kappa_{k,c} \mid \mathbf{y}, \boldsymbol{\kappa}_k, \theta_i, a_k\right) I\left(i \in \mathcal{A}_2\right), \tag{4.29}$$

for the set $\mathcal{A}_1 = \{i : Y_{ik} = c\}$ and $\mathcal{A}_2 = \{i : Y_{ik} = c+1\}$. Accordingly, the full conditional distribution of $\kappa_{k,c}$ is uniform using a (diffuse) prior with equal probability for each possible parameter value,

$$\kappa_{k,c} \mid \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\kappa}_{k(-c)}, a_k \sim \mathcal{U}\left(\Delta_l, \Delta_h\right), \tag{4.30}$$

where $\Delta_l = \max\left(\max_{i|Y_{ik}=c} z_{ik}, \kappa_{k,c-1}\right)$, $\Delta_h = \min\left(\min_{i|Y_{ik}=c+1} z_{ik}, \kappa_{k,c+1}\right)$, and $\boldsymbol{\kappa}_{k(-c)}$ is the set of threshold parameters for item k without $\kappa_{k,c}$.

This implementation (Albert and Chib, 1993) is straightforward since it consists of sampling from normal and uniform densities, but the convergence is very slow when the number of respondents is larger than 50 (see Chen et al., 2000, p. 37; Cowles, 1996). The interval in Equation (4.30) becomes narrow when increasing the number of respondents. As a result, the sampled threshold values differ slightly across iterations also due to the fact that the iterates are highly correlated. A slow convergence of the threshold parameters may also lead to a slow convergence of the other model parameters.

The convergence is accelerated by sampling threshold parameters via an M-H step (Cowles, 1996). This requires constructing a suitable proposal density. A new candidate is generated from a truncated normal density,

$$\kappa_{k,c}^{*} \sim \mathcal{N}\left(\kappa_{k,c}^{(m)}, \sigma_{mh}^{2}\right) I\left(\kappa_{k,c-1}^{*} < \kappa_{k,c}^{*} < \kappa_{k,c+1}^{(m)}\right), \tag{4.31}$$

where $\kappa_{k,c}^{(m)}$ is the value of $\kappa_{k,c}$ in the mth iteration of the sampler.

Chen et al. (2000) argued that the proposal density often is not spread out enough and, as a result, it will not generate candidate threshold values from the entire parameter space with nonzero posterior probability. Furthermore, the variance parameter of the proposal density is difficult to specify. The variance of the proposal density, σ_{mh}^{2}, must be specified appropriately to establish an efficient algorithm. Nandram and Chen (1996) defined a reparameterization that avoids the use of a truncated proposal for univariate response data. This approach is also interesting for multivariate ordinal data, in particular from an item-based test with three response options (see Exercise 4.8).

To avoid specifying any tuning parameters, an adaptive proposal is defined. The variance of this proposal density is adjusted within the sampling procedure. This fine-tuning of the proposal results in a good and efficient convergence of the algorithm without detailed prior information regarding the variance of the proposal. Specifically, say after each 50th iteration the acceptance rate regarding the threshold parameters is evaluated. If the acceptance rate is low, a high percentage of the sampled new candidates were rejected and the variance σ_{mh}^{2} is too high. The other way around, if the acceptance rate is high, a high percentage of the sampled new candidates were accepted and the variance σ_{mh}^{2} is too low.

In both situations, the variance is adjusted in the right direction. This way, a sequence of proposals is generated automatically that includes the fine-tuning of the proposal's variance parameter. In an exploratory phase, the proposal is calibrated to obtain an M-H step with an acceptable convergence rate. Gelman et al. (1996) recommended an acceptance rate close to 50% for one- or two-dimensional models.

A general MCMC scheme is presented for the graded response model. For ease of notation, the adjustment factor d is dropped, where $d = 1.7$ and $d = 1$ for the logistic and normal ogive graded response models, respectively.

MCMC Scheme 3 (Graded Response Model)

1. *Sample augmented data* $\mathbf{z}^{(m+1)}$ *for a logistic or normal ogive graded response model,*

$$Z_{ik} \mid Y_{ik}, \theta_i^{(m)}, \boldsymbol{\xi}_k^{(m)} \sim \begin{cases} \mathcal{L}(a_k\theta_i, 1) \\ \mathcal{N}(a_k\theta_i, 1), \end{cases} \tag{4.32}$$

where $Y_{ik} = c$ *if* $\kappa_{k,c-1} < Z_{ik} \leq \kappa_{k,c}$.

2. *Assume a hierarchical prior for* θ_i *(Equation* (2.8)*). Sample* $\theta_i^{(m+1)}$ *from*

$$\theta_i \mid \mathbf{z}_i^{(m+1)}, \mathbf{a}^{(m)}, \mu_\theta^{(m)}, \sigma_\theta^{2(m)} \sim \mathcal{N}\left(\Sigma_\theta \left(\mathbf{a}^t\mathbf{z}_i + \mu_\theta/\sigma_\theta^2\right), \Sigma_\theta\right), \tag{4.33}$$

where $\Sigma_\theta^{-1} = \mathbf{a}^t\mathbf{a}/\varphi + \sigma_\theta^{-2}$. *Candidates are generated and evaluated in an M-H step when* $\mathbf{Z}_i$ *is logistically distributed.*

3. *Draw candidates* $\boldsymbol{\kappa}_k^*$ *from the proposal density in Equation* (4.31)*. Sample* $U_k \sim \mathcal{U}(0,1)$*, and set* $\kappa_{k,c}^{(m+1)} = \kappa_{k,c}^*$ *for* $c = 1, \ldots, C_k - 1$ *when*

$$u_k \leq \min\left[\prod_i \frac{F\left(a_k\theta_i - \kappa_{k,y_{ik}-1}^*\right) - F\left(a_k\theta_i - \kappa_{k,y_{ik}}^*\right)}{F\left(a_k\theta_i - \kappa_{k,y_{ik}-1}\right) - F\left(a_k\theta_i - \kappa_{k,y_{ik}}\right)} \prod_{c=1}^{C_k-1} \frac{\Phi\left(\frac{\kappa_{k,c+1}-\kappa_{k,c}}{\sigma_{mh}}\right) - \Phi\left(\frac{\kappa_{k,c-1}^*-\kappa_{k,c}}{\sigma_{mh}}\right)}{\Phi\left(\frac{\kappa_{k,c+1}^*-\kappa_{k,c}^*}{\sigma_{mh}}\right) - \Phi\left(\frac{\kappa_{k,c-1}-\kappa_{k,c}^*}{\sigma_{mh}}\right)}, 1\right], \tag{4.34}$$

where F *denotes the corresponding cumulative logistic or normal distribution function.*

4. *Assume a positively truncated normal prior for* a_k,

$$a_k \mid \mu_a, \sigma_a^2 \sim \mathcal{N}\left(\mu_a, \sigma_a^2\right) I_{\mathcal{A}_k}(a_k),$$

where $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$. *Let* $\mathbf{x} = \boldsymbol{\theta}^{(m+1)}$*, and sample* $a_k^{(m+1)}$ *from*

$$a_k \mid \mathbf{z}_k^{(m+1)}, \boldsymbol{\theta}^{(m+1)}, \mu_a^{(m)}, \sigma_a^{2(m)} \sim \mathcal{N}\left(\mu_a^*, \Omega_a\right) I_{\mathcal{A}_k}(a_k), \tag{4.35}$$

where

$$\Omega_a^{-1} = \varphi^{-1}\left(\mathbf{x}^t\mathbf{x}\right) + \sigma_a^{-2}, \tag{4.36}$$

$$\mu_a^* = \Omega_a\left(\mathbf{x}^t\mathbf{z}_k + \mu_a\sigma_a^{-2}\right). \tag{4.37}$$

Candidates are generated when $\mathbf{Z}_k$ *is logistically distributed and evaluated in an M-H step.*

5. *Sample values of hyperparameters* $\left(\mu_\theta, \sigma_\theta^2\right)$ *from their full conditionals; see step 3 in MCMC scheme 1.*

In steps 2, 3, and 4, $\varphi = 1$ or $\varphi = \pi^2/3$ if $\mathbf{Z}_i$ is normally distributed or logistically distributed, respectively. In step 3, the candidate threshold parameters per item are evaluated and a variance parameter σ_{mh}^2 per item is

also calibrated. The first term of the acceptance probability in Equation (4.34) represents the ratio of posterior probabilities of the candidate threshold values to the threshold values from the last iteration. The second term in Equation (4.34) accounts for the nonsymmetric proposal density, which might favor some parameter values over others (see Equation (3.1)). This difference lies in the normalization constant of the proposal density according to Equation (4.31).

The positivity restriction on the discrimination parameters hinders a direct sampling approach for the hyperparameters μ_a and σ_a^2. Alternative sampling approaches are discussed in Exercises 3.7 and 4.4 and in Chapter 7.

4.4 Identification of Item Response Models

It was shown that the observed response data can be viewed as realizations of an underlying latent variable. Assume the two-parameter likelihood model for augmented response data according to Equation (4.7), which is given by

$$Z_{ik} = a_k\theta_i - b_k + \epsilon_{ik}, \tag{4.38}$$

where ϵ_{ik} are independent and standard normally or standard logistically distributed. This item response model is overparameterized since it has more parameters than can be estimated from the data. For example, for any constant c, $\widetilde{\theta}_i = a_k\theta_i + c$ and $\widetilde{b}_k = b_k + c$ give the same probability of a correct response since $\widetilde{\theta}_i - \widetilde{b}_k = a_k\theta_i - b_k$. That is, both parameterizations lead to the same success probability,

$$\begin{aligned} P\left(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k\right) &= P\left(Z_{ik} > 0 \mid \theta_i, \boldsymbol{\xi}_k\right) \\ &= P\left(a_k\theta_i - b_k + \epsilon_{ik} > 0\right) \\ &= P\left(\widetilde{\theta}_i - \widetilde{b}_k + \epsilon_{ik} > 0\right) \\ &= P\left(Z_{ik} > 0 \mid \widetilde{\theta}_i, \widetilde{b}_k\right) \\ &= P\left(Y_{ik} = 1 \mid \widetilde{\theta}_i, \widetilde{b}_k\right). \end{aligned}$$

In the same way, it can be shown that for any constant c, $\widetilde{\theta}_i = \theta_i/c$ and $\widetilde{a}_k = a_k c$ give the same probability of a correct response.

The metric (location and scale) of the person parameters is only known up to a linear transformation (Lord and Novick, 1968, p. 366). In the literature, several types of restrictions are proposed to anchor the metric. The location can be identified by fixing the ability level of a specific person, a so-called standard person (Rasch, 1960), or by fixing the mean population level of ability to zero. Items are often applied to different sets of individuals, and therefore it can be more convenient to put a constraint on the difficulty parameters. This is done by selecting one item as the standard item and fixing the item

difficulty parameter to a specific value, most often zero, or restricting the sum of item difficulty parameters to zero. If necessary, the scale of the metric is identified by fixing the population variance of the abilities. The scale can also be identified by restricting the discrimination parameter of a so-called standard item to a specific value, most often one, or restricting the product of discrimination parameters to one.

Bayesian item response models usually are not identified by restricting a single parameter since this complicates the specification of priors and conditional posteriors in MCMC algorithms. Furthermore, if the restricted single parameter is poorly identified, the standard errors of the estimated parameters go up. In an informal way, it will be shown that Bayesian item response models can be identified (1) by imposing restrictions on the hyperparameters or (2) via a (standard) scale transformation in the estimation procedure.

4.4.1 Data Augmentation and Identifying Assumptions

The introduction of an underlying latent variable Z_{ik} already induces an identification problem. Three identifying restrictions were introduced for the two-parameter likelihood model for augmented response data (Equation (4.38)). First, in general, the latent response data can be linked to the observed response data in such a way that $Z_{ik} > \kappa$ if $Y_{ik} = 1$ and $Z_{ik} \leq \kappa$ if $Y_{ik} = 0$, where κ is the cutpoint. In Equation (4.7), this cutpoint or threshold parameter is restricted to zero. For an unrestricted threshold parameter, the probability of a correct response is given by

$$\begin{aligned} P\left(Y_{ik} = 1 \mid \theta_i\right) &= P\left(Z_{ik} > \kappa \mid \theta_i\right) \\ &= P\left(a_k\theta_i - b_k + \epsilon_{ik} > \kappa \mid \theta_i\right) \\ &= P\left(\epsilon_{ik} > (b_k + \kappa) - a_k\theta_i \mid \theta_i\right). \end{aligned}$$

It can be seen that a change in the threshold parameter can always be compensated for by a corresponding change in the difficulty parameter, and the model is unidentified.

Second, the conditional mean of the error term is restricted to zero, $E(\epsilon_{ik} \mid \theta_i) = 0$. Assume that $E\left(\epsilon_{ik} \mid \theta_i\right) = \mu_\epsilon$. Then

$$\begin{aligned} P\left(Y_{ik} = 1 \mid \theta_i\right) &= P\left(\epsilon_{ik} > b_k - a_k\theta_i \mid \theta_i\right) \\ &= F\left(a_k\theta_i - (b_k - \mu_\epsilon) \mid \theta_i\right), \end{aligned}$$

where F is a standard cumulative (normal or logistic) distribution function. In the same way as above, a change in μ_ϵ can be compensated for by a similar change in b_k, and the model is not identified.

Third, the conditional variance of the residuals is restricted to one, $\mathrm{Var}(\epsilon_{ik} \mid \theta_i) = 1$ in the probit model and $\mathrm{Var}(\epsilon_{ik} \mid \theta_i) = \pi^2/3$ in the logit model. Assume that the variance is parameterized as φ. Then

$$\begin{aligned} P(Y_{ik} = 1 \mid \theta_i) &= P(\epsilon_{ik} > b_k - a_k\theta_i \mid \theta_i) \\ &= P(\epsilon_{ik} > a_k(b_k^* - \theta_i) \mid \theta_i) \\ &= F\left(\frac{a_k}{\sqrt{\varphi}}(\theta_i - b_k^*) \mid \theta_i\right), \end{aligned}$$

and the model is not identified since a change in the discrimination parameter can be compensated for by a similar change in $\sqrt{\varphi}$.

4.4.2 Rescaling and Priors with Identifying Restrictions

The two-parameter likelihood model for the augmented responses is identified due to the three assumptions mentioned given that the metric of θ_i is known. The metric of person parameters is usually unknown, and additional restrictions are needed to identify the model. One approach is to identify the metric by fixing the (hyper)parameters of the prior for the person parameters. A common assumption is to set $\mu_\theta = 0$ and $\sigma_\theta^2 = 1$ such that the person parameters are standard normally distributed. Then, the two-parameter model for the latent response is given by

$$Z_{ik} = a_k\epsilon_\theta - b_k + \epsilon_{ik},$$

where ϵ_θ and ϵ_{ik} are independent standard normally distributed. The random term ϵ_θ describes the between-individual heterogeneity and ϵ_{ik} the within-individual residual variation.

In a different approach, the hyperparameters are freely specified and the model is identified by establishing a metric in each MCMC iteration. That is, the sampled person parameter values are rescaled to an a priori specified metric. The metric of sampled values $\theta_i^{(m)}$ is easily changed via a linear transformation. When $\tilde{\theta}_i^{(m)} = \sigma_\theta\theta_i^{(m)} + \mu_\theta$, it follows that

$$\begin{aligned} E\left(\tilde{\theta}_i^{(m)}\right) &= E\left(\sigma_\theta\theta_i^{(m)} + \mu_\theta\right) = \sigma_\theta E\left(\theta_i^{(m)}\right) + \mu_\theta, \\ \mathrm{Var}\left(\tilde{\theta}_i^{(m)}\right) &= \mathrm{Var}\left(\sigma_\theta\theta_i^{(m)} + \mu_\theta\right) = \sigma_\theta^2\,\mathrm{Var}\left(\theta_i^{(m)}\right), \end{aligned}$$

and when $\theta_i^{(m)}$ is sampled from $f\big(\theta_i^{(m)} \mid \mathbf{y}\big)$, then $\tilde{\theta}_i^{(m)}$ is sampled from $f\big((\tilde{\theta}_i^{(m)} - \mu_\theta)/\sigma_\theta \mid \mathbf{y}\big)/\sigma_\theta$. It follows that each sampled value from the full conditional posterior in iteration m, say $\theta_i^{(m)}$, can be linearly transformed to have, for example, mean zero and variance one. In subsequent sampling steps but in the same iteration of the MCMC algorithm, the same rescaled vector is used. As a result, the identification problem is solved since the sampled person parameters have a fixed metric.

It is also possible to rescale the sample of item parameter values in each MCMC iteration. The sampled difficulty values can be rescaled such that, for example, their sum equals zero. Define $\mu_b = \sum_k b_k^{(m)}/K$ and transform the

sample at iteration m as $\tilde{b}_k^{(m)} = b_k^{(m)} - \mu_b$. This scale transformation sets the location. In the same way, the sampled discrimination values can be rescaled such that, for example, their product equals one. Let $\sigma_a = \prod_k a_k^{(m)}$, and transform the sample at iteration m as $\tilde{a}_k^{(m)} = a_k^{(m)}(1/\sigma_a)^{1/K}$. This restriction sets the scale of the metric. In the same iteration, the rescaled sample of item parameter values is used to sample the person parameters.

This identification procedure has the advantage that complicated response models with multistage priors can be identified via simple transformations of samples drawn at each MCMC iteration. Furthermore, it is difficult to identify a complex model by fixing parameters of a multistage prior since the location and scale are not directly parameterized (e.g., Chapters 6 and 7).

This technique corresponds closely to the identification procedure in Bilog-MG (Zimowski et al., 1996). The EM algorithm implemented uses Gauss-Hermite quadrature to integrate over the ability distribution. Initially the quadrature nodes and weights are based on an approximation to the prior ability distribution. In the first iteration, the metric of the estimated item parameters obtained in the M-step of the algorithm is identified by the metric of the prior ability distribution. In the next E-step, the posterior ability distribution is used to recompute new quadrature weights. The recomputed weights are adjusted to have a mean of zero and a variance of one. Thus, the metric of the estimated item parameters from the next M-step is identified by the rescaled posterior ability distribution, which has a location parameter of zero and a scale parameter of one. In each EM cycle, the recomputed quadrature weights based on the posterior ability distribution are rescaled to identify the model. In a similar way, in each iteration of the MCMC algorithm, the ability values drawn from the posterior ability distribution are rescaled to identify the model. Baker and Kim (2004, Chapter 6) give more details about how the identification problem is handled in the EM procedure.

4.5 Performance MCMC Schemes

4.5.1 Item Parameter Recovery

The performance of MCMC scheme 2 was compared with the performance of the WinBUGS (Lunn et al., 2000) and the Bilog-MG (Zimowski et al., 1996) programs. The convergence properties of the MCMC output of scheme 2 and WinBUGS were compared, and the accuracy of the item parameter estimates from the three programs was investigated. Data were generated under the two-parameter logistic model for K=10 items and N=1,000 persons. The prior model consists of the hierarchical prior for the item parameters (Equations (2.3)–(2.5)) and the hierarchical prior in (2.8) for the ability parameters. The parameters of the ability prior were fixed to mean zero and variance one to identify the model. Note that it is not possible in WinBUGS to fix the

mean and variance in each iteration since identification restrictions have to be implemented in the model.

Three separate runs of 20,000 iterations were performed using MCMC scheme 2 (implemented in Fortran) and WinBUGS to investigate the sensitivity of the starting values. For each run, starting values were generated from the priors.

In Figure 4.2, the first 10,000 MCMC iterates corresponding to the parameters of item 4 generated from scheme 2 of three different runs are plotted. In each iteration and each run, a running mean was computed that presents the mean of sampled values up to that iteration number. The three different lines in each plot present the running means as a function of the MCMC iterates drawn for the three different runs. It can be seen that within 1,000 iterations the running means from the different runs almost coincide. Similar plots were obtained for the other item parameters.

The first-order autocorrelation or serial correlation between the MCMC iterates was calculated for each MCMC chain, where the iterates from the burn-in period were ignored (see Equation 3.3). The averaged first-order autocorrelation was .827 for the sampled discrimination parameter values and .643 for the sampled difficulty parameter values. It takes less time to explore the entire posterior density of the difficulty parameter since the MCMC sample corresponding to the difficulty parameter exhibited less first-order autocorrelation. A high first-order autocorrelation indicates that longer runs are necessary to assure that all areas of the posterior density are reached. Furthermore, if the level of autocorrelation is high, a trace plot will be a poor diagnostic for convergence.

In Figure 4.3, the MCMC iterates of the parameters of item four are plotted from three different runs in WinBUGS. The running means of the three different runs differ greatly, and it takes at least 5,000 iterations for the running means to be of comparable size. After 5,000 iterations, further samples just slightly influence the calculation of the mean. Comparable plots were obtained using the samples of the other item parameters. It follows that in WinBUGS the starting values have a much higher impact on the values drawn in comparison with the values drawn from scheme 2. The averaged first-order autocorrelations of both MCMC samples from WinBUGS are comparable and indicate that a more efficient sample is obtained from scheme 2.

In Table 4.2, Geweke's and Gelman and Rubin's convergence diagnostic values and different levels of autocorrelations are given for the MCMC samples of all item parameters from WinBUGS and scheme 2. The convergence diagnostic of Geweke (1992) compares the means of the sampled values at two different stages of the sequence. The difference between the two means is assumed to be asymptotically normally distributed, and a difference that is located in the tail of the distribution provides evidence against convergence. The corresponding p-values in Table 4.2 show that the MCMC samples of discrimination parameter 7 and difficulty parameters 4, 7, and 8 may not have reached convergence when using a significance level of 5%. Gelman and Ru-

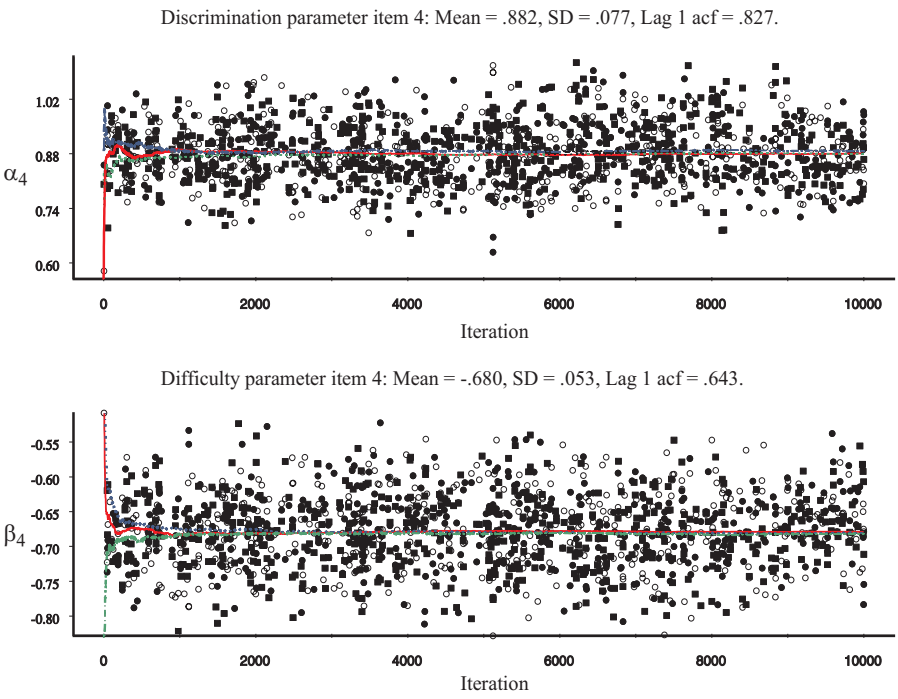

Fig. 4.2. MCMC iterates from three parallel chains of the discrimination and difficulty parameters from item 4 using scheme 2.

bin's convergence diagnostic shows no indication of nonconvergence for any of the MCMC chains. This scale reduction factor should be close to one if the sampler is close to the target distribution. For the discrimination parameter, comparable estimated autocorrelations within chains were found with both programs. After approximately a lag of 50, the iterations within a chain are independent. For the difficulty parameter, WinBUGS produces stronger within-chain autocorrelated samples when looking at the estimated autocorrelation of lags 1 and 5. Cross-correlations indicate the amount of correlation between the parameter values drawn, and high correlations lead to slow convergence. The MCMC samples did not exhibit high cross-correlations.

Finally, the parameter estimates and standard deviations from the different programs are reported in Table 4.3. Despite the autocorrelations, all sampled values obtained after a burn-in period of 5,000 iterations were used for posterior summarization. Bilog-MG assumes a lognormal and a normal prior for each discrimination and each difficulty parameter, respectively. This program computes MAP (maximum a posteriori, posterior mode) estimates for the item parameters. In Table 4.3, it can be seen that the estimated values do not differ much from the EAP estimates that are computed via WinBUGS and MCMC scheme 2, and most differences occur only in the second or third

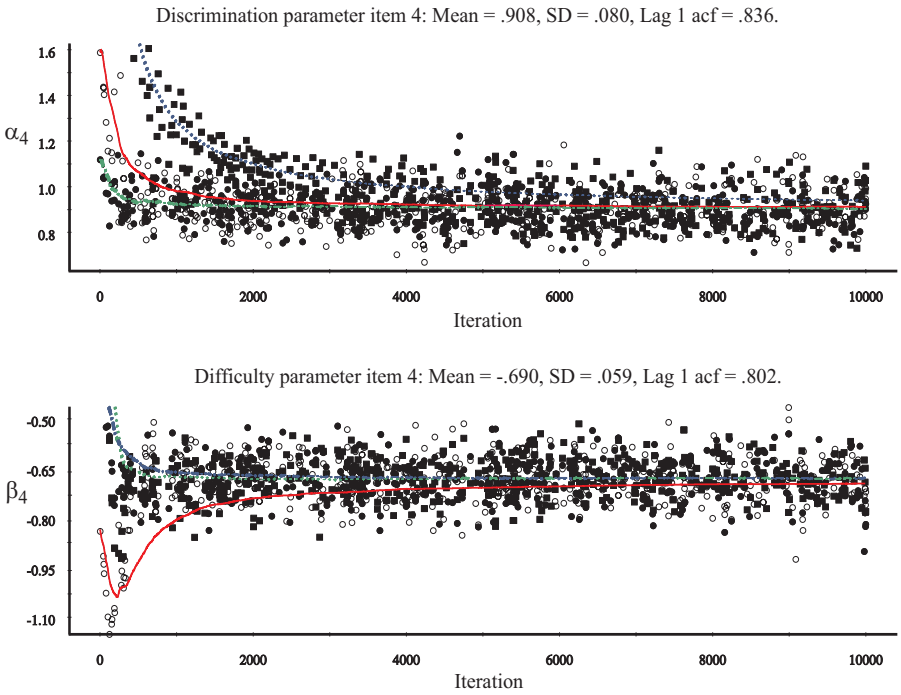

Fig. 4.3. MCMC iterates from three parallel chains of the discrimination and difficulty parameters from item 4 using WinBUGS.

decimal places. Baker and Kim (2004, Chapter 12) and Kim (2001), among others, compared the Gibbs sampling method of Albert (1992) with Bilog-MG and a predecessor of WinBUGS, and they obtained comparable item and ability estimates for different datasets. From the present study, it follows that MCMC scheme 2 produces a more efficient MCMC sample and uses less computation time (in Fortran). Inspection of the trace plots also showed that the choice of starting values hardly influences the run of the chain in contrast to the sequences produced by WinBUGS.

4.5.2 Hierarchical Priors and Shrinkage

Item response data were generated according to a normal ogive model with a standard normal prior for the ability parameters and the hierarchical prior for the item parameters (Equations (2.3)–(2.5)). A within-item correlation structure of .30 was assumed. Data were simulated for 100 persons responding to 50 items and 1,000 persons responding to 10 items.

The item parameters were obtained for each dataset using the normal ogive model. Two different priors for the item parameters were used: first, the hierarchical prior, where the hyperparameters were estimated from the data

Table 4.2. Convergence properties of the Gibbs sampling chains from scheme 2 and the MCMC chains from WinBUGS.

	MCMC scheme 2					WinBUGS				
	Geweke	Autocorrelation			G&R	Geweke	Autocorrelation			G&R
Item	p-value	Lag 1	Lag 5	Lag 50	SRF	p-value	Lag 1	Lag 5	Lag 50	SRF
Discrimination Parameter										
1	.590	.892	.603	−.001	1.000	.386	.844	.444	.036	1.001
2	.526	.866	.528	.000	1.000	.146	.836	.402	.029	1.003
3	.868	.923	.713	.081	0.999	.371	.885	.579	−.019	1.000
4	.105	.827	.453	−.028	1.000	.303	.836	.441	.000	1.002
5	.078	.794	.406	−.007	1.000	.837	.835	.441	−.017	1.001
6	.211	.722	.255	.016	1.000	.768	.797	.347	.005	1.004
7	.927	.777	.328	−.017	1.000	.995	.797	.308	.024	1.001
8	.500	.858	.506	−.014	0.999	.899	.819	.388	−.005	1.000
9	.065	.848	.520	.012	1.001	.247	.848	.477	−.005	1.001
10	.108	.793	.350	.020	1.001	.754	.821	.397	−.017	1.000
Difficulty Parameter										
1	.933	.645	.187	.071	1.000	.174	.833	.505	.039	1.001
2	.391	.603	.151	.046	1.000	.869	.809	.432	.045	1.001
3	.928	.776	.448	.318	1.000	.245	.860	.552	.038	1.000
4	.416	.643	.214	.100	1.000	.008	.802	.431	.041	1.000
5	.117	.646	.222	.116	1.000	.877	.810	.435	.026	1.001
6	.085	.501	.058	.015	1.000	.947	.760	.327	.030	1.002
7	.054	.487	.037	.000	1.000	.014	.760	.335	.034	1.000
8	.096	.609	.137	.063	1.000	.096	.803	.426	.032	1.000
9	.143	.703	.288	.150	1.000	.414	.829	.473	.046	1.003
10	.058	.588	.145	.050	1.000	.108	.809	.415	.043	1.000

that handle the amount of shrinkage in the posterior means, and second, an uninformative independent normal prior for the item parameters with a large variance, where the discrimination parameters were restricted to be positive. Thus, in the second case, the hyperparameters were fixed a priori, which means that the amount of shrinkage was also settled a priori.

Different runs of 20,000 iterations were performed using MCMC scheme 2. The first 1,000 iterations were considered as a burn-in period. The sampled item parameter values under the different priors were used to compute mean squared errors (MSEs). In Figure 4.4, the estimated MSEs of the item parameters for two different prior structures are plotted for the dataset of 50 items and 100 persons. It can be seen that estimated MSEs of the discrimination and difficulty parameters corresponding to the independent prior are in most cases greater than the MSE estimates corresponding to the hierarchical prior. The accuracy of each item parameter estimate was improved using a hierarchical prior by borrowing information from the other item parameter estimates,

Table 4.3. Item parameter estimates of the two-parameter logistic response model.

		MCMC 2		Bilog-MG		WinBUGS	
Item	Simulated	Mean	SD	Mean	SD	Mean	SD
Discrimination Parameter							
1	1.150	1.167	.096	1.132	.101	1.168	.110
2	1.034	1.067	.086	1.021	.092	1.061	.094
3	1.372	1.273	.114	1.233	.113	1.264	.114
4	.873	.882	.077	.869	.081	.908	.080
5	.856	.792	.070	.776	.073	.799	.072
6	.629	.623	.060	.605	.058	.640	.059
7	.890	.792	.067	.752	.068	.792	.074
8	1.057	1.061	.084	1.020	.093	1.074	.089
9	1.087	.936	.081	.922	.083	.951	.087
10	.838	.836	.070	.816	.075	.836	.071
Difficulty Parameter							
1	−.505	−.477	.063	−.485	.067	−.489	.068
2	−.376	−.390	.060	−.388	.061	−.391	.060
3	.796	.815	.067	.825	.081	.800	.076
4	−.560	−.680	.063	−.685	.063	−.690	.059
5	.768	.755	.063	.756	.062	.740	.062
6	.314	.419	.044	.416	.049	.409	.048
7	.163	.156	.044	.154	.050	.152	.051
8	−.519	−.444	.051	−.445	.062	−.454	.058
9	.807	.811	.058	.822	.069	.806	.067
10	.471	.528	.049	.526	.058	.513	.056

where the amount of shrinkage was inferred from the data. Furthermore, the within-item dependency improved the item parameter estimates. The independent prior induced almost no shrinkage, and the item parameter estimates were based only on the 100 item responses.

In Figure 4.5, the MSEs are plotted for the dataset of 10 items and 1,000 persons. The estimated differences in MSEs corresponding to the independent and hierarchical priors are very small and not systematic. The estimated shrinkage effects are very small for both priors. The accuracy of the item parameter estimates was substantially improved by increasing the number of respondents, and the prior influence became very small. Furthermore, a less accurate estimate of the within-item covariance structure was obtained by decreasing the number of items. It can be concluded that the hierarchical prior is useful for relatively small datasets when prior information significantly influences the item parameter estimates.

Assume the object was to estimate the person parameters. Then, a hierarchical prior for the person parameters supports shrinkage effects on the

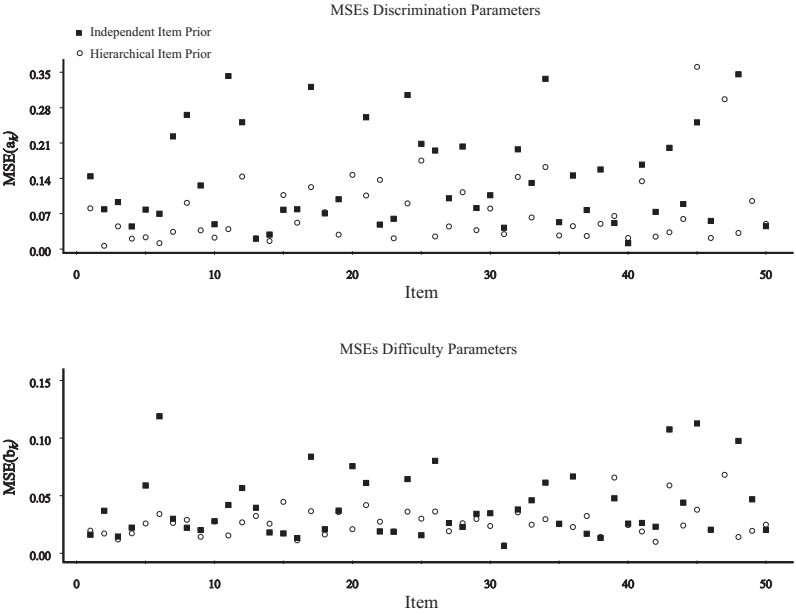

Fig. 4.4. MSEs of item parameter estimates using an independent item prior and a hierarchical item prior with N=100 and K=50.

person parameter estimates. There are many persons with 10 responses, and improved person parameter estimates will be obtained by pooling information from other respondents. This is in comparison with the dataset of 50 items, where accurate person estimates are obtained from the data and the prior influence is small.

4.6 European Social Survey: Measuring Political Interest

The European Social Survey (ESS; `http://www.europeansocialsurvey.org`) has taken place every two years starting in 2001 and covers over 30 countries. The objective of the ESS is to gather data about attitudes, attributes, and behavior patterns; measure and explain people's social values, cultural norms, and behavior patterns, and explore the way in which they differ within and between nations and the direction and speed at which they are changing.

Attention is focused on three items measuring political interest from the main questionnaire of 2006 on the block political engagement. The three items are: (1) How interested would you say you are in politics (very, quite, hardly, or not at all interested)? (2) How often does politics seem so complicated that you can't really understand what is going on (never, seldom, occasionally, regularly, frequently)? (3) Do you think that you could take an active role

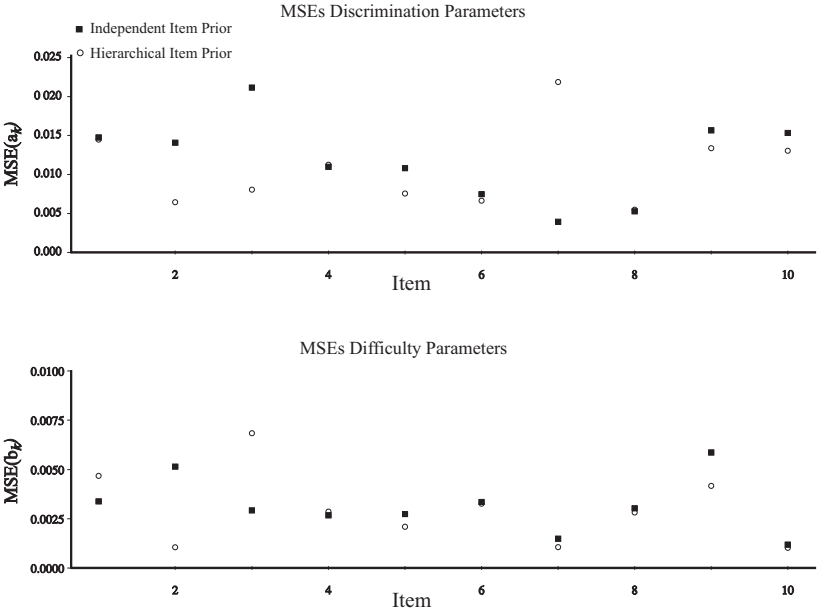

Fig. 4.5. MSEs of item parameter estimates using an independent item prior and a hierarchical item prior with N=1,000 and K=10.

in a group involved with political issues (definitely, probably, not sure either way, probably not, definitely not)?

The answers from the 2,364 Dutch respondents were analyzed with the graded response model. The latent variable, θ, presents the political interest and has a prior density according to Equations (2.8)–(2.10). The threshold parameters are order restricted a priori and uniformly distributed, and the discrimination parameters have a common positively truncated normal prior. MCMC scheme 3 was run for 10,000 iterations, where the first 1,000 iterations were regarded as a burn-in period.

In Figure 4.6, for each item, the first 1,000 MCMC iterates of the threshold parameters are plotted. The threshold parameters are ordered, which follows from the stepwise pattern, where the first threshold parameter separates the first two response options. The starting values were noninformative ordered integer values beginning at zero, and the proposal density variance was set at .05. It can be seen that item 1 has three threshold parameters (four response categories) and items 2 and 3 four threshold parameters each (five response categories). Although an M-H algorithm was used to sample values, the iterates of each threshold parameter converged very quickly to the highest posterior density region.

The estimated discrimination parameters are, respectively, .942 (.05), 1.040 (.06), and .974 (.05), where the posterior standard deviations are given in parentheses. In Figure 4.7, the estimated category response curves are plotted

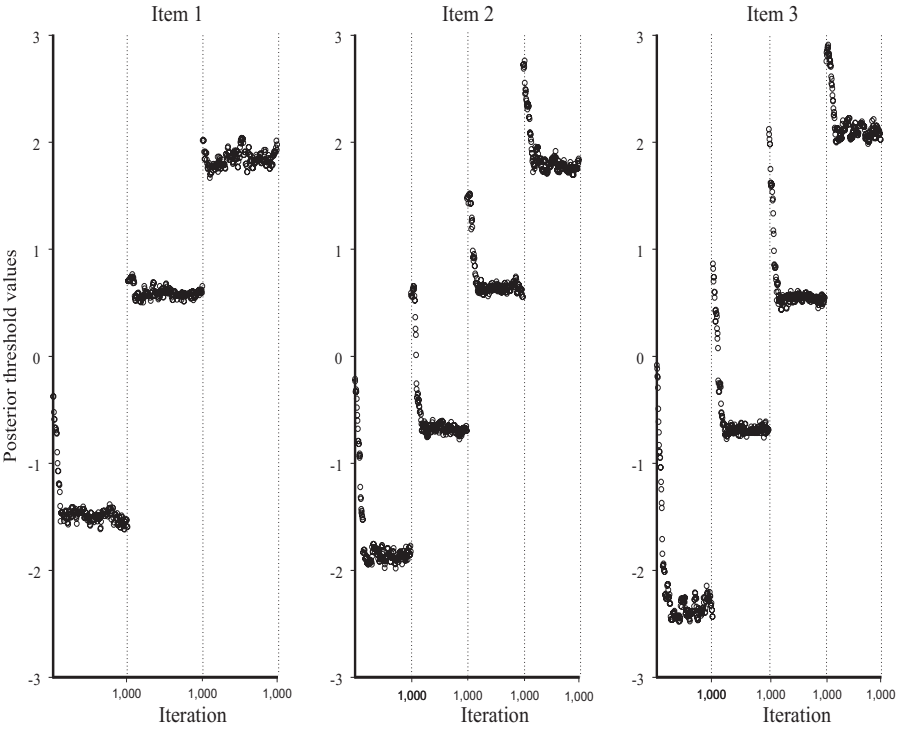

Fig. 4.6. ESS 2006: For each item, the first 1,000 MCMC iterates of the threshold parameters.

for three political items. Each category response curve represents the probability, denoted as $P_{kc}(\theta)$, of scoring in the particular category $(c = 1, \ldots, C_k)$ conditional on the level of θ. For the population mean level of $\theta = 0$, category 2 has the highest response probability for item 1 and category 3 for items 2 and 3. The estimated discrimination parameters are comparable, which means that the category response curves are similarly shaped (see Figure 4.7). Therefore, the category response curves differentiate across items in a similar way among levels of the latent variable.

In Figure 4.7, for each item, the density of the latent responses (see Equation (4.32)) is plotted (dotted lines). The corresponding estimated threshold parameters plotted show the range of latent response values that lead to a response in a certain category. It follows that only a few respondents selected the first and last response categories of item 3 such that only a few respondents are absolutely certain about playing an active role in a political group.

Posterior probability statements can be estimated using the sampled values. For example, the posterior probability of a Dutch person with below average interest in politics who finds politics never or rarely too complicated to understand equals

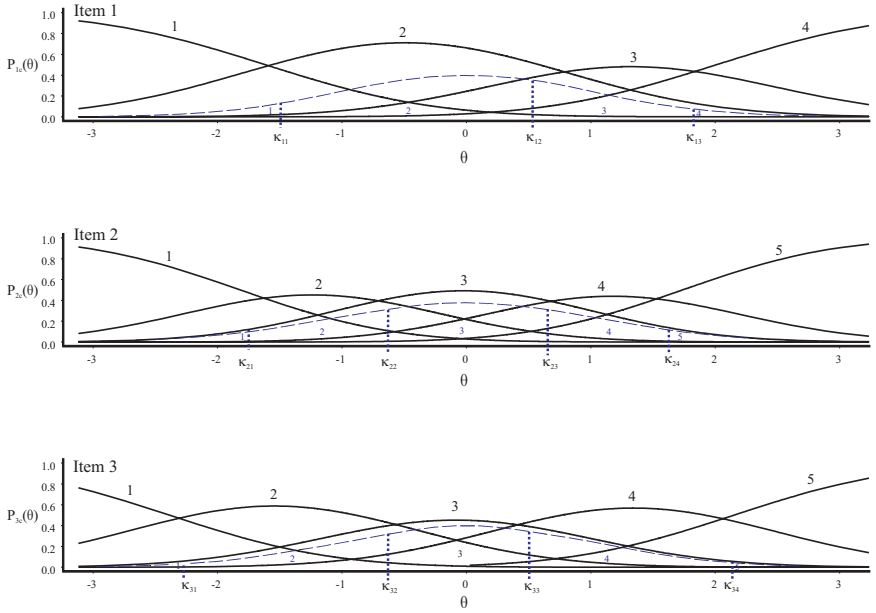

Fig. 4.7. ESS 2006: Category response curves of the three items under the graded response model.

$$
\begin{aligned}
P\left(Y_{rep} \leq 2 \mid \mathbf{y}\right) &= \int_0^\infty P\left(Z \leq \kappa_{2,3} - a_2\theta \mid \theta\right) p(\theta \mid \mathbf{y}) \mathrm{d}\theta \\
&= \int_0^\infty \Phi\left(\kappa_{2,3} - a_2\theta\right) p(\theta \mid \mathbf{y}) \mathrm{d}\theta \\
&\approx 1/M \sum_m \Phi\left(\kappa_{2,3} - a_2\theta^{(m)}\right) = .101, \qquad (4.39)
\end{aligned}
$$

where $\theta^{(m)} > 0$ $(m = 1, \ldots, M)$ is an MCMC sample from the marginal posterior density. The posterior probability is estimated given estimated item parameters. It is also possible to estimate the marginal posterior probability by integrating over the item parameter density using the MCMC samples. Note that in this case the normal distribution of the underlying latent response is used to compute the posterior probability. The observation y_{rep} is a (posterior) predictive observation under the model given the observed data $\mathbf{y}$. In Chapter 5, a more complete treatment of predictions based on posterior predictive densities is given.

4.7 Discussion and Further Reading

Bayesian estimation methods in the area of item response modeling started with applications of ability estimation under the assumption of known item

parameters. Bayes estimates of ability parameters were obtained by Birnbaum (1969) and Owen (1975). In the 1980s, Bayesian estimation methods were mainly focused on finding the posterior modes of both item and person parameters and relied heavily on numerical routines such as Gauss-Hermite quadrature and Newton-Raphson. The pioneering work of Mislevy (1986), Swaminathan and Gifford (1982, 1985, 1986), Tsutakawa (1984), Tsutakawa and Lin (1986), and Tsutakawa and Soltys (1988) stimulated the Bayesian approach to item response modeling and reflects the discussion of the marginal estimation approach versus the joint estimation approach. Baker and Kim (2004, Chapter 7) give an overview of Bayesian parameter estimation methods for item response models that were developed in the 1980s.

The feasibility of the Bayesian approach was often questioned due to the computational burden. This turned around in the 1990s. With the introduction of MCMC, the Bayesian methods became popular. In the 1990s, item response modeling applications using the MCMC methodology appeared (Albert, 1992; Bradlow et al., 1999; Patz and Junker, 1999a). The behavior of MCMC methods, parameter recovery, the computation time, and the convergence properties were investigated by Baker and Kim (2004), Kim (2001), and Patz and Junker (1999a), among others. As the complexity increases, MCMC methods become more attractive since they are based on simulations instead of exact numerical methods, and MCMC methods become particularly useful when data are sparse and/or asymptotic theory is unlikely to hold. In subsequent chapters, it will be shown that the MCMC schemes discussed are easily extended to more complex response models.

MCMC estimation methods for item response models have become increasingly common in different research areas, and several applications will be presented. Nowadays, different MCMC implementations are proposed for various item response models, and only a few are mentioned here besides the MCMC schemes discussed. There are several MCMC implementations for estimating the parameters of item response models for polytomous response data. Patz and Junker (1999a), among others, described an implementation of an M-H within Gibbs scheme for the generalized partial credit model (GPC; Muraki, 1992). Johnson and Albert (1999) generalized the M-H within Gibbs scheme of Cowles (1996) for the ordinal probit model to a scheme for the normal ogive graded response model. Béguin and Glas (2001) developed an MCMC algorithm for multidimensional item response models. Rupp, Dey and Zumbo (2004) give a short overview of item response modeling applications using MCMC estimation methods.

4.8 Exercises

4.1. According to (4.3), assume a lognormal prior for a discrimination parameter, $\log a_k \sim \mathcal{N}(\mu_a, \sigma_a^2)$.
(a) Show that the density function of a_k equals

$$p(a_k) = \frac{1}{a_k \sigma_a \sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma_a^2}\big(\log a_k - \mu_a\big)^2\right).$$

(b) Let $a_k = \exp\big(\tilde{a}_k\big)$. Then the moments of a_k can be obtained via the moment-generating function of a normal density, which is

$$E(\exp(t\tilde{a}_k)) = \exp(t\mu_a + t\sigma_a^2/2), \tag{4.40}$$

and the kth moment is equal to the kth derivative of the right-hand side of (4.40) evaluated at $t = 0$. Show that the mean and variance of the lognormally distributed variable a_k equal

$$E(a_k) = \exp\left(\mu_a + \sigma_a^2/2\right), \tag{4.41}$$

$$\text{Var}(a_k) = \exp\left(2\mu_a + \sigma_a^2\right)\left(\exp\big(\sigma_a^2\big) - 1\right). \tag{4.42}$$

(c) Derive equations for the prior parameters $\big(\mu_a, \sigma_a^2\big)$ from (4.41) and (4.42) that can be used to assign prior parameter values for specific values of $E(a_k)$ and $\text{Var}(a_k)$.
(d) Show how to generate values $a_k^{(m)}$ with mean 1 and variance .5 when $a_k = \exp\big(\tilde{a}_k\big)$ and $\tilde{a}_k \sim \mathcal{N}\big(\mu_a, \sigma_a^2\big)$.

4.2. According to Equations (2.8)–(2.10), let $g_1 = n_0/2$ and $g_2 = n_0\sigma_0^2/2$. It can be shown that the conditional density $p(\mu_\theta, \sigma_\theta^2 \mid \boldsymbol{\theta})$ is a normal inverse gamma.
(a) Show that the prior $p(\mu_\theta, \sigma_\theta^2)$ and the conditional density $p(\boldsymbol{\theta} \mid \mu_\theta, \sigma_\theta^2)$ are proportional to, respectively,

$$\big(\sigma_\theta^2\big)^{-\left(\frac{n_0+1}{2}+1\right)} \exp\left(-\frac{1}{2\sigma_\theta^2}\left(n_0\sigma_0^2 + n_0(\mu_\theta - \mu_0)^2\right)\right) \tag{4.43}$$

and

$$\big(\sigma_\theta^2\big)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_\theta^2}\left((n-1)s^2 + n(\overline{\theta} - \mu_\theta)^2\right)\right), \tag{4.44}$$

where $\bar{\theta} = \sum_i \theta_i/n$ and $s^2 = \sum_i (\theta_i - \mu_\theta)^2/(n-1)$.
(b) Combine the mean structures of Equations (4.43) and (4.44) and show that they can be rewritten as

$$n_0(\mu_\theta - \mu_0)^2 + n(\mu_\theta - \overline{\theta})^2 = (n + n_0)\left(\mu_\theta - \mu_n\right)^2 + \frac{nn_0}{n + n_0}(\overline{\theta} - \mu_0)^2, \tag{4.45}$$

where $\mu_n = \frac{n}{n+n_0}\overline{\theta} + \frac{n_0}{n+n_0}\mu_0$.

(c) Derive the parameters of the normal density $p(\mu_\theta \mid \sigma_\theta^2, \boldsymbol{\theta})$.
(d) Collect terms of interest from (4.43)–(4.45), and derive the parameters of the inverse gamma density $p(\sigma_\theta^2 \mid \boldsymbol{\theta})$.

4.3. Assume that $\tilde{\boldsymbol{\xi}}_k = (\log(a_k), b_k)^t$ and $\tilde{\boldsymbol{\xi}}_k \sim \mathcal{N}(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$, and let (2.4) and (2.5) specify the hyperprior densities.
(a) Show via matrix algebra that the conditional density

$$p\left(\tilde{\boldsymbol{\xi}} \mid \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right) \propto |\boldsymbol{\Sigma}_\xi|^{-K/2} \exp\left(-\frac{1}{2}\sum_k \left(\tilde{\boldsymbol{\xi}}_k - \boldsymbol{\mu}_\xi\right)^t \boldsymbol{\Sigma}_\xi^{-1}\left(\tilde{\boldsymbol{\xi}}_k - \boldsymbol{\mu}_\xi\right)\right)$$

can be factorized as

$$|\boldsymbol{\Sigma}_\xi|^{-K/2} \exp\left(\frac{K}{2}\left(-\operatorname{tr}\boldsymbol{\Sigma}_\xi^{-1}\mathbf{S} - \left(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_\xi\right)^t \boldsymbol{\Sigma}_\xi^{-1}\left(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_\xi\right)\right)\right), \quad (4.46)$$

where $\mathbf{S} = \sum_k (\tilde{\boldsymbol{\xi}}_k - \overline{\boldsymbol{\xi}})(\tilde{\boldsymbol{\xi}}_k - \overline{\boldsymbol{\xi}})^t/K$ and $\overline{\boldsymbol{\xi}} = \sum_k \tilde{\boldsymbol{\xi}}_k/K$.
(b) Show that the hyperprior density of $(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ can be written as

$$p\left(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, K_0, \nu\right) \propto |\boldsymbol{\Sigma}_\xi|^{-\left(\frac{\nu+2}{2}+1\right)} \exp\left(\frac{1}{2}\left(-\operatorname{tr}\left(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_\xi^{-1}\right)\right.\right.$$
$$\left.\left. -K_0\left(\boldsymbol{\mu}_\xi - \boldsymbol{\mu}_0\right)^t \boldsymbol{\Sigma}_\xi^{-1}(\boldsymbol{\mu}_\xi - \boldsymbol{\mu}_0)\right)\right). \quad (4.47)$$

(c) Derive the parameters of the normal density $p(\boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi, \tilde{\boldsymbol{\xi}})$ by collecting terms from (4.46) and (4.47).
(d) Given that $p(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi \mid \tilde{\boldsymbol{\xi}}) = p(\boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi, \tilde{\boldsymbol{\xi}})p(\boldsymbol{\Sigma}_\xi \mid \tilde{\boldsymbol{\xi}})$, derive the parameters of the inverse Wishart density $p(\boldsymbol{\Sigma}_\xi \mid \tilde{\boldsymbol{\xi}})$.

4.4. The prior for the item parameters $\boldsymbol{\xi}_k$ is normal with mean $\boldsymbol{\mu}_\xi = (\mu_a, \mu_b)$ and variance $\boldsymbol{\Sigma}_\xi$, and assume a normal prior for $\boldsymbol{\mu}_\xi$ with mean $\boldsymbol{\mu}_0$ and variance $\boldsymbol{\Sigma}_0$.
(a) Assume that $\boldsymbol{\Sigma}_\xi$ is known. Show that the density $p\left(\boldsymbol{\mu}_\xi \mid \boldsymbol{\xi}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ is normal with mean

$$\boldsymbol{\mu}^* = \left(\boldsymbol{\Sigma}_0^{-1} + K\boldsymbol{\Sigma}_\xi^{-1}\right)^{-1}\left(K\overline{\boldsymbol{\xi}}\boldsymbol{\Sigma}_\xi^{-1} + \boldsymbol{\mu}_0\boldsymbol{\Sigma}_0^{-1}\right)$$

and variance

$$\boldsymbol{\Sigma}^* = \left(\boldsymbol{\Sigma}_0^{-1} + K\boldsymbol{\Sigma}_\xi^{-1}\right)^{-1},$$

where $\overline{\boldsymbol{\xi}} = K^{-1}\sum_k \boldsymbol{\xi}_k$.
(b) Assume that each a_k is restricted to the set $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$ and $a_k \mid \mu_a, \sigma_a^2 \sim \mathcal{N}(\mu_a, \sigma_a^2)I(a_k \in \mathcal{A}_k)$. Show that the conditional density of $\boldsymbol{\mu}_\xi \mid \boldsymbol{\xi}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ is proportional to

$$p\left(\boldsymbol{\mu}_\xi \mid \boldsymbol{\xi}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \propto \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}_\xi - \boldsymbol{\mu}^*\right)^t \boldsymbol{\Sigma}^*\left(\boldsymbol{\mu}_\xi - \boldsymbol{\mu}^*\right)\right)}{\prod_k \Phi\left(\mu_a/\sigma_a\right)} \quad (4.48)$$

when $a_k \in \mathcal{A}_k$ and zero when $a_k \in \mathcal{A}_k^c$.

(c) Extend MCMC scheme 2 with an M-H step to draw samples from the density in Equation (4.48), where the parameter of interest is also located in the normalizing constant.
(d) Assume the lognormal prior in Equation (4.3) for the item parameters, with a normal inverse Wishart prior for the hyperparameters. Adjust MCMC scheme 2 by defining sampling steps for the item parameters and the hyperparameters.
(e) Related to (d), Gibbs sampling steps can be constructed to sample the item parameters. Show how the difficulty parameters can be sampled from a normal density via

$$p\left(b_k \mid \tilde{\mathbf{z}}_k, \boldsymbol{\theta}, a_k, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right) \propto p\left(\tilde{\mathbf{z}}_k \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}\right) p\left(b_k \mid a_k, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right).$$

Show that $\tilde{a}_k = \log(a_k)$ can be sampled from

$$\tilde{a}_k \mid \tilde{\mathbf{z}}_k, b_k, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi \sim \mathcal{N}\left(\mu_{\tilde{a}|b}, \Sigma_{\tilde{a}|b}\right) I_{(\Delta_l, \Delta_u)}(\tilde{a}_k),$$

where $\mu_{\tilde{a}|b}$ and $\Sigma_{\tilde{a}|b}$ are the prior parameters from the prior of a_k given b_k, and where

$$\Delta_l = \max\left(\max_{i|Y_{ik}=0,\theta_i<0} f(\tilde{z}_{ik}, \theta_i, b_k), \max_{i|Y_{ik}=1,\theta_i>0} f(\tilde{z}_{ik}, \theta_i, b_k)\right),$$
$$\Delta_u = \min\left(\min_{i|Y_{ik}=0,\theta_i>0} f(\tilde{z}_{ik}, \theta_i, b_k), \min_{i|Y_{ik}=1,\theta_i<0} f(\tilde{z}_{ik}, \theta_i, b_k)\right),$$

with $f(\tilde{z}_{ik}, \theta_i, b_k) = \log\left(\frac{b_k - \tilde{z}_{ik}}{\theta_i}\right)$.

4.5. Consider the data augmentation scheme defined in Equation (4.7), a normal prior for the ability parameters (Equation (2.8)), and common normal priors for the item parameters (Equation (4.23) and (4.24)). Sampling steps are derived that will make up Albert's MCMC scheme (Albert, 1992).
(a) Assume normally distributed augmented data. Show that the full conditional of θ_i is normal with mean

$$\left(\mathbf{a}^t\mathbf{a} + \sigma_\theta^{-2}\right)^{-1}\left(\mathbf{a}^t\left(\mathbf{z}_i + \mathbf{b}\right) + \mu_\theta/\sigma_\theta^2\right) \tag{4.49}$$

and variance

$$\left(\mathbf{a}^t\mathbf{a} + \sigma_\theta^{-2}\right)^{-1}.$$

(b) The model is identified by fixing prior parameters $\mu_\theta = 0$ and $\sigma_\theta^2 = 1$. Show that the conditional expected value of θ_i equals

$$E\left(\theta_i \mid \mathbf{z}_i, \boldsymbol{\xi}\right) = \left(\mathbf{a}^t\mathbf{a}\right)^{-1}\mathbf{a}^t\left(\mathbf{z}_i + \mathbf{b}\right) - \frac{1}{\mathbf{a}^t\mathbf{a} + 1}\left(\mathbf{a}^t\mathbf{a}\right)^{-1}\mathbf{a}^t\left(\mathbf{z}_i + \mathbf{b}\right). \tag{4.50}$$

(c) Explain that the EAP of θ_i in (4.50) has the form of a shrinkage estimator.

(d) Define the sampling step for the item parameters by specifying the full conditional density.
(e) Define sampling steps for the parameters of the priors for the item parameters using normal priors for the means and inverse gamma priors for the variances.

4.6. The logistic response model is very close to the normal ogive response model. The parameters of both models are defined on the same scale when using an adjustment factor d such that

$$p\left(Y_i = 1 \mid \theta_i, a, b\right) = \Phi\left(a\theta_i - b\right) = \frac{\exp\left(d\left(a\theta_i - b\right)\right)}{1 + \exp\left(d\left(a\theta_i - b\right)\right)}.$$

If the adjustment factor is considered to be a parameter, a sampling scheme can be defined to estimate its value. Therefore, simulate data under the normal ogive model, given values for $(a, b, \boldsymbol{\theta})$, by defining uniformly distributed variables $U_i \in (0, 1)$ $(i = 1, \ldots, N)$ in such a way that $Y_i = 1$ when

$$U_i < \Phi\left(a\theta_i - b\right)$$

and zero otherwise.
(a) Show that the conditional density of U_i is given by

$$U_i \mid \mathbf{y}, d \sim \begin{cases} \mathcal{U}\left(0, \frac{\exp(d(a\theta_i - b))}{1+\exp(d(a\theta_i - b))}\right) \text{ if } Y_i = 1 \\ \mathcal{U}\left(\frac{\exp(d(a\theta_i - b))}{1+\exp(d(a\theta_i - b))}, 1\right) \text{ if } Y_i = 0. \end{cases} \tag{4.51}$$

(b) Show that the conditional density in Equation (4.51) induces the restrictions

$$d \geq \log\left(\frac{u_i}{1 - u_i}\right) \Big/ \eta_i \text{ if } i \in \mathcal{A}_1,$$
$$d < \log\left(\frac{u_i}{1 - u_i}\right) \Big/ \eta_i \text{ if } i \in \mathcal{A}_2,$$

where the sets $\mathcal{A}_1$ and $\mathcal{A}_2$ equal

$$\mathcal{A}_1 = \left\{i : (Y_i = 1) \cup (\eta_i > 0) \cap (Y_i = 0) \cup (\eta_i < 0)\right\},$$
$$\mathcal{A}_2 = \left\{i : (Y_i = 1) \cup (\eta_i < 0) \cap (Y_i = 0) \cup (\eta_i > 0)\right\},$$

where $\eta_i = a\theta_i - b$.
(c) Given a noninformative prior $p(d) \propto 1$, deduce the conditional posterior

$$d \mid \mathbf{u}, \mathbf{y} \sim \mathcal{U}\left(\max_{i \in \mathcal{A}_1} \log\left(\frac{u_i}{1 - u_i}\right) \Big/ \eta_i, \min_{i \in \mathcal{A}_2} \log\left(\frac{u_i}{1 - u_i}\right) \Big/ \eta_i\right).$$

(d) Define a Gibbs sampling scheme for estimating the adjustment factor d.
(e) Explain how sample size N influences the convergence properties of the algorithm. (MCMC scheme 2 can be extended such that the adjustment factor becomes a nuisance parameter instead of using a fixed value of 1.7; see Lord, 1980, pp. 12–14).

4.7. Define a Gibbs sampler for the partial credit model based on data augmentation. Attention is focused on a Gibbs sampling step for the threshold parameter $\kappa_{k,c}$.
(a) Derive the conditional distribution of $\boldsymbol{\kappa}_k$ in terms of cumulative probabilities, where

$$\pi_{ik}(c) = P\left(Y_{ik} \leq c\right) = \frac{\sum_{s=0}^{c} \exp \sum_{l=0}^{s} \left(\theta_i - \kappa_{k,l}\right)}{\sum_{r=0}^{C_K} \exp \sum_{l=0}^{r} \left(\theta_i - \kappa_{k,l}\right)}.$$

(b) Derive the conditional distribution of $\boldsymbol{\kappa}_k$ given realizations of a random variable $U_{ik} \sim \mathcal{U}_{[0,1]}$, where

$$\pi_{ik}(c) = P\left(U_{ik} \leq \frac{\sum_{s=0}^{c} \exp \sum_{l=0}^{s} \left(\theta_i - \kappa_{k,l}\right)}{\sum_{r=0}^{C_K} \exp \sum_{l=0}^{r} \left(\theta_i - \kappa_{k,l}\right)}\right).$$

(c) Show that the conditional distribution of U_{ik} is given by

$$U_{ik} \mid \boldsymbol{\theta}, \boldsymbol{\kappa}, Y_{ik} = c \sim \mathcal{U}\left(\pi_{ik}(c-1), \pi_{ik}(c)\right).$$

(d) Show that the augmented data can be used to define a restriction on the parameter space of $\kappa_{k,c}$; that is,

$$\kappa_{k,c} \leq \sum_{l=0}^{c-1} \left(\theta_i - \kappa_{k,l}\right) + \theta_i - \log \left[\frac{u_{ik}}{1-u_{ik}} \left(\sum_{r \neq c}^{C_k} \exp \sum_{l=0}^{r} \left(\theta_i - \kappa_{k,l}\right) - \frac{1}{u_{ik}} \sum_{s \neq c}^{c-1} \exp \sum_{l=0}^{s} \left(\theta_i - \kappa_{k,l}\right) \right) \right] \quad \text{if } Y_{ik} = c, \tag{4.52}$$

$$\kappa_{k,c} > \sum_{l=0}^{c-1} \left(\theta_i - \kappa_{k,l}\right) + \theta_i - \log \left[\frac{u_{ik}}{1-u_{ik}} \left(\sum_{r \neq c}^{C_k} \exp \sum_{l=0}^{r} \left(\theta_i - \kappa_{k,l}\right) - \frac{1}{u_{ik}} \sum_{s=0}^{c-1} \exp \sum_{l=0}^{s} \left(\theta_i - \kappa_{k,l}\right) \right) \right] \quad \text{if } Y_{ik} = c+1. \tag{4.53}$$

(e) Show that the full conditional of $\kappa_{k,c}$ equals

$$\kappa_{k,c} \mid \mathbf{u}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\kappa}_{k(-c)} \sim \mathcal{U}(h_1, h_2) p(\kappa_{k,c}),$$

where

$$h_1 = \max_{i \mid Y_{ik} = c+1} \Delta_l,$$
$$h_2 = \min_{i \mid Y_{ik} = c} \Delta_u,$$

where Δ_l and Δ_u equal the right-hand sides of (4.53) and (4.52), respectively.

(f) Consider the logistic graded response model and the prior in Equation (2.8). In the same way as above, define a uniformly distributed random variable and show that the M-H step 2 of MCMC scheme 3 can be replaced by the Gibbs sampling step

$$\theta_i \mid \mathbf{y}, \mathbf{a}^{(m)}, \boldsymbol{\kappa}^{(m)}, \mu_\theta^{(m)}, \sigma_\theta^{2(m)} \sim \mathcal{N}\left(\mu_\theta, \sigma_\theta^2\right) I_{\mathcal{R}_\theta}(\theta_i)$$

for $\mathcal{R}_\theta = \{\theta_i \in \mathcal{R}, \Delta_l < \theta_i < \Delta_u\}$, where

$$\Delta_l = \max_{c|k\in\mathcal{A}} \left(\log\left(\frac{1-u_k}{u_k}\right) + \kappa_{k,c} \right) / a_k,$$
$$\Delta_u = \min_{c|k\in\mathcal{A}} \left(\log\left(\frac{1-u_k}{u_k}\right) + \kappa_{k,c-1} \right) / a_k,$$

and $\mathcal{A} = \{k : Y_{ik} = c\}$.

4.8. The object is to sample threshold parameters. Define $\delta = 1/\kappa_{k,C_k-1}$ for $C_k - 1 > 0$.
(a) Show that the normally distributed augmented data defined in Equation (4.25) when parameterized as $Z_{ik}^* = Z_{ik}\delta$ are distributed as

$$\mathbf{Z}_k^* \mid \delta, a_k, \boldsymbol{\theta} \sim \mathcal{N}\left(a_k\boldsymbol{\theta}\delta^2, \delta^2\right).$$

(b) Show how to sample threshold parameter values $\kappa_{k,c}$ from their conditional distributions, where δ functions as a variance parameter.
(c) The reparameterization eliminates all unknown threshold parameters for item k when $C_k = 3$ and $\kappa_{k,1}$ is fixed to zero to identify the model. Show how to sample the threshold parameters $\kappa_{k'1}$ for $k' \neq k$ by using the cumulative probabilities $P\left(\mathbf{Z}_{k'}^* \leq \kappa_{k'1}\right)$ and $P\left(\mathbf{Z}_{k'}^* > \kappa_{k'1}\right)$.

4.9. A hierarchical normal prior for the item parameters $\boldsymbol{\xi}_k = (a_k, b_k, \tilde{c}_k)^t$ is defined as

$$(a_k, b_k, \tilde{c}_k)^t \sim \mathcal{N}\left(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right) I_{\mathcal{A}_k}(a_k),$$

where $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$ and $\tilde{c}_k = \Phi^{-1}(c_k)$. Assume a normal inverse Wishart distribution for the hyperparameters.
(a) Show that the transformation $\tilde{c}_k = \Phi^{-1}(c_k)$ is monotone, where c_k and $\tilde{c}_k$ have density functions $p(.)$ and $g(.)$ with support sets

$$\mathcal{R}_{c_k} = \{c_k; p(c_k) > 0\},$$
$$\mathcal{R}_{\tilde{c}_k} = \{\tilde{c}_k; \tilde{c}_k = \Phi^{-1}(c_k) \text{ for } c_k \in \mathcal{R}_{c_k}\},$$

respectively.
(b) Show that the conditional distribution of $\tilde{c}_k$ given $a_k, b_k, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi$ is continuous on the support set $\mathcal{R}_{\tilde{c}}$ and that $\Phi^{-1}(c_k)$ has a continuous derivative on $\mathcal{R}_{\tilde{c}}$.

(c) Derive the conditional prior of c_k given $a_k, b_k, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi$ via

$$p\left(c_k \mid a_k, b_k, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right) = g\left(\Phi^{-1}\left(c_k\right) \mid a_k, b_k, \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right) \frac{\mathrm{d}\Phi^{-1}\left(c_k\right)}{\mathrm{d}c_k}. \quad (4.54)$$

(d) Define the acceptance probability of a candidate value c_k^* from a proposal $\mathcal{B}e(\alpha, \beta)$ distribution given the prior for c_k, defined in Equation (4.54), and binomially distributed augmented data $\mathbf{S}_k$, as defined above Equation (4.22).

4.10. An observed item response is denoted as y_{obs} and a missing item response as y_{mis}. Assume that data are missing at random (MAR) such that the probability of a value being missing depends only on observed values and not missing values (Rubin, 1976).[3]
(a) The observed-data posterior distribution of θ can be related to the complete-data posterior distribution:

$$p\left(\theta \mid \mathbf{y}_{obs}\right) = \int p\left(\theta \mid \mathbf{y}_{mis}, \mathbf{y}_{obs}\right) p\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}\right) \mathrm{d}\mathbf{y}_{mis}.$$

Explain how the posterior mean can be estimated using draws of $\mathbf{Y}_{mis}$ from $p\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}\right)$.
(b) Use the item response model for analysis to draw missing values. Verify that the probability function of $\mathbf{Y}_{mis}$ given $\mathbf{y}_{obs}$ can be written as

$$p\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}\right) = \int p\left(\mathbf{y}_{mis} \mid \theta\right) p\left(\theta \mid \mathbf{y}_{obs}\right) \mathrm{d}\theta.$$

(c) Consider MCMC scheme 1, and define an additional sampling step for generating realizations of $\mathbf{Y}_{mis}$ given $\mathbf{y}_{obs}$.
(d) Consider MCMC scheme 2, where a data augmentation step is used. Verify that the probability function of Y_{mis} given $\mathbf{y}_{obs}$ can be written as

$$p\left(y_{mis} \mid \mathbf{y}_{obs}\right) = \int\int p\left(y_{mis} \mid z\right) p\left(z \mid \theta, \mathbf{y}_{obs}\right) p\left(\theta \mid \mathbf{y}_{obs}\right) \mathrm{d}\theta \mathrm{d}z,$$

where z is an augmented item response.
(e) The following data augmentation scheme for MCMC scheme 2 is defined:

$$Z_{ik} \mid \theta_i, a_k, b_k, Y_{ik} \sim \begin{cases} \mathcal{N}\left(a_k\theta_i - b_k, 1\right) I\left(Z_{ik} \leq 0\right) & \text{if } Y_{ik} = 0 \\ \mathcal{N}\left(a_k\theta_i - b_k, 1\right) I\left(Z_{ik} > 0\right) & \text{if } Y_{ik} = 1 \\ \mathcal{N}\left(a_k\theta_i - b_k, 1\right) & \text{if } Y_{ik} \text{ is missing.} \end{cases}$$

Show how to compute the expected value of a missing observation.
(f) Assume that the (conditional) expected level of ability is given by Equation (4.49), and that the data augmentation scheme in (e) is used. Explain the effect of imputing missing responses on the estimated posterior mean level of ability.

[3] The missing-data mechanism is ignorable by assuming that the parameters of the item response model for the observed data and the parameters of the missingness mechanism are distinct (Rubin, 1987).

5

Assessment of Bayesian Item Response Models

The underlying assumptions of Bayesian item response models have to be examined to ensure their credibility and that meaningful inferences can be made. A set of tools will be discussed for testing model assumptions and hypotheses. This set of tools includes methods based on Bayesian residuals and predictive diagnostic checks. It will be shown that related computations can be done during an MCMC estimation procedure or afterwards using MCMC output.

5.1 Bayesian Model Investigation

A very powerful approach for model assessment is based on predictive assessment. The idea is to generate predictive data from the model, and the replicated data are compared with the observed data. Replicated datasets are compared with observed data with respect to features of interest, which are captured in a summary statistic. Different summary statistics of the data can be considered to investigate whether the model is able to describe the characteristics of the observed data. Summary statistics or discrepancy measures that are functions of the data only are also called test statistics. Discrepancy measures are chosen to detect systematic differences between model and data. It is possible to test various specific assumptions of a model due to the enormous flexibility in choosing discrepancy functions.

Two different sampling procedures for simulating replicated data will be considered. Data can be simulated from a fitted model and will be referred to as posterior predictive data since values are drawn conditional on the observed data. Data can also be sampled from a model without conditioning on observed data and will be referred to as prior predictive data. Prior predictive model checks will be used to test whether the model is appropriate for describing the data without needing posterior simulations. At this point, different priors can be considered to construct a complete model that yields sensible implications for observables. An MCMC algorithm can be developed for a complete model,

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_5,

and posterior predictive data can be used to detect inconsistencies between model and observed data and to better understand the implications of the model for observed data.

Although several predictive checks are explored to test the fit of an item response model, in this chapter only a brief overview is given since there are numerous possibilities and to date no definitive predictive tests exist in this area. Model fitting is an active area of current research, new developments are evolving quickly, and it seems that this will be a topic of further research.

Besides predictive assessment, a residual analysis is a common statistical tool for model validation. Residuals are easily obtained as by-products of an MCMC algorithm and will be used to investigate the fit of the model. When different models are fitted to the same data, they can be compared using a summary measure of fit. The DIC discussed in Section 3.2.3 will be used to compare different models for observed item response data.

5.2 Bayesian Residual Analysis

Bayesian residuals, also referred to as residuals, are viewed as random parameters with unknown values. The residuals need to be estimated from the data together with their uncertainties. Summary statistics of the posterior distributions of the residuals need to be calculated, and posterior probability statements can be made about the values of the realized errors (Box and Tiao, 1973; Zellner, 1971). For example, assume the normal ogive item response model for analyzing binary data. In that case, a residual is defined as $R_{ik} = Y_{ik} - \Phi(a_k\theta_i - b_k)$. The marginal posterior density of the residual, $p(r_{ik} \mid y_{ik})$, is a continuous-valued posterior, and the posterior mean is used to estimate the residual value.

In a standard residual analysis, residuals are usually transformed such that they approximately follow a normal distribution. In the case of discrete observations, such transformations result in poor approximations by the normal distribution. In a Bayesian residual analysis, attention can be focused on the posterior distribution of each residual, which can be estimated via MCMC.

Let $(\theta_i^{(m)}, a_k^{(m)}, b_k^{(m)})$ denote an MCMC sample from their joint posterior distribution. It follows that sampled values from the residual's posterior distribution corresponding to observation ik are given by

$$R_{ik}^{(m)} = Y_{ik} - \Phi\left(a_k^{(m)}\theta_i^{(m)} - b_k^{(m)}\right). \tag{5.1}$$

Although the Bayesian residuals are easily estimated within an MCMC scheme, the posterior variances of the residuals differ and the residuals' posterior densities are not directly comparable.

5.2.1 Bayesian Latent Residuals

Albert and Chib (1995) and Johnson and Albert (1999) introduced Bayesian latent residuals as an alternative to the Bayesian residuals. In Section 4.3, various data augmentation schemes were introduced such that a regression of an augmented variable Z_{ik} on the latent ability θ_i is linear and with the same error variance for all θ_i. The difference between the latent response, Z_{ik}, and the expected response is defined as a Bayesian latent residual. Specifically, according to the augmentation scheme in Section 4.3.2, the Bayesian latent residual is defined as the difference between the augmented normally distributed Z_{ik} (Equation (4.7)) and the expected mean $a_k\theta_i - b_k$. Note that each latent residual is standard normally distributed due to the identifying assumptions associated with the data augmentation scheme (see Section 4.4.1).

5.2.2 Computation of Bayesian Latent Residuals

According to Equation (4.38), the Bayesian latent residual corresponding to binary observations Y_{ik} is defined as

$$\varepsilon_{ik} = Z_{ik} - a_k\theta_i + b_k. \tag{5.2}$$

From the definition of the augmented data, it follows that, given $\boldsymbol{\xi}_k$ and θ_i, the Bayesian latent residual ε_{ik} is standard normally or logistically distributed. For polytomous response data,

$$\varepsilon_{ik} = Z_{ik} - a_k\theta_i, \tag{5.3}$$

where Z_{ik} is defined according to Equation (4.25). Both Bayesian latent residuals (Equations (5.2) and (5.3)) can be estimated as an average of computed residual values in each MCMC iteration.

A more efficient estimator is based on the conditional expectation given a sufficient statistic, which is called a Rao-Blackwellized estimator (Gelfand and Smith, 1990). When it is possible to draw independent samples, a Rao-Blackwellized estimator that is based on averaging conditional expectations instead of the original parameter (the empirical estimator) of interest can produce a large variance reduction and be more efficient. The variance reduction is not guaranteed when the estimator is based on samples that are drawn dependently using a Gibbs sampler. However, Liu, Wong and Kong (1994) proved that the Rao-Blackwellized estimator is better than the empirical estimator for data augmentation schemes.

The conditional expectation of a Bayesian latent residual is derived by integrating out the augmented response data. For binary response data (see Exercise 5.1), if $Y_{ik} = 1$,

$$\begin{aligned} E\left(\varepsilon_{ik} \mid Y_{ik} = 1, \theta_{ij}, \boldsymbol{\xi}_k\right) &= \int_0^\infty \varepsilon_{ik} \frac{p\left(z_{ik}, Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k\right)}{P\left(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k\right)} \mathrm{d}z_{ik} \\ &= \frac{\phi\left(b_k - a_k\theta_i\right)}{\Phi\left(a_k\theta_i - b_k\right)}, \end{aligned} \tag{5.4}$$

where $\phi(.)$ is the standard normal density function. For $Y_{ik} = 0$,

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

For ordinal response data using Equations (4.27) and (5.3), the conditional expectation of a Bayesian latent residual given $Y_{ik} = c$ equals

$$E\left(\varepsilon_{ik} \mid Y_{ik} = c, \theta_i, a_k\right) = \frac{\phi\left(\kappa_{k,c-1} - a_k\theta_i\right) - \phi\left(\kappa_{k,c} - a_k\theta_i\right)}{\Phi\left(\kappa_{k,c} - a_k\theta_i\right) - \Phi\left(\kappa_{k,c-1} - a_k\theta_i\right)}. \tag{5.6}$$

Some elementary calculations have to be done to find expressions for the posterior variances of the residuals (Exercise 5.1). Note that in a comparable way Rao-Blackwellized estimates can be obtained for logistically distributed latent residuals.

5.2.3 Detection of Outliers

The posterior distribution of a Bayesian latent residual can be used to calculate the posterior probability that the corresponding observation is an outlier. An observation is considered to be outlying if the posterior distribution of the corresponding residual is located far from its mean. Following Albert and Chib (1995), Chaloner and Brant (1988), and Zellner (1971), y_{ik} is an outlier if the absolute value of the residual is greater than some prespecified value q times the standard deviation. That is, observation y_{ik} is marked as an outlier if $P\left(|\varepsilon_{ik}| > q \mid y_{ik}\right)$ is large. The probability that an observation exceeds a prespecified value is called the outlying probability.

A Rao-Blackwellized estimate of the conditional probability that the absolute value of latent residual ε_{ik} exceeds a value q given y_{ik} can be derived. That is, if $Y_{ik} = 1$,

$$P\left(|\varepsilon_{ik}| > q \mid Y_{ik} = 1, \theta_i, a_k, b_k\right) = \frac{\Phi\left(-q\right)}{\Phi\left(a_k\theta_{ij} - b_k\right)} \tag{5.7}$$

for $q > -\left(a_k\theta_i - b_k\right)$, and if $Y_{ik} = 0$,

$$P\left(|\varepsilon_{ik}| > q \mid Y_{ik} = 0, \theta_i, a_k, b_k\right) = \frac{\Phi\left(-q\right)}{1 - \Phi\left(a_k\theta_i - b_k\right)} \tag{5.8}$$

for $q > a_k\theta_i - b_k$. The expressions can be used to estimate the outlying probabilities of the estimated Bayesian latent residuals given sampled values of the model parameters (see Exercise 5.2 for details).

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

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

Notice that other complex posterior probabilities can be computed with an MCMC algorithm by keeping track of all the possible outcomes of the relevant probability statement. However, this method has the drawback that a lot of iterations are necessary to get a reliable estimate. It could be possible, for example, that in the case of multiple outliers a test for a single outlier does not detect one outlier in the presence of another outlier. This so-called masking occurs when two outlying probabilities related to observations ik and sk do not indicate any outliers but the joint posterior probability

$$P\left(|\varepsilon_{ik}| > q, |\varepsilon_{sk}| > q \mid \mathbf{y}\right) \tag{5.9}$$

shows that y_{ik} and y_{sk} are both outliers. This joint probability can be estimated by counting the events where both absolute values of the residuals are greater than q times the standard deviation divided by the total number of iterations.

5.2.4 Residual Analysis: Dutch Primary School Mathematics Test

Item responses from 2,156 grade eight students, unequally spread over 97 schools, to 18 dichotomously scored mathematics items taken from the examination upon leaving school developed by the National Institute for Educational Measurement are considered (Doolaard, 1999). For the moment, the nesting of students in schools is ignored, but it will be discussed in Section 6.6.1.

A two-parameter normal ogive model was used as the measurement model, with a standard normal prior for the ability parameters and with the hierarchical prior in Equation (4.3) for the item parameters. In Listing 5.1, Rao-Blackwellized estimates of the latent residuals are specified. The expressions for the latent residual estimates are not dependent on the augmented variable and can be implemented in other model formulations such as Listing 1.1 (Exercise 5.3).

Listing 5.1. WinBUGS code: Estimating Bayesian latent residuals.

```
for (i in 1:N){
    for (k in 1:K){
        eta[i,k] <- a[k]*theta[i] - b[k]
        resid n[i,k]    <- 0.3989*exp(-.5*(pow(-eta[i,k],2)))
        residual[i,k] <- (resid n[i,k]/phi(eta[i,k]))*Y[i,k] +
                              (-resid n[i,k]/phi(-eta[i,k]))*(1-Y[i,k])
    }
}
```

In Figure 5.1, the marginal posterior densities of Bayesian latent residuals and Bayesian residuals corresponding to the same 25 randomly selected answers to item 17 are plotted (lower and upper plots, respectively). The order of the posterior means is the same in both plots. It can be seen that the marginal posterior distributions of the Bayesian residuals are defined on $(-1, 0)$ if the observation corresponding to item 17 equals zero and on $(0, 1)$ otherwise. The posterior mean of a Bayesian latent residual is positive (negative) when the answer is correct (incorrect). It is more difficult to assess the extremeness of the marginal posterior densities of the Bayesian residuals since they are different and defined on different domains. Subsequently, it is difficult to identify outliers from these marginal posterior distributions.

The outlying probabilities were computed for $q = 0$ using Equations (5.7) and (5.8) (see also Exercise 5.2). In Figure 5.1, the four smallest posterior means of the Bayesian latent residuals are significantly smaller than zero when using a 10% significance level. For $q = 1$, the outlying probability of a Bayesian latent residual is .982, and the corresponding response pattern showed that all items were scored correct except item 17, although this item was answered correctly by 88% of the students.

5.3 HPD Region Testing and Bayesian Residuals

As explained in Section 3.2.2, hypotheses can be tested using the posterior density of the parameters of interest. The posterior density is used to test whether or not a specific point lies inside or outside an HPD region. According to the usual form of a hypothesis that a parameter value or a function of parameter values is zero, the HPD interval is used to test if the parameter value differs significantly from zero.

Item and Person Fit

By means of a person-fit statistic, the fit of a score pattern can be determined given that the item response model holds. Investigation of the person fit may provide information about the response behavior of a person. There have been many statistics proposed to evaluate the fit of persons' response patterns. Glas and Meijer (2003) and Meijer and Sijtsma (2001) have given an overview of person-fit statistics for various item response models.

Here, the idea is to evaluate whether a set of Bayesian latent residuals corresponding to a specific response pattern is extreme under the item response model. Assume the data augmentation scheme in Equation (4.25) for the graded response model. The (person-fit) statistic $Q_{p,i}$ in (5.10) is defined for a set of Bayesian latent residuals (as defined in Equation (5.3)) belonging to the response pattern of a person indexed i as

$$Q_{p,i}\left(\mathbf{Z}_i\right) = \sum_k \left(Z_{ik} - a_k\theta_i\right)^2 = \sum_k \epsilon_{ik}^2, \tag{5.10}$$

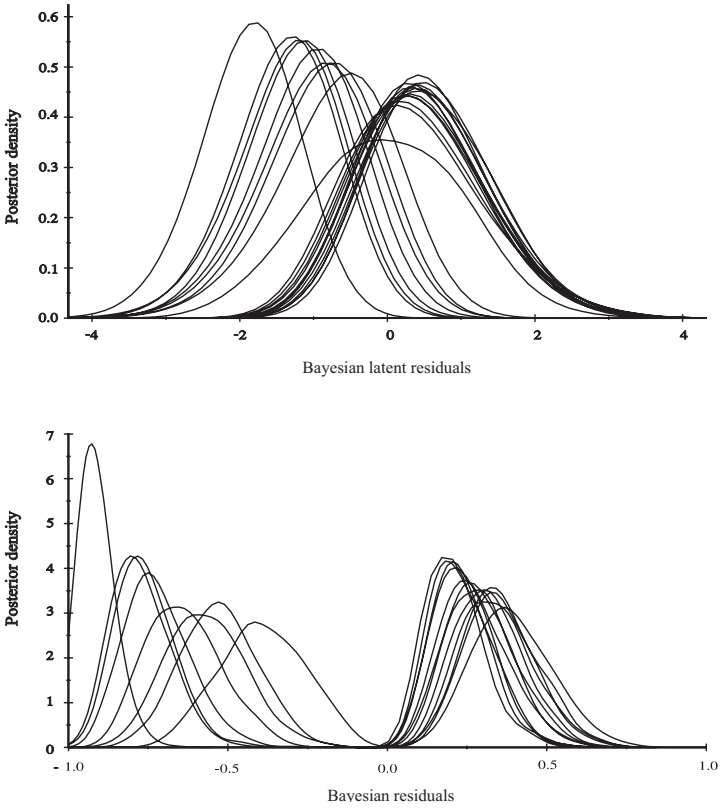

Fig. 5.1. Posterior densities of Bayesian latent residuals and Bayesian residuals corresponding to item 17.

where the dependence on item and person parameters is ignored. Each latent residual is standard normally distributed, and the sum of the K squared Bayesian latent residuals is chi-square distributed with K degrees of freedom.[1] This reference distribution is used to quantify the extremeness of the sum of squared latent residuals. As a result, a corresponding marginal posterior p-value (tail-area probability) is defined as

$$p_0(Q_{p,i}) = \int P\left(\chi^2_K > Q_p\left(\mathbf{z}_i\right)\right) p\left(\mathbf{z}_i \mid \mathbf{y}_i\right) \mathrm{d}\mathbf{z}_i, \tag{5.11}$$

and a small tail-area probability indicates that the estimated set of estimated latent residuals is extreme under the item response model for individual i. Each tail-area probability is easily computed within an MCMC scheme. In each iteration, after convergence, the conditional tail-area probability is com-

[1] Let $Z_1, \ldots, Z_K$ be independent normal random variables. The random variable $\chi^2_K = Z_1^2 + Z_2^2 + \cdots + Z_K^2$ has a chi-square distribution with K degrees of freedom.

puted and the mean of the conditional tail-area probabilities is an estimate of the corresponding marginal tail-area probability.

In the same way, an item fit statistic is defined to assess the fit of an item characteristic function under the graded response model. For each item, the corresponding Bayesian latent residuals are explored and interest is focused on assessing whether the sum of squared estimated Bayesian latent residuals is significantly large. This would indicate a poor fit of the item characteristic curve. The item fit statistic is defined as

$$Q_{item,k}\left(\mathbf{Z}_k\right) = \sum_i \left(Z_{ik} - a_k\theta_i\right)^2, \tag{5.12}$$

and the corresponding marginal tail-area probability equals

$$p_0(Q_{item,k}) = \int P\left(\chi^2_N > Q_{item,k}\left(\mathbf{z}_k\right)\right) p\left(\mathbf{z}_k \mid \mathbf{y}_k\right) \mathrm{d}\mathbf{z}_k, \tag{5.13}$$

where the dependence on person and item parameters is suppressed.

Detecting Discriminating Items

In an item analysis, interest is often focused on the discriminating power of the items. The more effectively an item differentiates between persons of higher and lower ability, the more useful that item is as an instrument for separating individuals. It is of interest to detect the most discriminating items but also to test whether items discriminate differently.

The one-sided hypothesis that the discriminating power of an item is above a mean level, say one, is tested by computing the posterior probability

$$P\left(a_k \geq 1 \mid \mathbf{y}\right) = \int_1^\infty p\left(a_k \mid \mathbf{y}\right) \mathrm{d}a_k. \tag{5.14}$$

The posterior probability can be computed via an MCMC sample of discriminating values from the marginal posterior distribution.

The marginal posterior probability in Equation (5.14) can also be computed via a data augmentation scheme. Assume a conditional normally distributed discrimination parameter according to Equation (4.35). The marginal posterior probability is expressed as

$$P\left(a_k \geq 1 \mid \mathbf{y}\right) = \int_1^\infty \int p\left(a_k \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \mathbf{y}\right) \mathrm{d}\mathbf{z}\mathrm{d}a_k,$$

which follows from properties of the augmented data scheme (see Exercise 5.5). The conditional distribution given augmented data is known, which makes it possible to derive a closed-form expression,

$$P\left(a_k \geq 1 \mid \mathbf{z}, \boldsymbol{\theta}, \mu_a, \sigma^2_a\right) = \frac{\Phi\left(\Omega_a^{-1/2}\left(\mu^*_a - 1\right)\right)}{\Phi\left(\Omega_a^{-1/2}\mu^*_a\right)}, \tag{5.15}$$

where Ω_a and μ_a^* are defined in Equations (4.36) and (4.37), respectively. As a result, the posterior probability of $a_k > 1$ can be accurately estimated using MCMC output also when the event is unlikely to occur.

The hypothesis that a discrimination parameter, or a set of discrimination parameters, equals a prespecified value is more difficult to test. The approach taken is to compute the posterior probability that a prespecified value a_k^0 is contained in the $1 - \alpha$ HPD region. According to Equation (3.16), the parameter value a_k^0 is included in the HPD interval if and only if

$$P\left(p\left(a_k \mid \mathbf{y}\right) \geq p\left(a_k^0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) \leq 1 - \alpha. \tag{5.16}$$

It is concluded that there is evidence that a_k differs significantly from a_k^0 when the probability that the point a_k^0 is included is greater than $1-\alpha$. In that case, the HPD interval needs to be stretched out to include a specific point that is unlikely to be covered. However, the corresponding marginal posterior distribution is unknown, which makes the procedure slightly more complicated. The idea is to condition on, among other things, augmented data, such that a closed-form expression of the conditional posterior probability that a prespecified value is covered by a $1 - \alpha$ HPD interval is obtained. The conditional posterior probability is easily computed in each MCMC iteration, and the average is an estimate of the corresponding marginal posterior probability.

Consider the conditional posterior distribution of a_k in Equation (4.35), which is normal with mean μ_{a^*} and variance Ω_a. The conditional posterior density $p(a_k \mid \mathbf{z}_k)$ is a monotonically decreasing function of

$$Q_a\left(a_k\right) = \Omega_a^{-1}\left(a_k - \mu_a^*\right)^2, \tag{5.17}$$

and the conditional joint posterior density $p(\mathbf{a} \mid \mathbf{z})$ is a monotonically decreasing function of

$$Q_a\left(a_1, \ldots, a_K\right) = \sum_k \Omega_a^{-1}\left(a_k - \mu_a^*\right)^2, \tag{5.18}$$

where, for notational convenience, the conditioning on the other parameters is suppressed. A large value of Q_a indicates that the point is not likely to be covered by the HPD interval. The Q_a is chi-square distributed when ignoring the positivity restriction on the discrimination parameters. In that case, it follows that a specific point is included when

$$\begin{aligned} P\left(Q_a\left(a_k^0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) &= P\left(Q_a\left(a_k \mid \mathbf{y}\right) \leq Q_a\left(a_k^0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) \\ &= \int P\left(Q_a\left(a_k \mid \mathbf{z}\right) \leq Q_a\left(a_k^0 \mid \mathbf{z}\right) \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \mathbf{y}\right) \mathrm{d}\mathbf{z} \\ &= \int P\left(\chi_1^2 \leq Q_a\left(a_k^0 \mid \mathbf{z}\right) \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \mathbf{y}\right) \mathrm{d}\mathbf{z} \\ &\leq 1 - \alpha, \end{aligned} \tag{5.19}$$

and a general point is included when

$$
\begin{aligned}
P\left(Q_a\left(\mathbf{a}^0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) &= P\left(Q_a\left(\mathbf{a} \mid \mathbf{y}\right) \leq Q_a\left(\mathbf{a}^0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) \\
&= \int P\left(Q_a\left(\mathbf{a} \mid \mathbf{z}\right) \leq Q_a\left(\mathbf{a}^0 \mid \mathbf{z}\right) \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \mathbf{y}\right) \mathrm{d}\mathbf{z} \\
&= \int P\left(\chi^2_K \leq Q_a\left(\mathbf{a}^0 \mid \mathbf{z}\right) \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \mathbf{y}\right) \mathrm{d}\mathbf{z} \\
&\leq 1 - \alpha. \qquad (5.20)
\end{aligned}
$$

In each MCMC iteration, a conditional p-value is calculated given augmented data using the property that Q_a is chi-square distributed. The average p-value is considered to be an estimate of the marginal p-value. Note that the p-values related to (5.19) and (5.20) refer to the probability of whether a point is included within an HPD interval. It is not based on the evaluation of a criterion using predictive or replicated data under the model. Model assessment using predictive data will be discussed in Section 5.4.

Common misinterpretations of the p-value are (1) that it would specify the posterior probability of the null hypothesis being true and (2) that a small p-value provides strong evidence against the null hypothesis. The p-value measures the surprise in the data and provides information to further develop or elaborate the model. At best, the p-value provides a rough indication that there is a certain mismatch between the model and the observed data. The hypothesis being tested is only a part of the actual test that is performed. This is one major point of criticism since with this kind of hypothesis testing it is not possible to address the question of interest directly. With respect to the second misinterpretation, Berger and Selke (1987) provided several examples showing that a small p-value does not necessarily mean that there is strong evidence against the null. Berger and Delampady (1987) and Berger and Selke (1987), among others, generally recommended the use of the Bayes factor when specific alternative hypotheses are available.

5.3.1 Measuring Alcohol Dependence: Graded Response Analysis

The College Alcohol Problem Scale (CAPS; O'Hare, 1997) was developed to serve as an initial screening instrument for students cited with a first offense for violating their university's rules concerning underage drinking. The items comprising the CAPS scale covered socioemotional problems (hangovers, memory loss, nervousness, depression) and community problems (drove under the influence, engaged in activities related to illegal drugs, problems with the law). Due to the high prevalence of alcohol abuse among college students, it is important that practitioners in student health services or counseling be able to identify students with drinking problems.

In 2002, 351 students from four colleges and universities in the state of North Carolina (Elon University, Guilford Technical Community College, University of North Carolina, Wake Forest University) were asked to respond to a questionnaire with 13 items from the CAPS instrument, with response

categories on a five-point scale (1=never/almost never to 5=almost always). The CAPS questionnaire is given in Section 5.9. It is assumed that a unidimensional latent variable representing alcohol dependence, denoted as θ, was measured by the items, where a higher level indicated that a participant was more likely to have a drinking problem.

The normal ogive graded response model was used to measure the individual alcohol-dependence levels. The fit of individual response patterns, items, and levels of item discrimination were examined.

For the CAPS data, the $Q_{p,i}$ statistic was evaluated for each person, and 3.4% of the response patterns can be characterized as improbable (aberrant) when using a significance level of 5%. This leads to the conclusion that the number of person misfits is relatively small. Two examples of the assessed aberrant response behavior are (1) a person answered "often" to items 5 ("Spent too much money on drugs") and 10 ("Drove under the influence") and "never" to all other items, and (2) a person answered "never" to items 5 ("Spent too much money on drugs") and 8 ("Caused others to criticize your behavior") and "often" to all other items. This suggests that these respondents were not willing to provide truthful answers and therefore gave inconsistent answers. The estimated observed score distribution is right-skewed, which indicates that a lot of respondents scored very low, and it might be possible that several respondents did not give honest answers due to the sensitive nature of the questions. More attention will be paid to this issue in Chapter 9.

From the 95% HPD intervals in Table 5.1 it follows that all discrimination parameters are located far away from zero. In the last column of Table 5.1, the tail-area probabilities, related to Equation (5.19), are given under the heading $p_0(Q_a)$, corresponding to the null hypothesis $a_k = 1$ $(k = 1, \ldots, K)$. It can be seen that items 8 and 11 are highly discriminating between persons with discrimination values significantly different from one (10% significance level). The tail-area probability of all discrimination parameters being equal to one (related to Equation (5.20)) is .030. It is concluded that there is no evidence that all CAPS items discriminate identically.

The $Q_{item,k}$ fit statistic was computed for each item. In the next to last column of Table 5.1, the tail-area probabilities are given for each item, and it can be concluded that the items fit the data. That is, the estimated sum of squared latent residuals cannot be considered extreme under the graded item response model.

5.4 Predictive Assessment

A standard statistical tool for model checking is based on a discrepancy measure (Gelman et al., 1996) or (departure) statistic (Bayarri and Berger, 2000) to investigate the compatibility of the model with the data. This discrepancy measure is chosen in such a way that large values indicate less compatibility.

Table 5.1. CAPS: Discrimination parameter estimates of the normal ogive graded response model.

Item	Mean	SD	HPD	$p_0(Q_{item})$	$p_0(Q_a)$
1	.645	.075	[.507, .796]	.224	.054
2	.727	.086	[.570, .902]	.511	.134
3	1.017	.101	[.824, 1.206]	.529	.598
4	.848	.113	[.629, 1.073]	.492	.390
5	.838	.124	[.578, 1.063]	.378	.407
6	.874	.099	[.682, 1.058]	.489	.448
7	1.212	.175	[.892, 1.555]	.612	.316
8	1.361	.145	[1.099, 1.647]	.374	.095
9	.916	.084	[.749, 1.074]	.342	.562
10	1.123	.113	[.886, 1.349]	.571	.436
11	1.395	.141	[1.144, 1.678]	.652	.055
12	1.257	.102	[1.064, 1.468]	.267	.164
13	1.229	.142	[.942, 1.503]	.535	.339

The reference distribution of the discrepancy measure is used to measure the extremeness of the observed discrepancy.

Consider an item response model $\mathcal{M}$ with parameters $(\boldsymbol{\theta}, \boldsymbol{\xi})$. Under the null hypothesis, it is assumed that the response data are conditionally distributed as $p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi})$ and the unknown parameters have a prior distribution $p(\boldsymbol{\theta}, \boldsymbol{\xi})$, without having a specific alternative hypothesis. Interest is focused on a statistic, denoted as $T(\mathbf{y})$, to investigate the compatibility of model $\mathcal{M}$ with the observed data $\mathbf{y}_{obs}$. Subsequently, a tail-area probability or p-value can be computed:

$$p_0(\mathbf{y}_{obs}) = P\left(T(\mathbf{Y}) \geq T(\mathbf{y}_{obs}) \mid \mathcal{M}\right). \tag{5.21}$$

The p-value can be computed in different ways when the parameters of the null model are unknown. In a prior predictive approach, the computation of p_0 is done with respect to the marginal distribution of $\mathbf{Y}$ (Box, 1980). In a posterior predictive approach, the computation is done with respect to the posterior predictive distribution of $\mathbf{Y}$ given $\mathbf{y}_{obs}$ (e.g., Rubin, 1984). Below, both approaches will be explored, and the advantages and disadvantages of both procedures will be discussed.

The predictive diagnostic checks utilized via p-values are specifically important when no fully specified alternative model is available. It is certainly not unusual that alternative models are not available and that checking the fit of the posited model needs to be done without the immediate availability of an alternative. This corresponds with the general beliefs about p-values (e.g., Bayarri and Berger, 2000; Gelman et al., 1996; Meng, 1994) that they are particularly interesting for investigating the compatibility of the model with

the data. That is, a set of T statistics can be used to test the incompatibility of the model with the data without needing specific alternatives.

5.4.1 Prior Predictive Assessment

Box (1980) recommended the use of the marginal predictive distribution for computing a p-value as defined in (5.21). Given that an assumed model $\mathcal{M}$ with parameters $(\boldsymbol{\theta}, \boldsymbol{\xi})$ is true, all possible data samples $\mathbf{y}$ that could occur are distributed as

$$p(\mathbf{y} \mid \mathcal{M}) = \int \int p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \mathcal{M}) p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathcal{M}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\xi}. \tag{5.22}$$

The left-hand side of (5.22) can be recognized as the marginal distribution of the data but can also be denoted as the prior predictive distribution. The prior predictive distribution has the same interpretation as the prior distribution. It does not depend on previous observations (prior distribution) and specifies the distribution of a quantity that is observable (predictive distribution). Equation (5.22) is the predictive distribution of $\mathbf{Y}$ that is needed to measure the surprise in the data. Box (1980) specified checking functions of the form $T(\mathbf{y})$ to investigate certain features in the data that are seldom extreme, if the model is true. The extremeness of each feature is measured by reference to $p(T(\mathbf{y}))$. This way, diagnostic checks of parametric as well as residual features of the model can be explored.

An overall natural predictive check for model $\mathcal{M}$ is defined by $T(\mathbf{y}) = p(\mathbf{y})^{-1}$ since high values of $T(\mathbf{y})$, and small values of $p(\mathbf{y})$, suggest that it is unlikely that the data are observed under the null model. The corresponding p-value equals

$$p_0(\mathbf{y}_{obs}) = P\left(p\left(\mathbf{y} \mid \mathcal{M}\right)^{-1} \geq p\left(\mathbf{y}_{obs} \mid \mathcal{M}\right)^{-1}\right), \tag{5.23}$$

where $\mathbf{y}_{obs}$ denotes the observed data. The prior predictive tail-area probability p_0 is an indication of the credibility of the model. A (very) low p_0 value indicates that it is unlikely that the observed data $\mathbf{y}_{obs}$ have occurred under model $\mathcal{M}$.

The prior predictive distribution of the discrepancy measure is difficult to obtain in closed form when it depends on unknown nuisance parameters. In that case, a forward simulation method can be used to estimate the prior predictive probability. The forward simulation method is easy to perform since it only requires sampling parameter values from the prior distributions and sampling observables from their conditional distribution given the sampled parameter values. Given population parameters $\boldsymbol{\theta}_P$ and $\boldsymbol{\xi}_P$, a forward simulator is defined by

$$\begin{aligned}
\boldsymbol{\xi}^{(m)} &\sim p\left(\boldsymbol{\xi} \mid \boldsymbol{\xi}_P\right), \\
\boldsymbol{\theta}^{(m)} &\sim p\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_P\right), \\
\mathbf{y}^{(m)} &\sim p\left(\mathbf{y} \mid \boldsymbol{\theta}^{(m)}, \boldsymbol{\xi}^{(m)}\right),
\end{aligned}$$

where the last step defines the sampling of response data from the prior predictive density (according to Equation (5.22)).

For example, let $T(\mathbf{y}_{obs})$ denote the sample skewness of the observed sum scores (see Exercise 1.3). A highly skewed distribution of observed sum scores may point out that a normal prior for the ability parameters is not appropriate and other population priors might be considered. It might also indicate that a test is too difficult (easy) for the sampled respondents, leading to excessively low (high) scores, and different priors for the item difficulty parameters might be considered. The prior distribution of the statistic $T(\mathbf{y})$ reveals whether the observed value $T(\mathbf{y}_{obs})$ is an extreme observation under the model. The sample skewness is computed for each draw of the prior predictive distribution. The fraction of draws leading to a value higher than $T(\mathbf{y}_{obs})$ is an estimate of the corresponding prior predictive p-value.

Considering data augmentation scheme (4.7), the prior predictive distribution of the observed data, Equation (5.22), can be written as,

$$p(\mathbf{y} \mid \mathcal{M}) = \int\int\int p(\mathbf{z} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \mathcal{M}) p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathcal{M}) \mathrm{d}\mathbf{z} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\xi}.$$

Then, a statistic can be defined based on the augmented data $T(\mathbf{z}_{obs})$. The following steps can be added to the forward simulator:

$$\begin{aligned} \mathbf{z}_{obs}^{(m)} &\sim p\left(\mathbf{z} \mid \mathbf{y}_{obs}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\xi}^{(m)}\right), \\ \mathbf{z}^{(m)} &\sim p\left(\mathbf{z} \mid \mathbf{y}^{(m)}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\xi}^{(m)}\right). \end{aligned}$$

Samples $\mathbf{z}_{obs}^{(m)}$, from the first step, are needed to estimate the value of the (observed) statistic, and in a similar way draws $\mathbf{z}^{(m)}$, from the second step, are needed to estimate the prior predictive p-value.

Again a discrepancy measure is defined to contrast the sample and prior information about the person parameter and to check their compatibility using terms of the posterior density. Assume the item parameters are known. Consider a normal prior for the person parameters, $p\left(\theta_i \mid \mu_\theta, \sigma_\theta^2\right)$. In Exercise 5.4, the terms in the exponent of the unnormalized posterior density of the person parameter are derived. Based on that expansion, the discrepancy measure

$$T\left(\mathbf{z}_i \mid \boldsymbol{\theta}_P\right) = (K-1)s_i^2 + \frac{\left(\hat{\theta}_i - \mu_\theta\right)^2}{\sum_k a_k^{-2} + \sigma_\theta^2} \tag{5.24}$$

is defined, where s_i^2 and $\hat{\theta}_i$ are defined in Equation (5.38) and (5.39), respectively. The corresponding prior predictive p-value for respondent i given augmented data is defined as

$$p_0(\mathbf{z}_{i,obs}) = P\left(\chi_K^2 \geq T\left(\mathbf{z}_{i,obs} \mid \boldsymbol{\theta}_P\right) \mid \mathbf{y}_{i,obs}\right). \tag{5.25}$$

The prior predictive p-value $p_0(\mathbf{y}_{i,obs})$ is computed by averaging over prior predictive p-values based on draws $\mathbf{z}_{i,obs}^{(m)}$.

The second term in (5.24) is focused on the discrepancy between the data-dependent least squares estimate and the prior estimate of θ_i. A predictive check can be defined to test the hypothesis that the sample and prior information are not in conflict with each other with respect to the mean. Note that the posterior distribution of θ_i combines the sample and prior information and it is used to construct a shrinkage estimate of θ_i (Exercise 5.4), which is a weighted combination of the (least squares) sample and prior estimate of θ_i. A large discrepancy between the sample and prior estimate of θ_i leads to a severely biased shrinkage estimate since it comprises two different thoughts about the true value of θ_i. Therefore, define the function

$$T\left(\hat{\theta}_i - \mu_\theta \mid \boldsymbol{\theta}_P\right) = \frac{\left(\hat{\theta}_i - \mu_\theta\right)^2}{\sum_k a_k^{-2} + \sigma_\theta^2}, \tag{5.26}$$

and the compatibility of the least squares estimate and the prior estimate is checked via

$$P\left(\chi_1^2 \geq T\left(\hat{\theta}_i - \mu_\theta \mid \boldsymbol{\theta}_P\right)\right). \tag{5.27}$$

Theil (1963) proposed a comparable test and denoted it as a comparability statistic. When the tail-area probability is small, it is concluded that the model is discredited by the data and that the posterior distribution of θ_i is not (closely) centered at the prior mean μ_θ. Note that the predictive check in Equation (5.27) can be extended to test simultaneously the compatibility of the sample and prior information with respect to the mean across all respondents. Similarly, the compatibility between prior and sample information with respect to the item parameters can be tested.

A general concern about the prior predictive test is its dependence on the prior distributions. An excellent model with very poor prior distributions may become suspicious during prior predictive checks. The computation of prior predictive probabilities requires large sets of sampled values when the prior distributions are proper but (very) vague. Also, the prior predictive distribution is improper when using improper prior distributions, and for that reason the prior predictive checks are restricted to proper prior distributions.

On the other hand, the dependency on the prior distributions makes the prior predictive checks particularly interesting for testing the compatibility between prior and sample information. A whole set of statistics may reveal the fit of the model without having to estimate any parameter, and they provide specific insights into the discrepancy between the model and the prior beliefs. When the prior predictive assessment reveals deficiencies, it depends on the investigator's belief in the prior distributions if the next step consists of altering the model or the prior specifications or both. Specific tests may provide useful clues for how to alter the model and/or priors.

5.4.2 Posterior Predictive Assessment

Geisser (1975), Guttman (1967), and Rubin (1984), among others, have focused on posterior predictive checks where the posterior predictive distribution of a statistic $T(\mathbf{Y})$ is used to test whether the observed statistic's value, $T(\mathbf{y}_{obs})$, appears to be typical. The computation of a p-value is done using the posterior predictive distribution of the predictive or replicated data $\mathbf{Y}_{rep}$. In correspondence with the prior predictive distribution of the data, Equation (5.22), the posterior predictive distribution is defined as

$$\begin{aligned} p\left(\mathbf{y}_{rep} \mid \mathbf{y}, \mathcal{M}\right) &= \int\int p\left(\mathbf{y}_{rep} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathcal{M}\right) p\left(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{y}, \mathcal{M}\right) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\xi} \\ &= \int\int p\left(\mathbf{y}_{rep} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \mathcal{M}\right) p\left(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{y}, \mathcal{M}\right) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\xi}, \end{aligned} \tag{5.28}$$

where the replicated or predictive data are independent of the observed data given the ability parameters due to the assumption of local independence. The distribution in (5.28) is a posterior predictive distribution since it concerns the distribution of future observables (predictive) and it conditions on observed values (posterior). In a posterior predictive assessment, the posterior predictive p-value (5.21) is computed using the posterior predictive density defined in (5.28).

Meng (1994) introduced an extension of the posterior predictive model check by allowing a statistic to depend on unknown model parameters. For example, the statistic in Equation (5.21) can be extended to depend on $\boldsymbol{\theta}$, which leads to a p-value

$$p_0(\mathbf{y}_{obs}) = P\left(T(\mathbf{y}_{rep}, \boldsymbol{\theta}) \geq T(\mathbf{y}_{obs}, \boldsymbol{\theta}) \mid \mathcal{M}\right), \tag{5.29}$$

where the probability is taken over the joint posterior distribution of $(\mathbf{y}_{rep}, \boldsymbol{\theta})$ under the null hypothesis.

There are two interpretations of the posterior predictive check with a parameter-dependent discrepancy measure. First, it allows one to measure directly the discrepancy between sample and unknown primary parameters $\boldsymbol{\theta}$, where the sampling distribution of the discrepancy measure may depend on unknown nuisance parameters. Second, the posterior predictive p-value in (5.29) can be computed for a reference distribution given primary parameters leading to $p_0(\mathbf{y}_{obs}, \boldsymbol{\theta})$. In the second stage, the expected value of the p-value is computed with respect to the marginal posterior distribution of $\boldsymbol{\theta}$, leading to the p-value defined in (5.29). As a result, the uncertainty in the primary parameters is taken into account in the computation of the p-value. This uncertainty is generated by the marginal distribution of the primary parameters.

Consider a data augmentation scheme based on the replicated response data. In that case, the posterior predictive density is defined as

$$p\left(\mathbf{y}_{rep} \mid \mathbf{y}\right) = \int\int\int p\left(\mathbf{z}_{rep} \mid \mathbf{y}_{rep}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) p\left(\mathbf{y}_{rep} \mid \boldsymbol{\theta}, \boldsymbol{\xi}\right) p\left(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{y}\right) \mathrm{d}\mathbf{z}_{rep}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\xi}.$$

A forward simulation method can be defined to obtain replicated data samples from the posterior predictive distribution. The forward simulation method for generating posterior predictive data requires sampling parameter values from the posterior distributions and sampling observables from their conditional distribution given the sampled parameter values. Given population parameters $\boldsymbol{\theta}_P$ and $\boldsymbol{\xi}_P$, a forward simulator is defined by

$$\begin{aligned}
\boldsymbol{\xi}^{(m)} &\sim p\left(\boldsymbol{\xi} \mid \mathbf{y}_{obs}, \boldsymbol{\xi}_P\right), \\
\boldsymbol{\theta}^{(m)} &\sim p\left(\boldsymbol{\theta} \mid \mathbf{y}_{obs}, \boldsymbol{\theta}_P\right), \\
\mathbf{y}_{rep}^{(m)} &\sim p\left(\mathbf{y}_{rep} \mid \boldsymbol{\theta}^{(m)}, \boldsymbol{\xi}^{(m)}\right), \\
\mathbf{z}_{rep}^{(m)} &\sim p\left(\mathbf{z}_{rep} \mid \mathbf{y}_{rep}^{(m)}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\xi}^{(m)}\right).
\end{aligned}$$

In the first two steps, parameter values are sampled from their marginal posterior distributions, which are the common output of an MCMC algorithm. The posterior predictive tests are easily evaluated within an MCMC algorithm. The sampling distribution of a discrepancy measure is usually unknown, and the MCMC algorithm is routinely used to derive a posterior predictive p-value; that is, the posterior predictive distribution is used to compute the probability that the discrepancy value obtained is extreme with respect to the observed value.

A common problem with the computation of posterior predictive p-values is its double use of the observed data, first to compute the posterior distribution of the model parameters for determining the posterior predictive distribution and second to compute the posterior predictive p-value. This double use of the data leads to unnatural behavior of the p-value. Bayarri and Berger (1999, 2000) showed in several examples that the distribution of the p-value is not uniformly distributed on $[0, 1]$ under the null. This complicates the interpretation and comparison of posterior predictive p-values. The literature shows two different ways of dealing with this problem.

One school of thought treats a posterior predictive p-value as statistical evidence for model misfit instead of using it for testing a specific null hypothesis (Gelman et al., 1996; Meng, 1994; Stern, 2000). It accepts the conservative behavior (more closely concentrated around .5) of the posterior predictive p-value (e.g., Bayarri and Berger, 2000; Sinharay and Stern, 2002). In this light, the posterior predictive model checks are viewed as diagnostic measures to evaluate model fit, in which the p-value is considered to be a useful summary. Attention is also focused on displaying diagnostic differences graphically. Stern (2000) noted that the posterior predictive p-value remains the conditional posterior probability of the event $T(\mathbf{Y}) \geq T(\mathbf{y}_{obs})$ and therefore provides useful model checking information with or without an (asymptotic) null distribution.

The other school is focused on establishing p-values on a calibrated scale (Aitkin, 1997; Bayarri and Berger, 2000). Bayarri and Berger (2000) proposed the conditional and partial predictive p-values, which are defined in such a way that the data are not used twice. However, posterior predictive model checking

is easy to perform via simulation-based techniques, but these proposed p-values are often difficult to compute. The recommended conditional posterior p-value is based upon the knowledge of a sufficient statistic such that the predictive distribution of a statistic is free from unknown model parameters, which may be hard to obtain.

Overview of Posterior Predictive Model Checks

Several posterior predictive checks have been developed for testing the fit of item response models. Statistics were developed to test the assumption of local independence (e.g., Hoijtink, 2001; Levy, 2006; Sinharay, 2005), person fit (Glas and Meijer, 2003), item fit (e.g., Sinharay, 2006), and differential item functioning (Hoijtink, 2001), among other things. The advantages of the proposed posterior predictive checks mentioned are (1) that the theoretical sampling distribution of the statistic does not need to be derived and (2) the discrepancy measures may depend on unknown parameters, and the associated uncertainty is explicitly taken into account in the computation of the p-value. Note that, in the frequentist framework, fit statistics that account for uncertainty in estimated item and ability parameters are complex, and several studies have shown that the theoretical sampling distribution of a statistic is affected due to plugged-in estimates depending on the degree of uncertainty (Molenaar and Hoijtink, 1990; Snijders, 2001; Stone and Hansen, 2000).

The assumption of local independence is often tested by investigating the assumption of weak local independence, which is a necessary but not a sufficient condition for (strong) local independence. The assumption of weak local independence can be stated as $\text{Cov}(Y_{ik}, Y_{ik'} \mid \theta_i) = 0$, which implies that the item responses to items k and k' $(k \neq k')$ are independent conditional on θ_i. The covariance between responses of item pairs can be used as a discrepancy measure to test local independence,

$$\begin{aligned} \text{Cov}(\mathbf{Y}_k, \mathbf{Y}_{k'}) &= \sum_i \left(Y_{ik} - \overline{Y}_{ik}\right)\left(Y_{ik'} - \overline{Y}_{ik'}\right)/N \\ &= \left(n_{11}n_{00} - n_{10}n_{01}\right)/N^2, \end{aligned} \tag{5.30}$$

where, for example, n_{11} is the number of subjects responding correctly to both items. Hoijtink (2001) proposed a discrepancy measure based on squared conditional covariances given the respondent's rest score $R_{kk'}$,

$$\sum_{R_{kk'}} \sqrt{n_{R_{kk'}}} \left(\text{Cov}\left(Y_{ik}, Y_{ik'} \mid R_{kk'}\right)\right)^2,$$

where $n_{R_{kk'}}$ equals the number of respondents with rest score $R_{kk'}$. Note that the sum of covariances is approximately zero if local independence holds. Edwards (1963) showed that the cross-product ratio or odds ratio can be taken as a measure of association when interest is focused on paired attributes with no fixed marginal totals. That is,

$$OR_{kk'} = \frac{n_{11}/n_{10}}{n_{01}/n_{00}} = \frac{n_{11}n_{00}}{n_{10}n_{01}}. \tag{5.31}$$

The odds ratio is invariant under the interchange of rows and columns and under row and column multiplications. Other measures of association and other properties of the odds ratio can be found in Bishop, Fienberg and Holland (1975). Sinharay, Johnson and Stern (2006) found the odds ratio to be a useful discrepancy measure for testing the assumption of local independence. Levy (2006) used the odds ratio to detect violations of local independence but with a specific interest in detecting a violation of the unidimensionality assumption due to the multidimensional nature of the observed data. Violations of local independence may be caused by assessment phenomena such as differential item functioning, testlet effects, and rater effects, and they can be framed in terms of multidimensionality (e.g., Mellenbergh, 1994a; Stout, Habing, Douglas, Kim, Roussos and Zhang, 1996; Yen, 1993).

Person-fit research is focused on determining the fit of individual response patterns. Response patterns can be regarded as unexpected, for example due to cheating, guessing, or plodding. Several person-fit statistics have been proposed for parametric and nonparametric item response models (e.g., Emons, Sijtsma and Meijer, 2005; Meijer and Sijtsma, 2001; Meijer, 2003). Glas and Meijer (2003) used most common person-fit tests as discrepancy measures and investigated the properties of the associated posterior predictive person-fit tests for binary response data. The various tests were used to detect the rate of guessing, item disclosure, and violations of local independence. With a simulation study, they showed that the detection rates for guessing and item disclosure were higher than for violations against local independence and concluded, that even for small sample sizes, accurate Type I error values were obtained. Across conditions, a discrepancy test from Tatsuoka (1984) had the highest power. This discrepancy measure takes on a large positive value for response patterns with correct answers to difficult items and incorrect answers to easy items. The nonstandardized version of the indices is given by

$$\sum_k \left(P_{ik}(\theta_i) - Y_{ik}\right)\left(P_{ik}(\theta_i) - Y_i/K\right), \tag{5.32}$$

where $P_{ik}(\theta_i)$ is the probability of a correct response to item k of a person indexed i and Y_i the number of correct responses. The theoretical distribution of Tatsuoka's standardized statistic is unknown, and often a normal distribution is assumed. As said, within the posterior predictive framework, the theoretical sampling distribution of the statistic does not need to be known explicitly.

Sinharay (2006) showed that the item-fit indices of Orlando and Thissen (2000, 2003) are promising discrepancy measures for unidimensional item response models for binary data. The fit indices are based on comparisons between the observed and expected proportions of correct scores for different examinee groups. Within a simulation study, it was shown that they have acceptable Type I error rates and considerable power for moderate and large sample sizes.

Two additional posterior predictive checks are based on weighted and unweighted mean squares. Béguin and Glas (2001) proposed a discrepancy measure for overall model fit to compare the observed score distribution with the posterior predictive score distribution. The discrepancy measure compares the expected, $E(n_r)$, and observed numbers of correct scores, n_r; that is,

$$X^2 = \sum_r \left(\frac{n_r - E(n_r)}{E(n_r)} \right)^2, \tag{5.33}$$

where $r = 0, 1, \ldots, K$. It is known that X^2 is not chi-square distributed, and the posterior predictive p-value is computed using samples from the posterior predictive distribution. Wright (1977) and Masters and Wright (1997) proposed so-called outfit and infit statistics without knowing the exact distributional properties. The outfit statistic is the unweighted mean squares of standardized residuals, where each residual is defined as the difference between the observed and the expected response to an item. The infit statistic is the weighted mean squares, where the weights are defined by the residual variances. This statistic can be computed for each person and each item. Both statistics summarize the variation between the observed response patterns and the expected response patterns under the model. The infit statistic is less sensitive to outliers since the influence of outliers is reduced by the weighting factor. The distributional properties of both statistics do need to be known explicitly when using them as posterior predictive checks (see Exercise 5.6).

5.5 Illustrations of Predictive Assessment

5.5.1 The Observed Score Distribution

Consider the CAPS response data in Section 5.3.1. The fit of the graded response model is investigated by comparing the observed score distribution with the posterior predictive score distribution. The discrepancy measure in Equation (5.33) is used to evaluate the extremeness of the observed score distribution where r ranges from $r = 13$ to $r = 65$ (13 items with response categories ranging from 1 to 5). The posterior predictive p-value equals

$$p_0\left(\mathbf{y}_{obs}\right) = P\left(X^2\left(\mathbf{y}_{rep}\right) \geq X^2\left(\mathbf{y}_{obs}\right)\right).$$

The corresponding posterior predictive p-value equals .039, which indicates that the observed score distribution is more extreme than expected under the model. This follows from the fact that the observed scores indicate a highly skewed (to the right) observed score distribution. A lot of respondents score relatively low, with 25% scoring between 13 and 14.5 (the first quartile) and 25% scoring between 23 and 52 (third quartile), with the median equal to 19. The discrepancy measure can also be computed for each single score level, which leads to a posterior predictive p-value for each of them. The

computations can be done in the same way. As expected, the p-values indicate that the observed score levels at the lower and higher ends are more extreme than expected under the model.

5.5.2 Detecting Testlet Effects

A Short Introduction to Testlet Effects

Rosenbaum (1988) argued that the assumption of local independence might be violated when there is a dependence within bundles of items that explicitly share material. This can be, for example, a reading passage. Respondents that have difficulties with certain words or sentences in the passage have greater difficulty with the entire bundle of items associated with the passage. It is also possible that items exhibit dependence without explicitly sharing material. Rosenbaum (1988) proved that the assumption of unidimensionality may hold between item bundles given the loss of local independence within the item bundles. As a result, the assumption of unidimensionality between item bundles holds when the assumption of conditional independence between item bundles holds. A unidimensional item response model is still appropriate when using the item bundle as the unit of measurement.

Bradlow et al. (1999) introduced a testlet parameter to model the additional dependence between items within the same item bundle, also called a testlet (for a more general overview of testlet models, see Wainer, Bradlow and Wang, 2007). Let l_k indicate the testlet of item k and l indicate a testlet, and assume that there are L testlets in total ($l, l_k \in 1, \ldots, L$). Then, let parameter $\upsilon_{i,l}$ denote a person-specific testlet effect that is independent of person parameters and item parameters. The testlet parameter, $\upsilon_{i,l}$, is considered to be a normally distributed random effects parameter with mean zero and variance $\sigma^2_{\upsilon_l}$. This specification allows for a testlet-specific variance parameter, and the testlet parameters are assumed to be independent from each other.

The random effects (testlet) parameter models the additional dependence between the individual responses to items within a testlet. To see this, assume binary response data, Y_{ik}, that are augmented by normally distributed latent continuous response data, Z_{ik}, with mean $a_k\theta_i - b_k - \upsilon_{i,l_k}$ and variance one, and Y_{ik} is the indicator that Z_{ik} is positive. The probability that person i responds correctly to item k is defined by

$$\begin{aligned} P\left(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k, \upsilon_{i,l_k}\right) &= P\left(Z_{ik} > 0 \mid \theta_i, \boldsymbol{\xi}_k, \upsilon_{i,l_k}\right) \\ &= \int_0^\infty \phi\left(x; a_k\theta_i - b_k - \upsilon_{i,l_k}\right) \mathrm{d}x \\ &= \Phi\left(a_k\theta_i - b_k - \upsilon_{i,l_k}\right). \end{aligned} \tag{5.34}$$

Thus the normal ogive model (see Equation (4.5)) is adjusted with a random effects parameter to account for a testlet effect. The sign of this testlet parameter leads to a higher (negative sign) or a lower (positive sign) success

probability. A respondent that scores well (poorly) on items in a particular testlet has higher (lower) success probabilities than the respondent's ability level would suggest, and this additional effect is implemented by a negative (positive) testlet parameter value. Now, for fixed item parameter values and $k \neq k'$, it follows that

$$\begin{aligned}\mathrm{Cov}\left(Z_{ik}, Z_{ik'} \mid \boldsymbol{\xi}_k, \boldsymbol{\xi}_{k'}\right) &= \mathrm{Cov}\left(a_k\theta_i, a_{k'}\theta_i\right) + \mathrm{Cov}\left(\upsilon_{i,l_k}, \upsilon_{i,l_{k'}}\right) \\ &= \begin{cases} a_k a_{k'}\sigma_\theta^2 + \sigma_{\upsilon_{l_k}}^2 & \text{if } l_k = l_{k'} \\ a_k a_{k'}\sigma_\theta^2 & \text{if } l_k \neq l_{k'} \end{cases} \end{aligned} \tag{5.35}$$

due to the various assumptions of independence. It can be concluded that when the items k and k' belong to the same testlet the additional dependence between the corresponding observations is captured by the testlet parameters. An MCMC algorithm for estimating the two-parameter item response model with testlet parameters can be constructed from MCMC scheme 2. This requires adding the steps for sampling testlet and testlet variance parameters from their conditional posterior distributions and adjusting a few original steps (see Exercise 5.7).

Simulation Study

Data are generated according to a normal ogive testlet model, Equation (5.34), with $N = 1{,}000$ and $K = 30$. The items are grouped by six testlets, each with five items. A common variance is assumed for the testlet parameters, $\sigma_{\upsilon_l}^2 = .20$. An exchangeable hierarchical prior density is specified for the item parameters,

$$\begin{aligned}\left(\log a_k, b_k\right)^t &\sim \mathcal{N}\left(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right), \\ \boldsymbol{\Sigma}_\xi &\sim \mathcal{IW}(2, \boldsymbol{\Sigma}_0), \\ \boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi &\sim \mathcal{N}(0, \boldsymbol{\Sigma}_\xi/2),\end{aligned}$$

and an exchangeable hierarchical prior for the testlet parameters

$$\begin{aligned}\upsilon_{i,l} &\sim \mathcal{N}(0, \sigma_{\upsilon_l}^2), \\ \sigma_{\upsilon_l}^2 &\sim \mathcal{IG}\left(3, 1\right).\end{aligned}$$

The parameters of a two-parameter normal ogive model are estimated, given the data generated via the normal ogive testlet model, using MCMC scheme 2. Violations of local independence are to be expected due to the additional dependence between individual responses to items in the same testlet. The odds ratio (Equation (5.31)) is used as a predictive check to test for violations of local independence. Replicated data are sampled from the prior predictive distribution (Equation (5.22)) and the posterior predictive distribution (Equation (5.28)) under the two-parameter normal ogive model. The odds

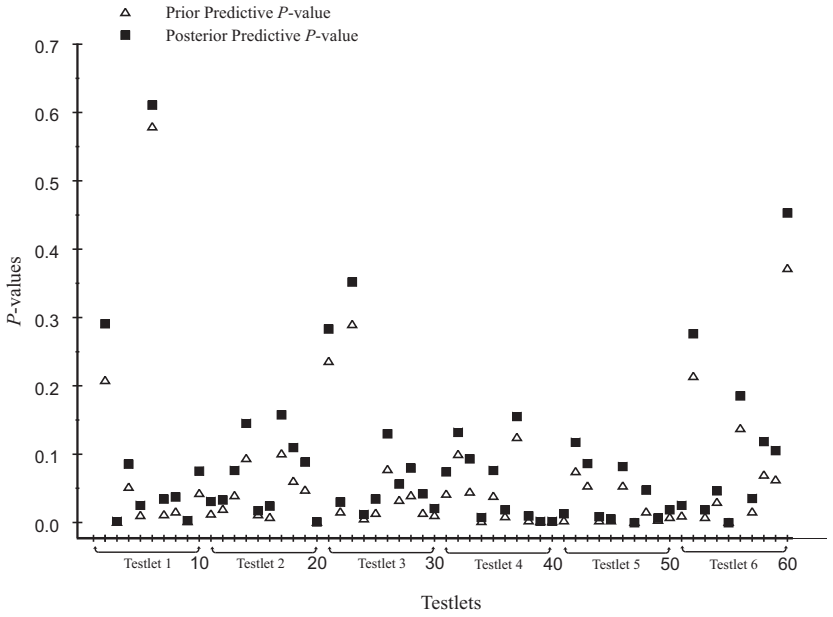

Fig. 5.2. Detecting violations of local independence: Prior and posterior predictive p-values related to item pairs where both items are nested in the same testlet.

ratio is evaluated as a prior predictive check by computing the prior predictive p-value for each item pair using replicated data from the prior predictive distribution of the data. A posterior predictive p-value is also computed for each item pair using replicated data from the posterior predictive distribution given the observed data.

In Figure 5.2, the estimated prior and posterior predictive p-values are plotted for item pairs where violations of local independence are to be expected. More specifically, the first five items are nested in the first testlet. Then, a total of 10 item-pairs are of interest: item 1 with 2, 3, 4, and 5; item 2 with 3, 4, and 5; item 3 with 4 and 5; and item 4 with 5. The estimated p-values are plotted in the same order. Each testlet consists of five items, and p-values are computed for the 10 item pairs.

It can be seen that each plotted prior predictive p-value is smaller than the corresponding posterior predictive p-value. This follows from the fact that, in general, the posterior predictive p-value is conservative and tends to 0.5 (for more details, see Robins, van der Vaart and Ventura, 2000). Subsequently, they are not exactly uniformly distributed, which makes it more difficult to say whether they are extreme or not. The prior predictive p-values are uniformly distributed and show more power in detecting violations of local independence. With a significance level of .05, 85% of the prior predictive p-values show a significant violation of local independence. For the posterior predictive p-values, 72% of the p-values detect a significant violation of local independence.

The p-values corresponding to item-pairs of items nested in different testlets should not show a significant violation of local independence. For this case, 99.7% of the posterior predictive p-values and 99.5% of the prior predictive p-values did not indicate a violation of local independence when violations were not to be expected. It can be concluded that the false-alarm rates are approximately the same but the detection rates differ (when violations of local independence are expected).

5.6 Model Comparison and Information Criteria

To compare different item response models fitted to the same data, the DIC can be used as a summary measure of fit. A general discussion of the DIC is given in Section 3.2.3.

The DIC of the two-parameter normal ogive model is described in more detail. Consider the deviance function $D(\boldsymbol{\theta}, \boldsymbol{\xi})$, which is based on the log-likelihood of the data given item and person parameters. This deviance function can be stated as

$$\begin{aligned} D(\boldsymbol{\theta}, \boldsymbol{\xi}) &= -2 \log p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi}) \\ &= -2 \log \prod_{i,k} \Big[\Phi(a_k \theta_i - b_k)^{y_{ik}} \left(1 - \Phi(a_k \theta_i - b_k)\right)^{1 - y_{ik}} \Big] \\ &= -2 \sum_{i,k} \Big[y_{ik} \log \Phi(\eta_{ik}) + (1 - y_{ik}) \log\left(1 - \Phi(\eta_{ik})\right) \Big], \end{aligned}$$

where $\eta_{ik} = a_k \theta_i - b_k$.

Assume MCMC samples $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$ from the marginal posterior densities. Then, the posterior mean of the deviance is estimated by

$$\overline{D(\boldsymbol{\theta}, \boldsymbol{\xi})} \approx \sum_m D\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\xi}^{(m)}\right) \Big/ M.$$

The deviance of the posterior means is estimated by substituting estimated posterior means for the item and person parameters. The penalty term p_D, the effective number of parameters, is easily evaluated as the posterior mean deviance minus the deviance of the posterior means.

The log-likelihood is known explicitly, which makes computing the marginal likelihood of the data easier when using MCMC samples (see Section 3.2.1). Subsequently, a Bayes factor can be used for model comparison. Spiegelhalter et al. (2002) argued that the Bayes factor and the DIC have different purposes. The Bayes factor summarizes how well priors have predicted the observed data, whereas the DIC summarizes how well the posterior density predicts future data generated by the same process as that which generated the obtained data. The DIC has a posterior predictive approach towards model selection, whereas the Bayes factor has a prior predictive approach. Nonetheless, in the next example, both methods are used to compare different response models fitted to the same data.

5.6.1 Dutch Math Data: Model Comparison

The use of the DIC is illustrated by comparing the fits of different response models. Therefore, the Dutch primary school math test data are considered (see Section 5.2.4). The fits of four item response models are investigated, where $\mathcal{M}_1$ and $\mathcal{M}_2$ denote the one-parameter and two-parameter normal ogive models, respectively. Let $\mathcal{M}_3$ and $\mathcal{M}_4$ denote the one-parameter and two-parameter logistic models, respectively. For all models, a standard normal prior is specified for the ability parameters. For the two-parameter models, a hierarchical prior is assumed for the item parameters. For the one-parameter models, a normal inverse gamma prior for the difficulty parameters is assumed.

Table 5.2 shows the posterior mean deviance, the deviance of the posterior means, the effective number of parameters, the DIC, and the marginal log-likelihood for each model. It follows from the estimated DICs that the two-parameter normal ogive model fits the data best. Furthermore, the two-parameter models are preferred over the one-parameter models, and the normal ogive models over the logistic models.

When looking at the estimated posterior mean deviances, it is remarkable that the one-parameter normal ogive model has the smallest posterior mean deviance. The corresponding DIC is relatively high due to the fact that the estimated effective number of parameters is considerably greater than that of the two-parameter models. In this case, the improvement of fit of the two-parameter models is enhanced by the reduction in the effective number of parameters and the DIC reduced. For the one-parameter models, the restriction on the item discrimination parameters leads to less pooled ability parameter estimates and a high number of effective parameters. That is, the ability parameter estimates are less pooled towards the prior mean in comparison with the estimates under the two-parameter model. The more complex modeling structure of the two-parameter model leads to a decrease in the number of effective parameters, ability estimates are shrunk towards the prior mean, and variability in the item discriminations is supported.

For each model, the estimated marginal log-likelihood is given in the last column of Table 5.2. A Bayes factor is used for model comparison. It follows that the two-parameter normal ogive model $\mathcal{M}_2$ is preferred over the one-parameter model $\mathcal{M}_1$ since $\hat{BF} = \exp((-18{,}529) - (-18{,}532)) = \exp(3)$. In the same way, it can be verified that the Bayes factor supports the conclusions that were made via the DIC.

5.7 Summary and Conclusions

The topic of model fitting is an active area of research with many diverse new methods. It is not possible to summarize all these methods in one chapter. In this chapter, attention was focused on Bayesian residual analysis, predictive assessment, and model comparison. Specifically, posterior predictive assessment constitutes a testing framework with many possibilities, but more

Table 5.2. Dutch math test: A model comparison.

Model		$\overline{D(\boldsymbol{\theta}, \boldsymbol{\xi})}$	$D(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\xi}})$	p_D	DIC	$\log p(\mathbf{y} \mid \mathcal{M})$
Normal Ogive	$\mathcal{M}_1$	37,062	35,168	1,894	38,956	−18,532
	$\mathcal{M}_2$	37,058	35,340	1,718	38,776	−18,529
Logistic	$\mathcal{M}_3$	37,104	35,220	1,884	38,989	−18,553
	$\mathcal{M}_4$	37,081	35,366	1,715	38,797	−18,539

research is needed to fully explore the possibilities of parameter-dependent discrepancy functions with known or unknown sampling distributions for assessing the fit of item response models.

Model criticism and selection is often focused on assessing the adequacy of a model in predicting the outcome of individual data points and summarizing the fit of the model as a whole. Carlin and Louis (1996), Gelfand, Dey and Chang (1992), and others advocated the use of cross-validation where the fitted value of an observation (or set of observations) is evaluated given all data except the observed value (set of observations). The posterior mean and variance are computed in relation to the conditional predictive distribution, also called the conditional predictive ordinate, which is the likelihood of each point given the rest of the data. Data points with low conditional predictive ordinates are not fitted well by the model. The conditional predictive distribution is often unknown and must be evaluated analytically, which can be time-consuming. An illustration of evaluating the fit of an item response model via cross-validation can be found in Fox (2005b).

The general fit of a model can also be examined via a model comparison approach. The adequacy of a baseline item response model can be evaluated by comparing it with competing models. However, as argued by Embretson and Reise (2000, p. 246), in straightforward item response analyses, the researcher's choices are already restricted to the standard models. They also remarked that, given sufficient response (unidimensional) data, the differences between competing models are probably very small. In subsequent chapters, more complex item response models will be discussed and more attention will be paid to the possibilities of evaluating a baseline model via Bayesian model choice methods as summarized in Section 3.2.3. In contrast to this, Bayesian item response modeling involves more issues than just choosing a likelihood model as in the frequentist framework. Prior and hyperprior distributions need to be specified, together with a likelihood model for the sample data, and they constitute the model. This complicates the Bayesian model assessment. Prior distributions may influence the results, and it is important to detect conflicts between prior and sample information. Also, the fit of a model can be improved by adjusting the prior distributions, a topic that will also be discussed in later chapters.

Finally, there is no standard procedure for evaluating the fit, and the variety of tests do not lead to a definite conclusion that a model does or does not fit the data. The outcomes of different test procedures need to be combined to come to a judgment about the fit of the model, and it is up to the researcher to make a proper judgment.

5.8 Exercises

5.1. Consider data augmentation scheme 4.7, and the object is to construct a Rao-Blackwellized estimator for the Bayesian latent residual in Equation (5.2).
(a) Verify that

$$P\left(Y_{ik}=1 \mid \theta_i, \boldsymbol{\xi}_k\right)=P\left(Z_{ik}>0 \mid \theta_i, \boldsymbol{\xi}_k\right)=\Phi\left(a_k \theta_i-b_k\right).$$

(b) For $Y_{ik}=1$, verify that the conditional expected value of ε_{ik} given θ_i and $\boldsymbol{\xi}_k$ equals

$$\begin{aligned} E\left(\varepsilon_{ik} \mid Y_{ik}=1, \theta_i, \boldsymbol{\xi}_k\right) &= \int_0^{\infty} \varepsilon_{ik} p\left(z_{ik} \mid \theta_i, \boldsymbol{\xi}_k, Y_{ik}=1\right) \mathrm{d} z_{ik} \\ &= \int_0^{\infty} \varepsilon_{ik} \frac{p\left(z_{ik} \mid \theta_i, \boldsymbol{\xi}_k\right)}{P\left(Y_{ik}=1 \mid \theta_i, \boldsymbol{\xi}_k\right)} \mathrm{d} z_{ik}, \end{aligned}$$

and derive the result in Equation (5.4).
(c) For $Y_{ik}=0$, show that the conditional expected value of ε_{ik} given θ_i and $\boldsymbol{\xi}_k$ is given by Equation (5.5).
(d) Derive a closed expression for the conditional variance of ε_{ik}; that is,

$$\operatorname{Var}\left(\varepsilon_{ik} \mid y_{ik}, \theta_i, \boldsymbol{\xi}_k\right)=\int_0^{\infty}\left(\varepsilon_{ik}-\varepsilon(y_{ik})\right)^2 p\left(z_{ik} \mid \theta_i, \boldsymbol{\xi}_k, y_{ik}\right) \mathrm{d} z_{ik},$$

where $\varepsilon(y_{ik})=E\left(\varepsilon_{ik} \mid \theta_i, \boldsymbol{\xi}_k, y_{ik}\right)$. For $Y_{ik}=1$, show that the conditional variance equals

$$1-\frac{\phi\left(-\eta_{ik}\right)}{\Phi\left(\eta_{ik}\right)}\left[\eta_{ik}+\frac{\phi(-\eta_{ik})}{\Phi(\eta_{ik})}\right],$$

and, for $Y_{ik}=0$, show that the conditional variance equals

$$1-\frac{\phi\left(-\eta_{ik}\right)}{1-\Phi\left(\eta_{ik}\right)}\left[\frac{\phi(-\eta_{ik})}{1-\Phi(\eta_{ik})}-\eta_{ik}\right],$$

where $\eta_{ik}=a_k \theta_i-b_k$.
(f) Construct a Rao-Blackwellized estimate of the variance of the point estimate defined in Equations (5.4) and (5.5).

5.2. The object is to find Rao-Blackwellized estimates of outlying probabilities, as defined in Equations (5.7) and (5.8), by integrating out the augmented response data.
(a) Let Z_{ik} be normally distributed according to Equation (4.7). Derive the conditional outlying probability of observation ik; that is, show that

$$P\left(|\varepsilon_{ik}| > q \mid Y_{ik}, \theta_i, a_k, b_k\right) = \frac{P\left(q < \varepsilon_{ik} \leq -q \mid \theta_i, a_k, b_k\right)}{P\left(Y_{ik} \mid \theta_i, a_k, b_k\right)} = \begin{cases} \frac{(1-\Phi(q))}{\Phi(a_k\theta_i - b_k)} & \text{if } Y_{ik} = 1 \\ \frac{(1-\Phi(q))}{\Phi(b_k - a_k\theta_i)} & \text{if } Y_{ik} = 0, \end{cases} \tag{5.36}$$

with the restriction that $q > b_k - a_k\theta_i$ when $Y_{ik} = 1$ and $q > a_k\theta_i - b_k$ when $Y_{ik} = 0$.
(b) Let Z_{ik} be logistically distributed according to Equation (4.7). Show that the outlying probability of observation ik equals

$$P\left(|\varepsilon_{ik}| > q \mid Y_{ik}, \theta_i, a_k, b_k\right) = \begin{cases} \frac{(1-\Psi(q))}{\Psi(a_k\theta_i - b_k)} & \text{if } Y_{ik} = 1 \\ \frac{(1-\Psi(q))}{\Psi(b_k - a_k\theta_i)} & \text{if } Y_{ik} = 0, \end{cases} \tag{5.37}$$

with the restriction that $q > b_k - a_k\theta_i$ when $Y_{ik} = 1$ and $q > a_k\theta_i - b_k$ when $Y_{ik} = 0$.

5.3. Consider the examinees' test result data in Section 1.4. An implementation of the normal ogive model via a latent variable specification is shown in Listing 5.2 based on the data augmentation step in Equation (4.7).

Listing 5.2. WinBUGS code: Latent variable specification of the normal ogive item response model.

```
model{
    for(i in 1:N){
      for(k in 1:K){
        eta[i,k] <- a[k]*theta[i]-b[k]
        Z[i,k] ~ dnorm(eta[i,k],1)
        P[i,k] <- step(Z[i,k])
        Y[i,k] ~ dbern(P[i,k])
       }
      theta[i] ~ dnorm(0,1)
   }
}
```

(a) Estimate the Bayesian residuals and plot them against the fitted probabilities. Indicate observations that can be marked as outliers.
(b) Estimate the Bayesian latent residuals simply as by-products of the MCMC algorithm and plot them against the fitted probabilities. Explain the differences with the plot of (a).
(c) Compute Rao-Blackwellized estimates of the Bayesian latent residuals, Equations (5.4) and (5.5). Plot the estimated Bayesian latent residuals against the estimates of (b) and explain the differences.

(d) Compute Rao-Blackwellized estimates of the outlying probabilities and compare the potential outliers detected with the results from (a).

5.4. The purpose is to derive the prior predictive checks in Equations (5.25) and (5.27). Assume normally distributed augmented response data, $\mathbf{Z}_{ik}$, with mean $a_k\theta_i - b_k$ and variance one. The augmented response data given observed binary responses are conditionally distributed according to Equation (4.7).
(a) Consider the augmented data likelihood $p\left(\mathbf{z}_i \mid \boldsymbol{\theta}, \boldsymbol{\xi}_k\right)$, expand the exponent with $a_k(\hat{\theta}_i - \hat{\theta}_i)$, and show that the likelihood is proportional to

$$p\left(\mathbf{z}_i \mid \boldsymbol{\theta}, \boldsymbol{\xi}_k\right) \propto \exp\left[-\frac{1}{2}\left((K-1)s_i^2 + \sum_k a_k^2\left(\hat{\theta}_i - \theta_i\right)^2\right)\right],$$

where

$$\hat{\theta}_i = \sum_k a_k^{-2} \sum_k a_k z_{ik}, \tag{5.38}$$

$$s_i^2 = (K-1)^{-1} \sum_k \left(z_{ik} - \left(a_k\hat{\theta}_i - b_k\right)\right)^2. \tag{5.39}$$

(b) Prove the following identity by completing the squares

$$\varphi_1(\theta-\theta_1)^2 + \varphi_2(\theta-\theta_2)^2 = (\varphi_1+\varphi_2)(\theta-\theta^*)^2 + \frac{(\theta_1-\theta_2)^2}{\varphi_1^{-1}+\varphi_2^{-1}},$$

where

$$\theta^* = \frac{(\varphi_1\theta_1 + \varphi_2\theta_2)}{\varphi_1+\varphi_2}.$$

(c) Assume a normal prior for θ_i (Equation (2.8)). Show that the unnormalized posterior $p(\mathbf{z}_i, \theta_i \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}_P) = p(\mathbf{z}_i \mid \theta_i, \boldsymbol{\xi}_k)p(\theta_i \mid \boldsymbol{\theta}_P)$ is proportional to

$$\exp\left[-\frac{1}{2}\left((K-1)s_i^2 + \left(\sum_k a_k^2 + \sigma_\theta^{-2}\right)\left(\theta_i - \theta_i^*\right)^2 + \frac{\left(\hat{\theta}_i - \mu_\theta\right)^2}{\sum_k a_k^{-2} + \sigma_\theta^2}\right)\right],$$

where

$$\theta_i^* = \frac{\sum_k a_k^2\hat{\theta}_i + \mu_\theta/\sigma_\theta^2}{\sum_k a_k^2 + \sigma_\theta^{-2}}.$$

(d) Derive the conditional posterior distribution of θ_i using the result of (c).
(e) Derive the conditional predictive distribution of $\mathbf{Z}_i$ given $(\boldsymbol{\xi}_k, \boldsymbol{\theta}_P)$ via

$$p\left(\mathbf{z}_i \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}_P\right) = \frac{p\left(\mathbf{z}_i, \theta_i \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}_P\right)}{p\left(\theta_i \mid \mathbf{z}_i, \boldsymbol{\xi}_k, \boldsymbol{\theta}_P\right)}.$$

(f) Derive the distribution of

$$T\left(\hat{\theta}_i - \mu_\theta \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}_P\right) = \frac{\left(\hat{\theta}_i - \mu_\theta\right)^2}{\sum_k a_k^{-2} + \sigma_\theta^2}.$$

5.5. Let observed ordinal data be described by a graded response model with prior distributions as defined in MCMC scheme 3. Assume normally distributed augmented data according to Equation (4.32).
(a) Prove that

$$\begin{aligned} P\left(a_k \geq 1 \mid \mathbf{y}\right) &= \int_1^\infty \int p\left(a_k \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \mathbf{y}\right) \mathrm{d}\mathbf{z}\mathrm{d}a_k \\ &= \int_1^\infty p\left(a_k \mid \mathbf{y}\right) \mathrm{d}a_k. \end{aligned}$$

(b) Derive the closed-form expression given by Equation (5.15), and show how MCMC output can be used to estimate the corresponding marginal probability.
(c) Consider the quantity $Q_a(a_k)$ in Equation (5.17) with conditional posterior density $p\left(Q_a \mid \mathbf{z}\right)$. Show that the posterior probability of $Q_a\left(a_k \mid \mathbf{y}\right) \geq Q_a\left(a_k^0 \mid \mathbf{y}\right)$ can be expressed as

$$1 - P\left(Q_a\left(a_k^0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) = \int \int_{Q_a\left(a_k^0\right)}^\infty p(Q_a \mid \mathbf{z}) p(\mathbf{z} \mid \mathbf{y}) \mathrm{d}Q_a \mathrm{d}\mathbf{z},$$

suppressing the conditioning on the other parameters.
(d) Show how MCMC output can be used to estimate the posterior probability with and without the fact that the $Q_a(a_k)$ is chi-square distributed given augmented data $\mathbf{z}$.

5.6. Binary response data are modeled with the normal ogive item response model. Assume the item parameters are known. The outfit person statistic of Wright (1977) is given by

$$T(\mathbf{Y}_i, \theta_i) = K^{-1} \sum_k \left(\frac{Y_{ik} - P_{ik}(\theta_i)}{\sqrt{(P_{ik}(\theta_i)(1 - P_{ik}(\theta_i)))}} \right)^2. \tag{5.40}$$

(a) Consider the outfit statistic $T(\mathbf{Y}_i, \theta_i)$ as a discrepancy measure, and define the corresponding posterior predictive p-value.
(b) Show that the outfit person statistic based on Bayesian latent residuals given normally distributed augmented data $\mathbf{Z}_i$, Equation (4.7), is given by

$$T(\mathbf{Z}_i, \theta_i) = K^{-1} \sum_k \left(Z_{ik} - (a_k \theta_i - b_k)\right)^2.$$

(c) Substantiate that the conditional posterior predictive p-value given $\mathbf{Z}_i$ and θ_i equals

$$p_0\left(\mathbf{Z}_i, \theta_i\right) = P\left(\chi^2_K \geq K\, T\left(\mathbf{Z}_i, \theta_i\right) \mid \mathbf{Z}_i, \theta_i\right), \tag{5.41}$$

and show how to compute the posterior predictive p-value using MCMC output.
(d) In the same way as in (b) and (c), derive the outfit item statistic based on Bayesian latent residuals.
(e) Argue that the posterior predictive tests via Bayesian residuals and Bayesian latent residuals are different tests and will lead to different outcomes.

5.7. Assume that response data are distributed according to the normal ogive testlet model, Equation (5.34). The object is to develop an MCMC scheme for estimating the model parameters by extending MCMC scheme 2 with steps for sampling testlet parameters $\upsilon_{i,l}$ and $\sigma^2_{\upsilon_l}$.
(a) Show that the full conditional distribution of $\upsilon_{i,l}$ is normal with mean and variance

$$\begin{aligned} E\left(\upsilon_{i,l} \mid \mathbf{z}_i, \boldsymbol{\xi}, \theta_i, \sigma^2_{\upsilon_l}\right) &= \frac{-\sum_{k:l_k=l}\left(z_{ik} - \left(a_k\theta_i - b_k\right)\right)}{n_l + \sigma^{-2}_{\upsilon_l}}, \\ \operatorname{Var}\left(\upsilon_{i,l} \mid \mathbf{z}_i, \boldsymbol{\xi}, \theta_i, \sigma^2_{\upsilon_l}\right) &= \left(n_l + \sigma^{-2}_{\upsilon_l}\right)^{-1}, \end{aligned}$$

where $n_l = \sum_k I\left(l_k = l\right)$ is the number of items in testlet l.
(b) Given an inverse gamma prior $\mathcal{IG}(g_1, g_2)$ for $\sigma^2_{\upsilon_l}$, show that the full conditional distribution of $\sigma^2_{\upsilon_l}$ is an inverse gamma with shape parameter $g_1 + n/2$ and scale parameter $g_2 + \sum_i \upsilon^2_{i,l}/2$.
(c) Show how to adjust MCMC scheme 2 to obtain a scheme for estimating the normal ogive testlet model.

5.8. A Bayesian residual can be defined for polytomous item response data as the difference between the observed and the expected responses. Following Masters and Wright (1997), define the expected response and variance as

$$\begin{aligned} \mu_{r_{ik}} &= E\left(Y_{ik} \mid \theta_i, \boldsymbol{\xi}_k\right) = \sum_{c=1}^{C_k} cP\left(Y_{ik} = c \mid \theta_i, \boldsymbol{\xi}_k\right), \\ \sigma^2_{r_{ik}} &= \operatorname{Var}\left(Y_{ik} \mid \theta_i, \boldsymbol{\xi}_k\right) = \sum_{c=1}^{C_k} \left(c - \mu_{r_{ik}}\right)^2 P\left(Y_{ik} = c \mid \theta_i, \boldsymbol{\xi}_k\right), \end{aligned}$$

and, subsequently, the standardized Bayesian residual as

$$R_{ik} = \left(Y_{ik} - \mu_{r_{ik}}\right)/\sigma_{r_{ik}}.$$

(a) Reason that the sum of squared standardized Bayesian residuals $Q_k(\mathbf{y}) = \sum_k R^2_{ik}$ and $Q_i(\mathbf{y}) = \sum_i R^2_{ik}$ can be used as a measure to detect item and person misfits, respectively.
(b) Define posterior predictive checks using the parameter-dependent discrepancy functions in (a).

(c) Compare the posterior predictive tests constructed with the posterior predictive tests in Equations (5.11) and (5.13) and explain the differences.
(d) Let $\mathbf{r}_k(\boldsymbol{\theta}, \mathbf{y})$ denote the vector of residuals for item k given $\boldsymbol{\theta}$; that is,

$$\mathbf{r}_k(\boldsymbol{\theta}, \mathbf{y}) = \int \mathbf{r}_k(\boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\xi} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\xi}.$$

Let $\sigma_{r_{k,k'}}$ denote the covariance between two vectors of residuals corresponding to items k and k'. Construct a test for local independence based on the covariance term $\sigma_{r_{k,k'}}$, and show how to use an MCMC algorithm to perform the computations.

5.9. The examinees' test result data in Section 1.4 are examined. Posterior predictive tests are performed in WinBUGS to detect person misfit.
(a) Extend Listing 1.1 with code from Listing 5.3 to simulate posterior predictive response data. Argue that posterior predictive data are sampled.

Listing 5.3. WinBUGS code: Simulating posterior predictive data.

```
theta.ppf <- cut(theta)
a.ppf <- cut(a)
b.ppf <- cut(b)

for(i in 1:N){
        for(k in 1:K){
            P[i,k] <- phi(a.ppf[k]*theta.ppf[i]-b.ppf[k])
            rep[i,k] ~ dbern(P[i,k])
        }
}
```

(b) Implement and evaluate Wright's outfit person statistic (Equation (5.40)).
(c) Explain detected misfits by examining the corresponding response patterns.
(d) Implement and evaluate Tatsuoka's nonstandardized person statistic (Equation (5.32)).
(e) Evaluate both statistics under the two-parameter logistic model, and compare the results with (c) and (d).

5.9 Appendix: CAPS Questionnaire

How often (almost always (5), often (4), sometimes (3), seldom (2), almost never (1)) have you had any of the following problems over the past year as a result of drinking too much alcohol?

1. Feeling sad, blue, or depressed.
2. Nervousness or irritability.
3. Hurt another person emotionally.
4. Family problems related to your drinking.
5. Spent too much money on drugs.
6. Badly affected friendship or relationship.
7. Hurt another person physically.
8. Caused others to criticize your behavior.
9. Nausea or vomiting.
10. Drove under the influence.
11. Spent too much money on alcohol.
12. Feeling tired or hung over.
13. Illegal activities associated with drug use.

6

Multilevel Item Response Theory Models

The item response data structure is hierarchical since item responses are nested within respondents. Often respondents are also grouped into larger units and variables are available that characterize the respondents and the higher-level units. An item response modeling framework is discussed that includes a multilevel population model for the respondents and takes such a hierarchical data structure into account. An important application area is in education, where response observations are grouped in students and students grouped in schools. Several school effectiveness research studies are discussed. The hierarchical item response model is extended in several directions to handle latent explanatory variables, model latent individual growth, and identify clusters of respondents.

6.1 Introduction: School Effectiveness Research

School effectiveness research is concerned with exploring differences within and between schools. The objective is to investigate the relationship between explanatory and outcome factors. This involves choosing an outcome variable, such as student's ability, and studying differences among schools after adjusting for relevant background variables. Interest is focused on the relative size of school differences and the factors that explain these differences and influence student learning.

Typically, in school effectiveness research, students are nested in classrooms, classrooms in schools, schools within school systems, and so on. A generally acceptable statistical model in the assessment of school effectiveness therefore requires the deployment of multilevel analysis techniques. A multilevel model takes the hierarchical structure into account, and variance components are modeled at each sampling level. As a result, homogeneity of results of individual pupils in the same school is accounted for since pupils in the same school share common experiences. Specifically, a multilevel model

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_6,

describes relationships between one or more "outcome" variables (examination results, attitudes), school and teacher characteristics (teacher's attitude, financial resources, class size), and student characteristics (achievements, social background). In the study of school effectiveness research, multilevel modeling has become so conspicuous since it allows for the analysis of individual and group-level effects and cross-level interactions.

The appropriateness of multilevel models in the assessment of school effectiveness was shown by Aitkin and Longford (1986). Since that time, most of the research has focused on multilevel modeling of hierarchically structured educational data and the assessment of relevant input and output indicators (e.g., Goldstein, 2003; Longford, 1993; Raudenbush and Bryk, 2002).

6.2 Nonlinear Mixed Effects Models

Two approaches for analyzing variables from different levels at one single level have been criticized. The first disaggregates all higher-order variables to the individual level. That is, data from higher levels are assigned to a much larger number of units at level 1. All disaggregated values are assumed to be independent of each other, which is a misspecification that threatens the validity of the inferences. In the second approach, observations at level 1 are aggregated to the higher level. As a result, all within-group information is lost. Relations between aggregated variables can be much stronger and different from the relations between nonaggregated variables. Snijders and Bosker (1999) give a complete overview of potential (statistical) errors when the clustered structure of the data is ignored.

Statistical models under a broad variety of names have been developed that can handle the different levels of the data. The models capture the between- and within-subject variances by modeling the data in two stages. At the first stage, a regression function is specified for the observations for each subject. Each subject has its own regression function. The same covariates are used across subjects, but the regression coefficients are allowed to vary. At the second stage, the regression coefficients that are allowed to vary are considered to be random outcome variables. These random outcome variables are referred to as random (regression) effects or random coefficients. This explains the term random coefficient models that is often used in sociological research and econometrics (De Leeuw and Kreft, 1986; Longford, 1993). However, there are a variety of names in the literature to describe versions of the same model. In educational and sociological research, the name multilevel model is often used (Goldstein, 2003; Snijders and Bosker, 1999). In biometric research, mixed effects model or random effects model are common terms (Laird and Ware, 1982; Hedeker and Gibbons, 2006; Longford, 1987). Other common names are variance component models (Dempster, Rubin and Tsutakawa, 1981) and hierarchical linear models (Raudenbush and Bryk, 2002).

In a nonlinear mixed effects model, the related regression function depends nonlinearly on fixed and random effects parameters. This is in contrast to a linear mixed effects model, where the regression function is a linear combination of fixed and random effects parameters. The integrated likelihood function of the nonlinear mixed effects model does not have a closed-form expression, which complicates the estimation algorithms. Nonlinear (mixed effects) models are referred to as generalized linear (mixed effects) models when the observations are distributed according to an exponential family distribution and when the predictor can be expressed as a linear term.

Typical for mixed effects regression models is the incorporation of random effects parameter(s) in regression models that account for dependencies between level-1 observations. The correlation structure of the level-1 observations is described by the random effects parameters, and the observations are assumed to be conditionally independent given the random effects parameters. The random effects parameters are assumed to be distributed according to a common population distribution.

In this light, an item response model can be recognized as a nonlinear mixed effects model. A characteristic element of item response data is that the observations from each respondent are not mutually independent. The individual's responses share a common underlying ability parameter and therefore the responses are said to be dependent. That is, the item responses are nested within subjects, which leads to a two-level structure. Note that the local independence assumption states that the observations are (only) conditionally independent given the ability parameter.

In the nonlinear mixed effects framework, the two-parameter model is stated as

$$h\left(E\left(Y_{ik} \mid \theta_i, b_k\right)\right) = \alpha_k \theta_i - b_k \tag{6.1}$$

where $h(.)$ is the so-called link function and equals the probit function $\Phi(.)^{-1}$ or the logit function $\Psi(.)^{-1}$. It can be seen that the success probabilities are transformed such that (1) they are linearly related to the term $a_k\theta_i - b_k$ and (2) the transformed success probabilities are mapped from the unit interval onto the whole real line. Given ability parameters, the model is a member of the class of generalized linear models (McCullagh and Nelder, 1989).

Let the ability parameters be normally distributed with unknown mean and variance. Then, a nonlinear mixed effects model is defined where the correlation structure of the (level-1) item responses is described by the random person parameters. Note that the model in terms of the underlying latent variable (Equation (4.7)) can also be recognized as a nonlinear mixed effects model. The right-hand side of Equation (4.38) contains a product of parameters, and the model is therefore not within the class of generalized linear mixed effects models. The Rasch model can be recognized as a generalized linear random intercept model.

Liu and Hedeker (2006) showed that within this framework the parameters can be estimated via marginal maximum likelihood estimation. The likelihood

equations are solved via multidimensional Gauss-Hermite quadrature. Liu and Hedeker mentioned that three levels of random effects can be handled and that the number of level-3 random effects is limited to three or four when Gauss-Hermite quadrature is used to numerically integrate over the distribution of random effects. The nonlinear mixed effects modeling framework, Equation (6.1), requires specific priors for the item parameters, and this limits its usefulness in more complex settings.

An intraclass correlation coefficient is easily computed as the proportion of (unexplained) variance that is between subjects, also called the between-subjects effect. Consider the unobserved continuous response vector $\mathbf{Z}_k$ for item k, where the level-1 variance is restricted to be one and the level-2 variance equals a_k^2 which leads to an intraclass correlation coefficient of $\rho_I = a_k^2/(1 + a_k^2)$. A highly discriminating item leads to a high proportion of variance between subjects. This makes sense since a steep ICC leads to more diverse responses in comparison with a flat ICC, which leads to more similar responses from respondents of different ability levels. That is, for various ability levels, the variation in success probabilities is higher for highly discriminating items (see Figure 1.2), and more variation is explained between subjects, than for lowly discriminating items. Note that ρ_I is a correlation coefficient since it defines the correlation between two random continuous responses of the same subject. Other definitions of the intraclass correlation coefficient are possible (Commenges and Jacqmin, 1994) and will lead to somewhat different results (Snijders and Bosker, 1999, p. 224).

The level-1 variance is fixed, which causes the nonlinear mixed effects model to act differently in comparison with the linear mixed effects model. In the linear case, the level-1 variance will reduce when incorporating explanatory variables in the model, but this cannot happen in this nonlinear model. Snijders and Bosker (1999) noted that in this particular case the variance of the random effects and the estimated regression effects will tend to become larger when adding a level-1 explanatory variable.

This generalized linear mixed effects model representation (which includes Rasch-type models) has some advantages. In some ways, standard Rasch-IRT models can be extended within this modeling framework, and standard software for generalized linear mixed effects models can be used. Current software packages using a likelihood-based estimation method for (nonlinear) mixed effects models include GLLAMM (generalized linear latent and mixed models; Skrondal and Rabe-Hesketh, 2004), HLM (hierarchical linear models; Raudenbush, Bryk, Cheong and Congdon, 2000), MIXOR (mixed effect ordinal regression; Hedeker and Gibbons, 1996), MIXNO (mixed effect nominal logistic regression; Hedeker, 1999), MLwiN (Goldstein et al., 1998), Mplus (Muthén and Muthén, 1998), NLME library in S+ (nonlinear mixed effects model; Pinheiro and Bates, 2000), and SAS PROC NLMIXED. The software package MLwiN also includes a Bayesian estimation method.

Adams, Wilson and Wu (1997) defined a Rasch model with a population model for the ability parameters that describes the between-subject variation

using explanatory information at the level of subjects as a nonlinear mixed effects logistic model. Kamata (2001), Maier (2001), and Raudenbush and Sampson (1999), among others, defined multilevel formulations of the Rasch model within this framework to study relationships between background information and levels of ability and to examine whether items function differently across groups of respondents. Rijmen, Tuerlinckx, De Boeck and Kuppens (2003) give an overview of item response models that are covered within the general class of nonlinear mixed effects models.

The framework of the generalized linear mixed effects model is too restrictive for the class of models that will be studied in this and subsequent chapters. First, it can only handle one-parameter (Rasch) models since the extended two- and three-parameter models are not within this class. Second, the class of models does not allow free prior choices for the model parameters. Simply stated, more complex priors may lead to a model that is not within this class. Moreover, MCMC methods make it possible to estimate simultaneously parameters of more complex models. Several MCMC algorithms have been developed for parameter estimation of nonlinear mixed effects models (e.g., Albert and Chib, 1993; Karim and Zeger, 1992; Zeger and Karim, 1991).

6.3 The Multilevel IRT Model

6.3.1 A Structural Multilevel Model

The student's ability is considered to be an outcome variable of the multilevel regression model. This outcome variable is not directly observable but is known to be a latent variable. The multilevel model including a latent variable with a known parametric distribution is called a structural multilevel model. The multilevel model describes the structure of the individual abilities. Before engaging in measuring the latent variable, assume for the moment that the abilities can be directly observed.

The population of level-2 clusters, say schools, are indexed $j = 1, \ldots, J$. The students at level 1 are nested in schools and indexed $i = 1, \ldots, n_j$. Let level-1 student-specific covariates be denoted by $\mathbf{x}_{ij} = (x_{0ij}, x_{1ij}, \ldots, x_{Qij})^t$, where x_{0ij} usually equals one. In general, the level-1 model is represented by

$$\theta_{ij} = \beta_{0j} + \ldots + \beta_{qj}x_{qij} + \ldots + \beta_{Qj}x_{Qij} + e_{ij}, \tag{6.2}$$

where the errors are independently and identically distributed with mean zero and variance σ_θ^2. The regression parameters are allowed to vary across schools. Level-2 school-specific covariates are denoted by $\mathbf{w}_{qj} = (w_{0qj}, w_{1qj}, \ldots, w_{Sqj})^t$, where w_{0qj} typically equals one. The random regression coefficients defined in Equation (6.2) are considered to be outcomes in the linear regression at level 2,

$$\beta_{qj} = \gamma_{q0} + \ldots + \gamma_{qs}w_{sqj} + \ldots + \gamma_{qS}w_{Sqj} + u_{qj}, \tag{6.3}$$

for $q = 0, \ldots, Q$, and where level-2 error terms, $\mathbf{u}_j$, are multivariate normally distributed with mean zero and covariance matrix $\mathbf{T}$. The elements of $\mathbf{T}$ are denoted by $\tau^2_{qq'}$ for $q, q' = 0, \ldots, Q$.

In the formulation above, the coefficients of all level-1 predictors are treated as random; that is, as varying across level-2 units. In certain applications, it can be desirable to constrain the effects of one or more of the predictors to be identical across level-2 units. In that case, the corresponding regression effect is considered to be a fixed effect.

The multilevel model in Equations (6.2) and (6.3) is more easily explained without explanatory variables. Therefore, consider the empty structural multilevel model,

$$\theta_{ij} = \beta_{0j} + e_{ij}, \tag{6.4}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{6.5}$$

where θ_{ij} is modeled by a school-specific intercept β_{0j} and the within-school level-1 error term e_{ij}. The school-specific intercept is broken down into a general mean γ_{00} and the deviation, u_{0j}, of the school-specific intercept from the general mean. The schools in a sample are thought to be representative of a larger population of schools, and the school-specific intercepts are treated as random effects. Most common is to assume a normal distribution. Therefore, the error term at level 2, u_{0j}, is assumed to be normally distributed with mean zero and variance τ^2_{00}. The error term at level 1 is normally distributed with mean zero and variance σ^2_θ.

It is typically assumed that any two abilities within the same school are correlated because they share the same random component, and two abilities related to different schools are uncorrelated. It follows that

$$\begin{aligned}
\text{Cov}\left(\theta_{ij}, \theta_{i',j'}\right) &= \text{Cov}\left(\beta_{0j} + e_{ij}, \beta_{0j'} + e_{i'j'}\right) \\
&= \text{Cov}\left(u_{0j} + e_{ij}, u_{0j'} + e_{i'j'}\right) \\
&= \text{Cov}\left(e_{ij}, e_{i'j'}\right) + \text{Cov}\left(u_{0j}, u_{0j'}\right) \\
&= \begin{cases} \sigma^2_\theta + \tau^2_{00} & \text{for } i = i', j = j' \\ \tau^2_{00} & \text{for } i \neq i', j = j' \\ 0 & \text{for } j \neq j', \end{cases}
\end{aligned}$$

where it is assumed that the school-specific intercepts are uncorrelated, and level-1 and level-2 residuals are uncorrelated. The covariance can be expressed as a correlation coefficient, which leads to the intraclass correlation coefficient

$$\begin{aligned}
\rho_I &= \frac{\text{Cov}\left(\theta_{ij}, \theta_{i'j'}\right)}{\sqrt{\text{Var}\left(\theta_{ij}\right)}\sqrt{\text{Var}\left(\theta_{i'j'}\right)}} \\
&= \frac{\tau^2_{00}}{\tau^2_{00} + \sigma^2_\theta},
\end{aligned}$$

when $i \neq i'$, and $j = j'$. This correlation coefficient presents the proportion of variance in individual abilities that is attributable to schools.

The empty multilevel model is a linear mixed effects model since the model can be presented in a fixed and a random part in addition to the level-1 residual e_{ij}. The total random part is composed of level-1 and level-2 residuals; that is,

$$\theta_{ij} = \underbrace{\gamma_{00}}_{\textit{fixed part}} + \underbrace{u_{0j} + e_{ij}}_{\textit{random part}}. \tag{6.6}$$

The variance components of the random part are regarded as between-school and within-school variances. Therefore, the model is also referred to as a variance components model.

The structural multilevel model with explanatory variables can also be presented as a linear mixed effects model. Therefore, stack the row vectors $\mathbf{x}_{ij}$ in the matrix $\mathbf{x}_j$ that represents the explanatory information of individuals in school j. The matrix $\mathbf{w}_j$ represents explanatory information on school j, and it contains the stacked vectors $\mathbf{w}_{qj}$ $(q = 0, \ldots, Q)$. A matrix of covariates will also be referred to as a design matrix without a direct reference to an experimental setting. The level-1 part of the model can be written as

$$\begin{bmatrix} \theta_{1j} \\ \vdots \\ \theta_{n_j j} \end{bmatrix} = \begin{bmatrix} 1 & x_{11j} & \ldots & x_{Q1j} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_j j} & \cdots & x_{Qn_j j} \end{bmatrix} \begin{bmatrix} \beta_{0j} \\ \vdots \\ \beta_{Qj} \end{bmatrix} + \begin{bmatrix} e_{0j} \\ \vdots \\ e_{n_j j} \end{bmatrix}$$

and the level-2 part as

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \vdots \\ \beta_{Qj} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{0j}^t & 0 & \ldots & 0 \\ 0 & \mathbf{w}_{1j}^t & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{w}_{Qj}^t \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \vdots \\ \vdots \\ \gamma_{QS} \end{bmatrix} + \begin{bmatrix} u_{0j} \\ u_{1j} \\ \vdots \\ u_{Qj} \end{bmatrix}.$$

In matrix notation, the model is written as

$$\boldsymbol{\theta}_j = \mathbf{x}_j \boldsymbol{\beta}_j + \mathbf{e}_j, \tag{6.7}$$

$$\boldsymbol{\beta}_j = \mathbf{w}_j \boldsymbol{\gamma} + \mathbf{u}_j. \tag{6.8}$$

Now, the single matrix equation for school j becomes

$$\begin{aligned} \boldsymbol{\theta}_j &= \mathbf{x}_j \left(\mathbf{w}_j \boldsymbol{\gamma} + \mathbf{u}_j \right) + \mathbf{e}_j \\ &= \mathbf{x}_j \mathbf{w}_j \boldsymbol{\gamma} + \mathbf{x}_j \mathbf{u}_j + \mathbf{e}_j \\ &= \tilde{\mathbf{x}}_j \boldsymbol{\gamma} + \mathbf{x}_j \mathbf{u}_j + \mathbf{e}_j, \end{aligned} \tag{6.9}$$

where $\mathbf{e}_j \sim \mathcal{N}\left(0, \sigma_\theta^2 \mathbf{I}_{n_j}\right)$ and $\mathbf{u}_j \sim \mathcal{N}\left(0, \mathbf{T}\right)$. The fixed effects design matrix $\tilde{\mathbf{x}}_j$ contains student-level, school-level, and the product of student- and school-level covariates that represent cross-level interactions. Such a cross-level interaction refers to the modification of the effect of a level-1 variable

by characteristics of the level-2 school to which an individual belongs (or vice versa). The term cross-level effect is used to denote a main effect of a level-2 variable on the level-1 outcome as well as the modification of the effect of a level-1 variable by a level-2 variable.

Finally, the (conditional) covariance structure of $\boldsymbol{\theta}_j$ can be written as

$$\text{Var}\left(\boldsymbol{\theta}_j \mid \mathbf{x}_j, \mathbf{w}_j\right) = \mathbf{x}_j \mathbf{T} \mathbf{x}_j^t + \sigma_\theta^2 \mathbf{I}_{n_j}.$$

The conditional covariance matrix of $\boldsymbol{\theta}_j$ is modeled in terms of a component that includes the random school effects and a component that includes the within-school error structure. The first component deals with the heterogeneity in the population of schools, and the second component posits an error structure that is the same for all schools.

This two-level framework is easily generalized to more levels (e.g., Goldstein, 2003; Raudenbush and Bryk, 2002).

6.3.2 The Synthesis of IRT and Structural Multilevel Models

The success of the structural multilevel model (Equations (6.2) and (6.3)) with a latent outcome variable depends partly on the appropriate handling of measurement issues regarding this latent dependent variable. In practice, often error prone measures are used, and ignoring measurement error in the estimated multilevel outcome variable can lead to biased structural multilevel parameter estimates. Since each student can be presented only a limited number of items, inference about his or her ability is subject to considerable uncertainty. This also includes response error due to the unreliability of the measurement instrument, and moreover, human response behavior is stochastic in nature.

In the present approach, the idea is to integrate an item response model for measuring the individual abilities with a structural multilevel model that explains differences at different levels of ability. The measurement model defines the relationship between the ability and the corresponding observed response data. The structural multilevel model describes the hierarchical structure of individuals in the population.

The observations are considered to be nested within the individuals and individuals nested in schools. At level 1, measurement error is defined as being associated with the individual ability. At level 2, differences in ability among individuals within the same school are modeled given student-level characteristics. At level 3, the variation across schools is modeled given background information at the school level. In Figure 6.1, a path diagram of this multilevel IRT (MLIRT) model is given. Three boxes are plotted, where the box indexed item is plotted in the box indexed individual, which is plotted in the box indexed school to illustrate the nesting of item observations in individuals and individuals in schools. The ellipse represents the item response model for measuring the multilevel outcome variable θ_{ij}. The two boxes indexed individual and school constitute the structural multilevel model for θ_{ij}. It can be

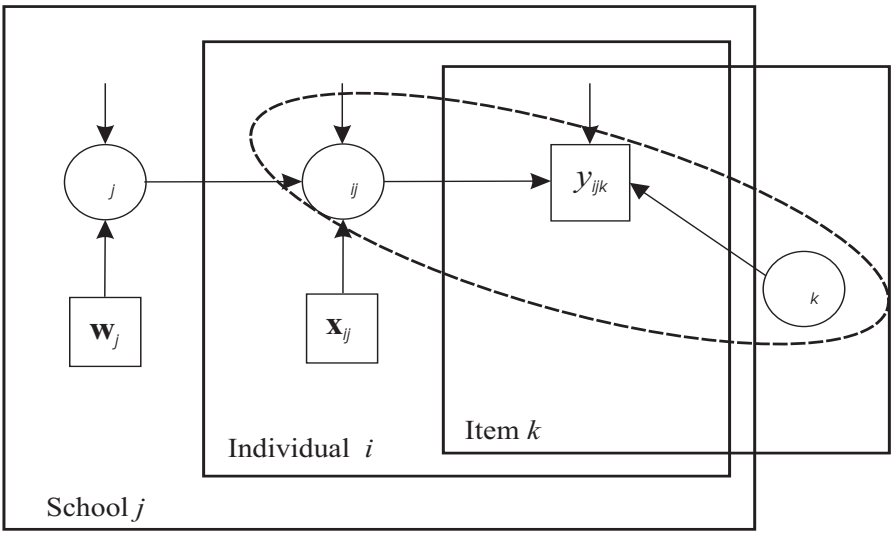

Fig. 6.1. Path diagram for the MLIRT model.

seen that there are three levels of uncertainty; (1) at the level of observations, (2) at the student level, and (3) at the school level. The explanatory information $\mathbf{x}_{ij}$ and $\mathbf{w}_j$ at the individual and school levels explain variability in the latent abilities within schools and the random regression effects across schools, respectively. As noted, the observations are nested within the students, and this is depicted as a box within the individual box. An ellipse is drawn to stress the nonlinear structure and the different meaning of the measurement part of the model. The item parameters are not influenced by the nested structure and are placed outside the boxes. The structural multilevel part of the model is focused on explaining variation via covariates and hierarchical levels, where the measurement part of the model, depicted as an ellipse, is focused on quantifying the measurement error corresponding to the nonlinear relationship between the discrete observations and the abilities.

The likelihood of the MLIRT model is obtained by integrating over the random effects distribution. This likelihood equation can be derived by following the structure given in Figure 6.1. Let $\mathbf{y}$ be the matrix of observed response data. Then the likelihood is a product of the likelihood for the J groups. Taking the nested structure into account and integrating over the random effects distributions, it follows that

$$p\left(\mathbf{y} \mid \boldsymbol{\xi}, \sigma_\theta^2, \boldsymbol{\gamma}, \mathbf{T}\right) = \prod_{j=1}^{J} \left[\int \left[\prod_{i=1|j}^{n_j} \int p\left(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}_k\right) p\left(\theta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j, \sigma_\theta^2\right) \mathrm{d}\theta_{ij} \right] \right.$$
$$\left. p\left(\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{T}\right) \right] \mathrm{d}\boldsymbol{\beta}_j. \tag{6.10}$$

The three levels of the model in Figure 6.1 can also be recognized from the likelihood in Equation (6.10). The distribution of the observations, $p(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}_k)$, represents the item response model at the lowest level. The second level is represented by the conditional distribution of the ability parameters given the random school effects parameters, $p\left(\theta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j, \sigma_\theta^2\right)$, and the highest level is described by the distribution of the random school effects, $p\left(\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{T}\right)$.

The MLIRT modeling framework has additional advantages. First, the structural multilevel model parameters are estimated from the item response data without having to condition on estimated person parameters. In "traditional" multilevel studies, estimated ability parameters considered to be outcome variables are treated as known. That is, they are assumed to be measured without an error. In the first stage, the abilities are estimated given a set of item responses using an item response model or by simply counting the number of correct responses. In the second stage, relationships between the estimated outcome variable and observed student variables and other group characteristics are analyzed using multilevel analysis techniques. Such a two-stage estimation procedure can cause serious underestimation of the standard errors of the model parameters due to the fact that some parameters are held fixed at values estimated from the data. Ignoring the uncertainty regarding the abilities within the model may lead to biased parameter estimates, and the statistical inference may be misleading (see also Section 6.6.3). In conclusion, this MLIRT framework leads to a proper treatment of the measurement error associated with the ability parameter.

Second, the MLIRT modeling framework allows the incorporation of explanatory variables at different levels of hierarchy. The inclusion of explanatory information can be important in various situations. Mislevy and Sheehan (1989) described different cases of using collateral information about examinees. A particular case is concerned with using collateral information to sample examinees and assign items to them. In this case, inconsistent item parameter estimates will be obtained when the parameters are estimated via integration over the examinees' population distribution and ignoring the collateral information. This follows from the fact that the missing explanatory information about the examinees cannot be seen as missing at random (MAR); the probability of the observed pattern of missingness is not the same for all values of the missing variables. A striking example is given in targeted testing, where examinees of different grade levels obtain different test items that are selected on the basis of the collateral information. In this case, the item parameter estimates will be inconsistent when the explanatory information is ignored.

Third, another related advantage of the model is that it can handle incomplete data in a very flexible way. To ease the notation, the index K is not given a subscript i, but variation across individuals in terms of the number of completed items is allowed. That is, there are no restrictions on the number of observations per individual, and it is possible that there are only a few common items across individuals. The number of individuals may differ across schools. Each school j may vary in the number of respondents since n carries

the subscript j. There are no restrictions on the number of students per school. This model feature follows in a straightforward way from properties of item response and multilevel models. Note that this flexibility assumes that the missing data are missing completely at random, also known as MCAR (Rubin, 1976). That is, item responses and subjects are missing for completely random reasons. The missing-data mechanism that characterizes the reasons for the missingness is known to be MCAR and constitutes patterns of missing data that are independent of both the observed data and the missing data.

Fourth, the MLIRT modeling framework provides shrinkage estimators that are based on an efficient combination of the response data and the collateral information. In general, a shrinkage estimator is biased, but a reduction in mean squared error can be achieved when this estimator has a smaller variance in comparison with an unbiased estimator. The shrinkage estimator of the ability parameter illustrates the combined use of response data and student-level collateral information. Therefore, consider the two-parameter response (probit) model formulation (Equation (4.38)) and the structural multilevel model for the ability parameters (Equations (6.2) and (6.3)). The latent response data $\mathbf{Z}_{ij}$ and the ability parameter θ_{ij} are bivariate normally distributed. That is,

$$\begin{bmatrix} \mathbf{Z}_{ij} \\ \theta_{ij} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{a}\theta_{ij} - \mathbf{b} \\ \mathbf{x}_{ij}^t \boldsymbol{\beta}_j \end{bmatrix}, \begin{bmatrix} \sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K & \sigma_\theta^2 \mathbf{a} \\ \sigma_\theta^2 \mathbf{a}^t & \sigma_\theta^2 \end{bmatrix} \right). \tag{6.11}$$

From a property of the bivariate normal distribution,[1] it follows that the conditional posterior expectation of ability can be expressed as (suppressing the conditioning for notational convenience)

$$\begin{aligned} E\left(\theta_{ij}\right) &= \mathbf{x}_{ij}^t \boldsymbol{\beta}_j + \sigma_\theta^2 \mathbf{a}^t \left(\mathbf{I}_K + \sigma_\theta^2 \mathbf{a}\mathbf{a}^t\right)^{-1} \left(\mathbf{z}_{ij} - \left(\mathbf{a}\mathbf{x}_{ij}^t \boldsymbol{\beta}_j - \mathbf{b}\right)\right) \\ &= \mathbf{x}_{ij}^t \boldsymbol{\beta}_j + \sigma_\theta^2 \mathbf{a}^t \left(\mathbf{I}_K - \frac{\mathbf{a}\mathbf{a}^t}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}}\right) \left(\left(\mathbf{z}_{ij} + \mathbf{b}\right) - \mathbf{a}\mathbf{x}_{ij}^t \boldsymbol{\beta}_j\right) \\ &= \mathbf{x}_{ij}^t \boldsymbol{\beta}_j - \frac{\mathbf{a}^t\mathbf{a}}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}} \mathbf{x}_{ij}^t \boldsymbol{\beta}_j + \frac{\mathbf{a}^t \left(\mathbf{z}_{ij} + \mathbf{b}\right)}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}} \\ &= \frac{\sigma_\theta^{-2} \mathbf{x}_{ij}^t \boldsymbol{\beta}_j}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}} + \frac{\mathbf{a}^t \left(\mathbf{z}_{ij} + \mathbf{b}\right)}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}} \\ &= \left(\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}\right)^{-1} \left(\left(\mathbf{a}^t\mathbf{a}\right) \hat{\theta}_{ij} + \sigma_\theta^{-2} \mathbf{x}_{ij}^t \boldsymbol{\beta}_j\right), \qquad (6.12) \\ &= \frac{\mathbf{a}^t\mathbf{a}}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}} \hat{\theta}_{ij} + \left(1 - \frac{\mathbf{a}^t\mathbf{a}}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}}\right) \mathbf{x}_{ij}^t \boldsymbol{\beta}_j, \qquad (6.13) \end{aligned}$$

[1] Let (X_1, X_2) be bivariate normally distributed with means (μ_1, μ_2) and variances $(\sigma_{x_1}^2, \sigma_{x_2}^2)$ and correlation ρ. The conditional distribution of X_1 given $X_2 = x_2$ is normal with mean $\mu_{x_1} + \rho\sigma_{x_2}^{-2}(x_2 - \mu_{x_2})$ and variance $\sigma_{x_1}^2 - \rho^2\sigma_{x_2}^{-2}$ (Anderson, 2003, pp. 33–36).

where $\hat{\theta}_{ij} = (\mathbf{a}^t\mathbf{a})^{-1}\,\mathbf{a}^t\,(\mathbf{z}_{ij} + \mathbf{b})$ and represents the least squares estimate for θ_{ij} given the latent response data and item parameters. In the second step for deriving the expression in (6.13), a specific expression for the so-called Schur complement of $\sigma_\theta^2\mathbf{a}\mathbf{a}^t + \mathbf{I}_K$ of the covariance matrix in (6.11) is used. In this case, the inverse of the Schur complement is given by

$$\left(\mathbf{A} + \lambda\mathbf{v}\mathbf{v}^t\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^t\mathbf{A}^{-1}}{1/\lambda + \mathbf{v}^t\mathbf{A}^{-1}\mathbf{v}}; \tag{6.14}$$

see Searle et al. (1992, p. 453).

The posterior mean of the corresponding posterior distribution is considered to be an estimate of the ability parameter. In this light, the result in Equation (6.13) is viewed as a shrinkage estimate of the ability parameter, which consists of a linear combination of two weighted estimates, the ordinary least squares estimate $\hat{\theta}_{ij}$ and, given an estimate of random regression effect $\boldsymbol{\beta}_j$, the (prior) level-2 estimate $\mathbf{x}_{ij}^t\boldsymbol{\beta}_j$. The weights are proportional to the variance of the ordinary least squares estimate and the variance σ_θ^2. When reducing the number of items (level 1), the variance of the $\hat{\theta}_{ij}$ increases and $\mathbf{a}^t\mathbf{a}$ decreases, and the posterior mean moves towards the level-2 mean. When increasing the number of (informative) items, the least squares estimate is less corrected by the level-2 mean. The shrinkage estimate can be viewed as a compromise between the within-subject estimate that is based on the individual's item response data and the between-subject estimate that is based on the information from the structural population model. It can be concluded that the use of the examinees' explanatory information may lead to a more accurate estimate of the ability parameters.

Finally, in the same way, the use of explanatory information may lead to more accurate item parameter estimates. In that case, the posterior distribution of the ability parameters is based on item response data and collateral information. Both provide information about the examinees' ability parameters. Subsequently, a more accurate marginal posterior density of the item parameters can be obtained when integrating over the more informative posterior density of ability parameters that is based on collateral and response information. Note that the marginal posterior density of the item parameters can be expressed as

$$\begin{aligned} p\left(\boldsymbol{\xi}_k \mid \mathbf{y}, \mathbf{x}\right) &= \int p\left(\boldsymbol{\xi}_k \mid \mathbf{y}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}; \boldsymbol{\beta}, \sigma_\theta^2, \mathbf{x}\right) \mathrm{d}\boldsymbol{\theta} \\ &= \int \frac{p\left(\mathbf{y} \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}\right) p(\boldsymbol{\xi}_k) p\left(\boldsymbol{\theta}; \boldsymbol{\beta}, \sigma_\theta^2, \mathbf{x}\right)}{p(\mathbf{y})} \mathrm{d}\boldsymbol{\theta} \\ &= \int p\left(\mathbf{y} \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}\right) p(\boldsymbol{\xi}_k) \frac{p\left(\boldsymbol{\theta}; \boldsymbol{\beta}, \sigma_\theta^2, \mathbf{x}\right)}{p(\mathbf{y})} \mathrm{d}\boldsymbol{\theta} \\ &= \int p\left(\mathbf{y} \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}\right) p(\boldsymbol{\xi}_k) \frac{p\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\beta}, \sigma_\theta^2, \mathbf{x}\right)}{p\left(\mathbf{y} \mid \boldsymbol{\theta}\right)} \mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{6.15}$$

$$= \int \frac{p\left(\mathbf{y} \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}\right) p(\boldsymbol{\xi}_k)}{p\left(\mathbf{y} \mid \boldsymbol{\theta}\right)} p\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\beta}, \sigma_\theta^2, \mathbf{x}\right) \mathrm{d}\boldsymbol{\theta}$$

$$= \int p\left(\boldsymbol{\xi}_k \mid \mathbf{y}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\beta}, \sigma_\theta^2, \mathbf{x}\right) \mathrm{d}\boldsymbol{\theta}, \tag{6.16}$$

where the semicolon notation is used to stress that the subsequent parameters are assumed to be known.

In Equation (6.15), the integration is performed with respect to the *prior* density of the ability parameters using the collateral information. After some transformations, it follows that the marginal posterior density of the item parameters can also be obtained via integration using the conditional *posterior* density of the ability parameters, Equation (6.16). As a result, the marginal posterior density of the item parameters, which is obtained via prior or posterior integration, might be improved by using collateral information.

Note that the use of an item response model, in contrast to an observed (sum) score, also has some advantages. First, the structural multilevel model with a latent outcome variable measured with an item response model has the advantage that latent rather than observed scores are used as dependent variables, which offers the possibility of separating the influences of item difficulty and ability level and modeling response variation and measurement error. Second, contrary to observed scores, latent scores are test-independent, which offers the possibility of using results from different tests in one analysis where the parameters of the item response model and the multilevel model can be estimated concurrently.

6.4 Estimating Level-3 Residuals: School Effects

There is increasing interest in the accountability of educational institutions. Research has focused on measuring the "quality" of schools and making quantitative comparisons between schools. So-called performance indicators can be used to judge school effectiveness and determine what measures can be taken for improvement, obtains knowledge about the relative size of school differences, and determine the extent to which other indicators may explain differences. Within this context, attention has focused on contextual differences and the appropriate specification of a statistical model.

Aitkin and Longford (1986), among others, modeled school effects as higher-level residuals and showed interest in the actual realized values to compare schools and in the distributional properties of the residuals. For example, a significant school-level variance indicates the existence of differences among schools. This method for comparing schools is not without criticism. Obviously, the residuals behave like noise, which complicates the inferences. The random error term may contain a contextual part that should be excluded when the purpose is assessing the effectiveness of a school. A comprehensive review of school effectiveness research can be found in Scheerens (1992).

Although several student outcome measures need to be used, for a full understanding of school effectiveness research, one kind of measure is used and attention is focused on the reliability of this outcome measure. In the present approach, different sources of variation are taken into account when estimating the school effects: within-subject, between-subject or within-school, between-school variability.

For illustration, the contribution of a school, the so-called school effect, to the abilities of its students is analyzed by a multilevel model, Equations (6.4) and (6.5). The MLIRT model can be presented as one equation by entering the multilevel expressions in the normal latent response formulation of Equation (4.38). Then, the stacked individual response vectors, each related to K items, of the n_j persons in school j, denoted as $\mathbf{Z}_j$, can be represented as

$$\begin{bmatrix} Z_{1j1} \\ \vdots \\ Z_{1jK} \\ \vdots \\ Z_{n_jjK} \end{bmatrix} = \begin{bmatrix} a_1\gamma_{00} - b_1 + a_1u_{0j} + a_1e_{1j} + \epsilon_{1j1} \\ \vdots \\ a_K\gamma_{00} - b_K + a_Ku_{0j} + a_Ke_{1j} + \epsilon_{1jK} \\ \vdots \\ a_K\gamma_{00} - b_K + a_Ku_{0j} + a_Ke_{n_jj} + \epsilon_{n_jjK} \end{bmatrix}, \tag{6.17}$$

where three independent random error terms can be recognized. The ϵ_{ijk} represents the random deviation between the Z_{ijk} and the mean term $a_k\theta_{ij} - b_k$. This within-individual residual variation is fixed at one due to an identification restriction (Section 4.4). The e_{ij} represents the between-individual or within-school residual variation given the school effect and reflects the variation around the mean $\gamma_{00} + u_{0j}$. The third residual term, u_{0j}, represents the random school effect. The normally distributed random error terms are independent and have a mean of zero. The within-school variance is denoted by σ_θ^2 and the between-school variance by τ^2.

The equations for the n_j times K latent observations $\mathbf{Z}_j$ can be presented in matrix notation by utilizing a summing vector $\mathbf{1}_{n_j} = (1, \ldots, 1)^t$ of n_j elements of one. A matrix of ones is denoted as $\mathbf{J}_{n_j}$, which equals $\mathbf{1}_{n_j}\mathbf{1}_{n_j}^t$. Subsequently, the vector $\mathbf{1}_{n_j} \otimes \mathbf{a}$ is the stacked vector of n_j vectors $\mathbf{a} = (a_1, \ldots, a_K)^t$, where matrix operation $\otimes$ is known as the direct product or Kronecker product. Now, Equation (6.17) can be presented as

$$\mathbf{Z}_j = \mathbf{1}_{n_j} \otimes (\mathbf{a}\gamma_{00} - \mathbf{b}) + \mathbf{1}_{n_j} \otimes \mathbf{a}u_{0j} + \mathbf{e}_j \otimes \mathbf{a} + \boldsymbol{\epsilon}_j. \tag{6.18}$$

Interest is focused on estimating the random error term u_{0j} via the posterior expectation of u_{0j}. Therefore, consider the multivariate normally distributed latent response data and the normally distributed random effects, $(\mathbf{Z}_j, \boldsymbol{\theta}_j, u_{0j})^t$. The nested structure of observations within individuals and of individuals within schools becomes apparent from the variance-covariance structure. Via Equation (6.18), it follows that, suppressing the conditioning,

$$\begin{aligned}\text{Var}\left(\mathbf{Z}_j\right) &= \text{Var}\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right) u_{0j} + \text{Var}\left(\mathbf{e}_j \otimes \mathbf{a}\right) + \text{Var}\left(\boldsymbol{\epsilon}_j\right) \\ &= \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right) \tau^2 \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t + \mathbf{I}_{n_j} \otimes \sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_{(n_j \cdot K)} \\ &= \mathbf{J}_{n_j} \otimes \left(\tau^2 \mathbf{a}\mathbf{a}^t\right) + \mathbf{I}_{n_j} \otimes \left(\sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K\right).\end{aligned}$$

The covariance structure of the observations $\mathbf{Z}_j$ consists of three terms. The first term reflects the covariance structure induced by the nesting of individuals in schools. The second term reflects the nesting of observations within individuals. The discrimination parameters function as weights in such a way that high (low) discriminating parameter values cause a strong (weak) covariance structure in individual observations. The last term is a unity matrix. In the same way, the entire dispersion matrix of the random variables, $\left(\mathbf{Z}_j, \boldsymbol{\theta}_j, u_{0j}\right)^t$, equals

$$\left[\begin{array}{c|c|c} \mathbf{J}_{n_j} \otimes \left(\tau^2 \mathbf{a}\mathbf{a}^t\right) + \mathbf{I}_{n_j} \otimes \left(\sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K\right) & \mathbf{J}_{n_j} \otimes \tau^2 \mathbf{a} + \mathbf{I}_{n_j} \otimes \sigma_\theta^2 \mathbf{a} & \mathbf{1}_{n_j} \otimes \tau^2 \mathbf{a} \\ \mathbf{J}_{n_j} \otimes \tau^2 \mathbf{a}^t + \mathbf{I}_{n_j} \otimes \sigma_\theta^2 \mathbf{a}^t & \sigma_\theta^2 \mathbf{I}_{n_j} + \tau^2 \mathbf{J}_{n_j} & \mathbf{1}_{n_j} \tau^2 \\ \mathbf{1}_{n_j}^t \otimes \tau^2 \mathbf{a}^t & \mathbf{1}_{n_j}^t \tau^2 & \tau^2 \end{array}\right].$$

For the moment, assume that the student abilities are measured exactly. Interest is often focused on the population of schools; in particular, the variability of the school effects. Here, attention is focused on the size of the school effects given the individual abilities. The abilities and random school effects are bivariate normally distributed. The conditional expected value of the school effects given the abilities, suppressing the conditioning on other model parameters, equals

$$\begin{aligned}E\left(u_{0j} \mid \boldsymbol{\theta}_j\right) &= E\left(u_{0j}\right) + \text{Cov}\left(u_{0j}, \boldsymbol{\theta}_j\right) \left(\text{Var}\left(\boldsymbol{\theta}_j\right)\right)^{-1} \left(\boldsymbol{\theta}_j - E(\boldsymbol{\theta}_j)\right) \\ &= \mathbf{1}_{n_j}^t \tau^2 \left(\tau^2 \mathbf{J}_{n_j} + \sigma_\theta^2 \mathbf{I}_{n_j}\right)^{-1} \left(\boldsymbol{\theta}_j - \mathbf{1}_{n_j} \gamma_{00}\right) \\ &= \frac{n_j \tau^2}{\sigma_\theta^2} \left(\overline{\theta}_j - \gamma_{00}\right) - \frac{n_j \tau^2}{\sigma_\theta^2} \left(\frac{n_j \tau^2}{\sigma_\theta^2 + n_j \tau^2}\right) \left(\overline{\theta}_j - \gamma_{00}\right) \\ &= \frac{n_j \tau^2}{\sigma_\theta^2 + n_j \tau^2} \left(\overline{\theta}_j - \gamma_{00}\right) \\ &= \left(\overline{\theta}_j - \gamma_{00}\right) - \frac{\sigma_\theta^2}{\sigma_\theta^2 + n_j \tau^2} \left(\overline{\theta}_j - \gamma_{00}\right), \end{aligned} \tag{6.19}$$

using that

$$\left(\tau^2 \mathbf{J}_{n_j} + \sigma_\theta^2 \mathbf{I}_{n_j}\right)^{-1} = \sigma_\theta^{-2} \left(\mathbf{I}_{n_j} - \frac{\tau^2}{\sigma_\theta^2 + n_j \tau^2} \mathbf{J}_{n_j}\right). \tag{6.20}$$

In Equation (6.19), the conditional expected posterior value is expressed in a shrinkage representation. The expected posterior value is a shrinkage estimate of the school effect. In estimating the school effect, u_{0j}, the overall mean, in this case zero, is biased, but the unbiased estimate $\overline{\theta}_j - \gamma_{00}$ has a larger variance. For the shrinkage estimate, Equation (6.19), it follows that when $\overline{\theta}_j$

exceeds γ_{00}, the expected school effect is less than $\overline{\theta}_j - \gamma_{00}$, whereas for $\overline{\theta}_j$ less than γ_{00}, the expected school effect exceeds $\overline{\theta}_j - \gamma_{00}$. But the expected school effect is only corrected by a fraction of $\overline{\theta}_j - \gamma_{00}$ and not by $\overline{\theta}_j - \gamma_{00}$ itself. For large within-school variances and small numbers of students per school, the shrinkage estimate is much more efficient than the overall mean or $\overline{\theta}_j - \gamma_{00}$. In this particular case, the unbiased within-sample mean has a large variance.

The fraction defined in Equation (6.19) is a combination of within-school variance and between-school variance; that is, the empty multilevel model takes these different sources of variation into account. The within-subject variance is the additional source of variation that is taken into account. Therefore, the conditional expected posterior value of u_{0j} given the latent response data is considered to be an estimate of the school effect. The conditional expected posterior value of u_{0j} can be derived in a comparable way. From Appendix 6.9, it follows that

$$\begin{aligned} E\left(u_{0j} \mid \mathbf{z}_j\right) &= \left(\mathbf{1}_{n_j}^t \otimes \tau^2 \mathbf{a}^t\right) \left[\left(\mathbf{J}_{n_j} \otimes \tau^2 \mathbf{a}\mathbf{a}^t\right) + \left(\mathbf{I}_{n_j} \otimes \left(\sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K\right)\right)\right]^{-1} \\ &\quad \cdot \left(\mathbf{z}_j - \left(\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)\gamma_{00} - \mathbf{1}_{n_j} \otimes \mathbf{b}\right)\right) \qquad (6.21) \\ &= \left(\frac{n_j \tau^2}{((\mathbf{a}^t\mathbf{a})^{-1} + \sigma_\theta^2) + n_j \tau^2}\right) \hat{u}_{0j} \\ &= \hat{u}_{0j} - \left(\frac{(\mathbf{a}^t\mathbf{a})^{-1} + \sigma_\theta^2}{((\mathbf{a}^t\mathbf{a})^{-1} + \sigma_\theta^2) + n_j \tau^2}\right) \hat{u}_{0j}, \qquad (6.22) \end{aligned}$$

where

$$\hat{u}_{0j} = (\mathbf{a}^t\mathbf{a})^{-1}\mathbf{a}^t\left(\overline{\mathbf{z}}_j - (\mathbf{a}\gamma_{00} - \mathbf{b})\right)$$

and $\overline{\mathbf{z}}_j = \sum_{i=1}^{n_j} \mathbf{z}_{ij}/n_j$.

The estimator is a weighted sum of the mean of weighted latent responses of the students of a school indexed j and the overall population mean, which in this case is zero. The shrinkage factor that determines the movement of this least squares estimator to the overall mean consists of the between-school and within-school variances and the inner product of the discrimination parameters, called measurement variance. Relatively high discrimination parameters indicate that the abilities of the examinees can be distinguished quite well from each other given their response patterns. When the information regarding the abilities of the students is relatively high, the estimated school effect shrinks less towards the overall mean. In that case, schools can be better distinguished from each other with respect to the estimated school effects. Relatively low discrimination parameters move the estimate of the school effect towards the overall mean since the estimated abilities of the students cannot be distinguished accurately from each other.

The least squares estimator $\hat{u}_{0j}$ is constructed as the mean over all n_j individual least squares estimators. The variance of the least squares estimator equals $(\mathbf{a}^t\mathbf{a})^{-1}$, and this measurement variance is influenced by the number of

items and the values of the discrimination parameters. The expected posterior mean shrinks towards the overall mean when decreasing the between-school variance, τ^2, or increasing the within-school variance, σ_θ^2, and/or the measurement variance $(\mathbf{a}^t\mathbf{a})^{-1}$.

The shrinkage factor in Equation (6.19) equals the shrinkage factor in Equation (6.22) when the inner product of discrimination parameters equals zero such that the measurement variance is ignored. In Figure 6.2, both shrinkage factors are plotted for various conditions. The discrimination parameter values equal one such that the inner product of discrimination values equals K. Population parameter γ_{00} equals zero. The (horizontal) x-axis presents the number of items from three to ten and the number of respondents in school j from 30 to 100. The shrinkage factor of Equation (6.22), plotted as a straight line and labeled random, decreases when the number of items or the number of individuals increases. The shrinkage factor of Equation (6.19), plotted as a broken line and labeled fixed, decreases when the number of individuals increases.

For relatively small within- and between-school variances, the upper figure in Figure 6.2 shows that the measurement variance has a large effect on the value of the shrinkage factor. The variance of the least squares estimator increases due to the small number of items and as a result the expected school effect is shrunk towards zero. When ignoring the measurement variance, the corresponding shrinkage factor is small due to the small values for τ^2 and σ_θ^2, which leads to a relatively high estimated school effect. Ignoring the measurement variance leads to higher school effects in absolute values and a sharper distinction between schools. Furthermore, covariates that explain variation in school effects are more likely to be significant when ignoring the measurement variance. Note that the small values of τ^2 and σ_θ^2 correspond to the realistic situation where additional background information explains a large part of the within-school and between-school variations. The middle and lower figures show the shrinkage factors for increasing values of the within- and between-school variations. It follows that the measurement variance part has a smaller impact on the shrinkage factor when the other sources of variance increase. In that case, the straight lines and the broken lines are closer to each other but differences remain visible.

The plotted shrinkage factors are based on the true known values of other model parameters, which are usually unknown. Ignoring measurement variance leads to higher estimated school effects in absolute values. This will lead to a biased between-school variance estimate since the schools appear to differ more with respect to the estimated school effects. The degree of shrinkage depends on unknown model parameters, and they need to be estimated from the data as well. Ignoring the measurement variance influences the estimates of the school effects and the other model parameters.

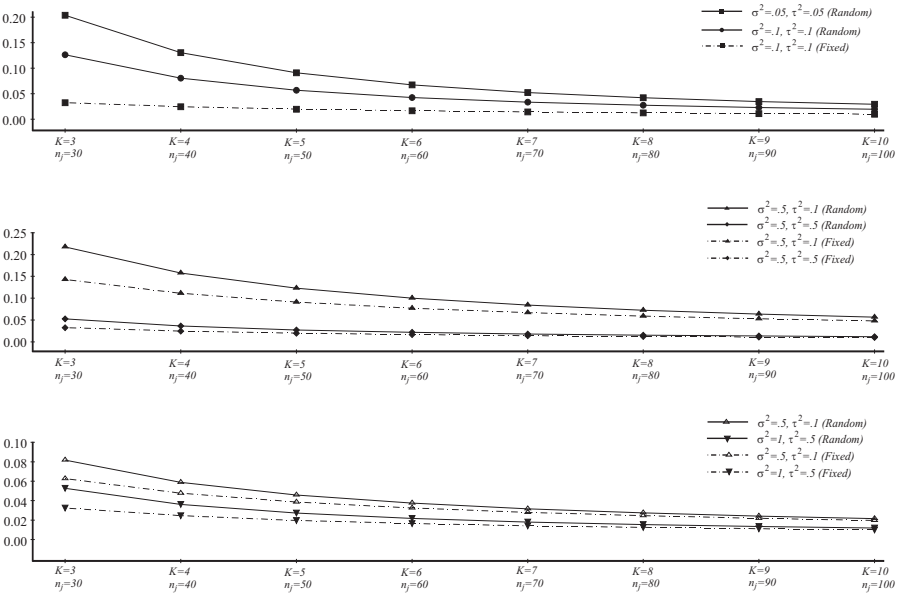

Fig. 6.2. Shrinkage factors of the conditional expected (residual) school effect U_{0j}.

6.5 Simultaneous Parameter Estimation of MLIRT

A Markov chain Monte Carlo (MCMC) method is constructed to estimate all model parameters simultaneously. Assume the structural multilevel model in Equations (6.2) and (6.3) in combination with a two-parameter model for binary discrete item observations or a graded response model for ordinal response data.

Assume the following prior distributions for the structural multilevel model parameters. The fixed effects, $\boldsymbol{\gamma}$, are assumed to have an independent normal prior, with mean zero and variance σ_γ, and the hyperparameter σ_γ equals a large number to specify a noninformative prior. The prior for the covariance matrix $\mathbf{T}$ is taken to be an inverse Wishart density,

$$p\left(\mathbf{T} \mid n_q, \mathbf{S}_T\right) \propto |\mathbf{T}|^{-(n_q+Q+2)/2} \exp\left(-\operatorname{tr}\left(\mathbf{S}_T\mathbf{T}^{-1}\right)/2\right),$$

with, for example, unity matrix $\mathbf{S}_T$ and n_q $(n_q \geq Q+1)$ equal to a small number to specify a proper diffuse prior. The conventional prior for σ_θ^2 is the inverse gamma with parameters g_1 and g_2 with density function

$$p\left(\sigma_\theta^2\right) \propto \left(\sigma_\theta^2\right)^{-(g_1+1)} \exp\left(-\frac{g_2}{\sigma_\theta^2}\right).$$

A proper noninformative prior is specified with $g_1 = 1$ and a small value for g_2. The following MCMC scheme for the MLIRT model contains slightly modified steps of schemes 2 and 3.

MCMC Scheme 4 (MLIRT)

A1) Binary response data

1. *Sample augmented data* $\mathbf{z}^{(m+1)}$ *according to Equation* (4.7).
2. *Assume an exchangeable multivariate normal prior for* $\boldsymbol{\xi}_k$ *with mean* $\boldsymbol{\mu}_\xi$ *and variance* $\boldsymbol{\Sigma}_\xi$. *Let* $\mathbf{H} = \left(d\boldsymbol{\theta}^{(m)}, -d\mathbf{1}_n\right)$, *and sample* $\boldsymbol{\xi}_k^{(m+1)}$ *from the conditional density*

$$\boldsymbol{\xi}_k \mid \mathbf{z}_k^{(m+1)}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\mu}_\xi^{(m)}, \boldsymbol{\Sigma}_\xi^{(m)} \sim \mathcal{N}\left(\boldsymbol{\mu}_\xi^*, \boldsymbol{\Omega}_\xi\right), \tag{6.23}$$

where

$$\begin{aligned}\boldsymbol{\Omega}_\xi^{-1} &= \varphi^{-1}\left(\mathbf{H}^t\mathbf{H}\right) + \boldsymbol{\Sigma}_\xi^{-1},\\ \boldsymbol{\mu}_\xi^* &= \boldsymbol{\Omega}_\xi\left(\mathbf{H}^t\mathbf{z}_k + \boldsymbol{\mu}_\xi\boldsymbol{\Sigma}_\xi^{-1}\right),\end{aligned}$$

with $\varphi = 1$ *if* $\mathbf{Z}_k$ *is normally distributed and* $\varphi = \pi^2/3$ *if* $\mathbf{Z}_k$ *is logistically distributed. The density in Equation* (6.23) *is used to generate candidates evaluated in an M-H step when* $\mathbf{Z}_k$ *is logistically distributed.*

3. *Sample* $\boldsymbol{\mu}_\xi^{(m+1)}, \boldsymbol{\Sigma}_\xi^{(m+1)}$ *from the conditional densities*

$$\begin{aligned}\boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi^{(m)}, \boldsymbol{\xi}^{(m+1)} &\sim \mathcal{N}\left(\frac{K_0}{K_0+K}\boldsymbol{\mu}_0 + \frac{K}{K_0+K}\overline{\boldsymbol{\xi}}, \boldsymbol{\Sigma}_\xi/(K+K_0)\right)\\ \boldsymbol{\Sigma}_\xi \mid \boldsymbol{\xi}^{(m+1)} &\sim \mathcal{IW}\left(K+\nu, \boldsymbol{\Sigma}^*\right),\end{aligned}$$

where $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0 + K\mathbf{S} + \frac{KK_0}{K+K_0}\left(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_0\right)\left(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_0\right)^t$.

A2) Polytomous response data

1. *Sample augmented data* $\mathbf{z}^{(m+1)}$ *given* $\boldsymbol{\theta}^{(m)}, \boldsymbol{\xi}^{(m)}$ *according to Equation* (4.32).
2. *Sample threshold parameter values* $\boldsymbol{\kappa}^{(m+1)}$ *given* $\mathbf{z}^{(m+1)}, \boldsymbol{\theta}^{(m)}, \mathbf{a}^{(m)}$ *according to step 3 of MCMC scheme 3.*
3. *For each* k, *sample parameters* $a_k^{(m+1)}$ *given* $\mathbf{z}_k^{(m+1)}, \boldsymbol{\theta}^{(m)}, \mu_a^{(m)}, \sigma_a^{2(m)}$ *according to step 4 of MCMC scheme 3.*

B) Sample the multilevel parameter values

4. *For each* i *and* j, *sample* θ_{ij} *from the conditional normal density*

$$\theta_{ij} \mid \mathbf{z}_{ij}^{(m+1)}, \boldsymbol{\xi}^{(m+1)}, \boldsymbol{\beta}_j^{(m)}, \sigma_\theta^{2(m)} \sim \mathcal{N}\left(\mu_\theta, \Omega_\theta\right), \tag{6.24}$$

where

$$\mu_\theta = \Omega_\theta\left(\left(\mathbf{a}^t\mathbf{a}\right)\hat{\theta}_{ij} + \sigma_\theta^{-2}\mathbf{x}_{ij}^t\boldsymbol{\beta}_j\right), \tag{6.25}$$

$$\Omega_\theta = \left(\mathbf{a}^t\mathbf{a} + \sigma_\theta^{-2}\right)^{-1}. \tag{6.26}$$

The density in (6.24) *becomes a proposal density when the latent responses are logistically distributed.*

5. *For each j, sample $\boldsymbol{\beta}_j^{(m+1)}$ from the full conditional*

$$\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j^{(m+1)}, \sigma_\theta^{2(m)}, \mathbf{T}^{(m)}, \boldsymbol{\gamma}^{(m)} \sim \mathcal{N}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Omega}_\beta\right), \tag{6.27}$$

where

$$\boldsymbol{\mu}_\beta = \boldsymbol{\Omega}_\beta \left(\boldsymbol{\Sigma}_j^{-1} \hat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1} \mathbf{w}_j \boldsymbol{\gamma}\right),$$
$$\boldsymbol{\Omega}_\beta = \left(\boldsymbol{\Sigma}_j^{-1} + \mathbf{T}^{-1}\right)^{-1},$$

with $\boldsymbol{\Sigma}_j = \sigma_\theta^2 \left(\mathbf{x}_j^t \mathbf{x}_j\right)^{-1}$ and $\hat{\boldsymbol{\beta}}_j = \left(\mathbf{x}_j^t \mathbf{x}_j\right)^{-1} \mathbf{x}_j^t \boldsymbol{\theta}_j$.

6. *Sample $\boldsymbol{\gamma}^{(m+1)}$ from the full conditional*

$$\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(m+1)}, \mathbf{T}^{(m)}, \sigma_\gamma \sim \mathcal{N}\left(\boldsymbol{\mu}_\gamma, \boldsymbol{\Omega}_\gamma\right),$$

where

$$\boldsymbol{\mu}_\gamma = \boldsymbol{\Omega}_\gamma \sum_j \mathbf{w}_j^t \mathbf{T}^{-1} \boldsymbol{\beta}_j,$$
$$\boldsymbol{\Omega}_\gamma = \left(\sum_j \mathbf{w}_j^t \mathbf{T}^{-1} \mathbf{w}_j + \mathbf{I}_\nu \sigma_\gamma^{-1}\right)^{-1},$$

where unity matrix $\mathbf{I}_\nu$ is of dimension $\nu = (Q+1)(S+1)$.

7. *Sample $\sigma_\theta^{2(m+1)}$ given $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\beta}^{(m+1)}$ from an inverse gamma density with shape parameter $g_1 + \sum_j n_j/2$ and scale parameter*

$$\sum_j \left(\boldsymbol{\theta}_j - \mathbf{x}_j \boldsymbol{\beta}_j\right)^t \left(\boldsymbol{\theta}_j - \mathbf{x}_j \boldsymbol{\beta}_j\right)/2 + g_2.$$

8. *Sample $\mathbf{T}^{(m+1)}$ given $\boldsymbol{\gamma}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}$ from an inverse Wishart density with $n_q + J$ degrees of freedom and scale matrix*

$$\sum_j \left(\boldsymbol{\beta}_j - \mathbf{w}_j \boldsymbol{\gamma}\right)\left(\boldsymbol{\beta}_j - \mathbf{w}_j \boldsymbol{\gamma}\right)^t + \mathbf{S}_T.$$

The MCMC scheme of Fox and Glas (2001) only considers binary response data and the normal ogive item response model. In this scheme, logistic and normal ogive item response models are considered and more advanced prior distributions for the item parameters are used, as discussed in Chapter 2. MCMC scheme 4 is easily extended to handle a structural multilevel model with more than two levels. Restricting the mean and variance of the latent ability scale is a common way of identifying the MLIRT model. This can be done during the MCMC estimation procedure by transforming each drawn vector of ability values in such a way that the mean and variance of the scaled vector correspond with the a priori specified mean and variance. This

procedure corresponds to fixing the mean and variance of the conditional posterior distribution of the latent variable. The MLIRT model can also be identified via restrictions on the item parameters (see Section 4.4).

The log of the complete-data likelihood of the MLIRT model can be presented as

$$\log p\big(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{\xi}, \boldsymbol{\gamma}, \sigma_\theta^2, \mathbf{T}\big) = \sum_{i,j} \Bigg(\underbrace{\sum_k \log p\left(y_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k\right)}_{\textit{Item response part}} + \underbrace{\log p\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma_\theta^2\right) + \log p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right)}_{\textit{Structural multilevel part}} \Bigg).$$

This complete-data log-likelihood of the MLIRT model consists of two parts, a part following from the item response model and a part following from the structural multilevel model. The idea is that model changes in the multilevel part can be tested conditional on the item response part such that relatively small changes in the log-likelihood of the multilevel part can be detected. In practice, MLIRT models with different structural multilevel parts and equivalent item response parts are often compared with each other. Therefore, attention is focused on the multilevel part, and the likelihood of interest is defined as

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \sigma_\theta^2, \mathbf{T}\right) = \frac{p\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \sigma_\theta^2\right) p\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{T}\right)}{p\left(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\theta^2, \mathbf{T}\right)} \tag{6.28}$$

$$= \prod_j \left(2\pi\sigma_\theta^2\right)^{-n_j/2} |\boldsymbol{\Omega}_\beta|^{1/2} |\mathbf{T}|^{-1/2} \exp\left[\frac{-1}{2\sigma_\theta^2}\left(\boldsymbol{\theta}_j - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right)^t \left(\boldsymbol{\theta}_j - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right) - \frac{1}{2}\left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1}\left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)\right], \tag{6.29}$$

where $\tilde{\boldsymbol{\beta}}_j$ is normally distributed with mean $\boldsymbol{\mu}_\beta$ and variance $\boldsymbol{\Omega}_\beta$ according to Equation (6.27). Equation (6.28) holds for all $\boldsymbol{\beta}_j$ and in particular when $\boldsymbol{\beta}_j = \tilde{\boldsymbol{\beta}}_j$ (Dempster et al., 1981). The log-likelihood term of Equation (6.29) can be used as a deviance term for model comparison; that is,

$$D\left(\boldsymbol{\gamma}, \sigma_\theta^2, \mathbf{T}; \boldsymbol{\theta}\right) = N \log\left(2\pi\sigma_\theta^2\right) - J \log|\boldsymbol{\Omega}_\beta| + J \log|\mathbf{T}| + \frac{1}{\sigma_\theta^2}\left(\boldsymbol{\theta} - \mathbf{x}\tilde{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{\theta} - \mathbf{x}\tilde{\boldsymbol{\beta}}\right) + \sum_j \left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1}\left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right). \tag{6.30}$$

The Bayesian information criterion (BIC) can be computed using the deviance defined in Equation (6.30) since the number of parameters is specified directly. The number of model parameters is needed to define the penalty

term in the BIC. When the deviance term contains random effects parameters, a deviance information criterion (DIC) can be computed using output from MCMC scheme 4. The DIC consists of two terms. One term is the deviance of the multilevel part, which is evaluated at the posterior means. The other term is the expected posterior deviance, where the expectation is taken with respect to the posterior distribution of the structural multilevel model parameters.

The DIC of the structural multilevel part conditions on the latent dependent variable, which can be integrated out via its marginal posterior distribution. That is, for both terms of the DIC, the posterior expectation can be computed with respect to the marginal posterior density of $\boldsymbol{\theta}$.

An integrated likelihood of an MLIRT model is derived in Appendix 6.10. The integrated likelihood in Equation (6.44) can be used for model comparison without having to condition on random effects parameter estimates. When a deviance term is defined based on the likelihood in Equation (6.44), the DIC based on this deviance is a posterior expected DIC where the expectation is taken with respect to the marginal posterior densities of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

6.6 Applications of MLIRT Modeling

The applications in Sections 6.6.1, 6.6.2, and 6.6.3 focus on school effectiveness research with an interest in the development of the knowledge and skill of individual students in relation to school characteristics. Data are analyzed at the individual level, and it is assumed that classrooms, schools, and experimental interventions have an effect on all students exposed to them. In school or teacher effectiveness research, both levels of the structural multilevel model are of importance because the objects of interest are schools and teachers as well as students. There is interest in the effect on student learning of the organizational structure of the school, characteristics of a teacher, and characteristics of the student. In Section 6.6.3, the MLIRT model is extended by incorporating latent explanatory variables. The application is focused on estimating the effect of educational leadership (school level) on math achievement (student level) after adjusting for student characteristics. In the application in Section 6.6.4, longitudinal item response data are analyzed where responses are nested within measurement occasions that are nested within subjects. A growth mixture modeling approach will be pursued to estimate individual trajectories of cognitive impairment for patients and controls.

6.6.1 Dutch Primary School Mathematics Test

In Section 5.2.4, test data for students leaving primary school were analyzed with a two-parameter item response model that ignored the nesting of the students in schools and the background information of students and schools. The 2,156 grade eight students are unequally spread over 97 schools and responded

to dichotomously scored mathematics items taken from the examination leaving school developed by Cito (Netherlands national institute for educational measurement). The 97 schools were fairly representative of all Dutch primary schools. Of the 97 schools sampled, 72 schools regularly participated in the examination leaving primary school. These are denoted as Cito schools, and the remaining 25 schools are denoted as the non-Cito schools. A school-level indicator variable (End) equaled one if the school participated in the test leaving primary school and zero if this was not the case.

Several student-intake characteristics were measured: socioeconomic status (SES), standardized scores on a nonverbal intelligence test (ISI-NV), and gender. The standardized SES scores were based on four indicators: the education and occupation level of both parents (if present). Gender was coded as zero for males and one for females.

The data were analyzed with an MLIRT model consisting of a normal ogive item response model at level 1 for measuring students' math abilities. A structural multilevel model defined the level-2 (student-level) and level-3 (school-level) parts. The MLIRT analysis was compared with a linear multilevel analysis using observed sum scores, the sum of the number of correct answers, as a measurement for the math abilities.

Interest was focused on (1) the effect of schools' participation in the examination leaving primary school on students' math abilities after adjusting for individual differences and (2) differences in parameter estimates when using sum scores compared with using an item response model for measuring math achievement. Therefore, the structural multilevel model $\mathcal{M}_1$

$$\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_{1j} ISI_{ij} + \beta_{2j} SES_{ij} + \beta_{3j} Female_{ij} + e_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} End_j + u_{0j}, \\
\beta_{1j} &= \gamma_{10}, \\
\beta_{2j} &= \gamma_{20}, \\
\beta_{3j} &= \gamma_{30},
\end{aligned}$$

was considered, where $e_{ij} \sim \mathcal{N}\left(0, \sigma_\theta^2\right)$ and $u_{0j} \sim \mathcal{N}\left(0, \tau_{00}^2\right)$.

In Table 6.1, the estimates of the parameters issued from MCMC scheme 4 are given. Under the heading Linear Multilevel Model, parameter estimates of an empty multilevel model and multilevel model $\mathcal{M}_1$ are given using standardized observed sum scores as the outcome variable. Under the heading MLIRT Model, parameter estimates of MLIRT models are given with the same structural multilevel part as the linear multilevel models. The MLIRT models were identified by fixing the mean and variance of the vector of sampled math abilities to zero and one, respectively, in each MCMC iteration. This way, all model parameter estimates are directly comparable. The reported standard deviations and HPD regions are the posterior standard deviations and the 95% highest posterior density intervals, respectively.

The grouping of students in schools explained a lot of the variance in math ability. A higher proportion of the variance is explained at the school

level according to the MLIRT analysis in comparison with the linear multilevel analysis. It followed that 28% of the individual variance according to the MLIRT analysis and 21% according to the linear multilevel analysis is explained at the school level. The estimated posterior standard deviations are larger in the MLIRT analysis since the measurement error in the dependent variable is taken into account. The posterior standard deviations from the MLIRT analysis provide a more reliable basis for testing, for example, the importance of predictor variables, since the inaccuracy of the estimated math achievements is taken into account.

From the parameter estimates of model $\mathcal{M}_1$ it follows that conditioned on SES, ISI, and Female, Cito schools performed better than non-Cito schools. That is, a significant positive effect on the students' math abilities is found in schools that participate on a regular basis in the central test leaving primary school. The general mean ability, γ_{00}, of the students attending non-Cito schools is significantly different from zero, and the positive significant value of γ_{01} indicates a positive effect on the students' math abilities due to the school's regular participation in the central exam leaving school. The positive effects of the predictors ISI and SES indicate that a student with a higher score (socioeconomic status or nonverbal intelligence test) performed better on the math test than a student with a lower score. The effect of Female is also significant and negative, meaning that boys outperformed girls on the math test.

The estimated predictor effects of the MLIRT analysis are higher in absolute value than the estimated effects of the linear multilevel analysis. Therefore, differences between students' math abilities and the explanatory effects of individual and school characteristics both become more apparent. The measurement of math abilities based on the two-parameter item response model resulted in a sharper distinction between students' outcomes than the observed sum scores. This is caused by the fact that the response patterns contain more information about the students' math abilities than the sum scores. The sum score is often used for unidimensional ability estimation, and under the Rasch model the sum score is statistically sufficient in estimating the true ability (Lord, 1980). However, the sum scores are biased estimates of ability when the items have different discrimination values. Errors in the sum scores are ignored, and they attenuate the effects of the predictors towards zero. The variance explained due to grouping is considerably lower in the linear multilevel analysis. The proportion of variance explained by the explanatory variables in the MLIRT analysis is 38% at the student level and 39% at the school level and in the linear multilevel analysis 29% and 46%, respectively.

For each model, the log-likelihood of the multilevel model part is computed given the outcome variable (for the linear multilevel models) or by integrating over the density of the latent outcome variable (for the MLIRT models). The DICs of the linear multilevel models and the MLIRT models show that in both cases model $\mathcal{M}_1$ is preferred. The dependent variable of the linear multilevel model is different from that of the MLIRT model, but they are scaled the same

way and are both considered estimates of the true ability values. However, the standardizing factors, the marginal posterior distributions $p(\boldsymbol{\theta})$ and $p(\overline{\mathbf{y}})$ (where $\overline{\mathbf{y}}$ are the observed sum scores), were excluded from the deviance terms, which complicates a direct comparison of the linear multilevel and MLIRT models using the estimated DICs. The reduction in DIC values is independent of the standardizing factor.

It can be seen that this reduction of 1,048 from the MLIRT analyses is much larger than the reduction of 736 from the linear multilevel analyses. This indicates that there is more statistical evidence to prefer MLIRT model $\mathcal{M}_1$ over the empty MLIRT model in comparison with preferring linear multilevel model $\mathcal{M}_1$ over the empty linear multilevel model. This conclusion is supported by the fact that the (fixed) effects (in absolute value) of the student- and the school-level variables are higher. The estimated effective number of parameters is smaller for the linear multilevel models. This indicates that the random intercepts have shrunk more towards the prior mean and schools appear to be more similar in their effects on student achievement in comparison with the results from the MLIRT analyses.

6.6.2 PISA 2003: Dutch Math Data

The Programme for International Student Assessment (PISA) launched by the Organisation for Economic Co-operation and Development (OECD) is conducted to assess student performance and collect data on student and institutional factors that can explain differences in performance. The PISA 2003 results can be found in OECD (Organisation for Economic Co-operation and Development) (2004), and the PISA 2003 data can be found at `http://pisa2003.acer.edu.au/downloads.php` (February 2010).

In 2003, 41 countries participated, and the survey covered mathematics (the main focus in 2003), reading, science, and problem solving. Attention is focused on the mathematics (literacy) abilities of 15-year-old Dutch students. Student performance in mathematics is measured via 85 items. Students were given credit for each item that they answered with an acceptable response. In this example, the item responses were coded as zero (incorrect) or one (correct).

In PISA 2003, each student was given a test booklet with clusters of items. Each mathematics item appeared in the same number of test booklets. This (linked) incomplete design makes it possible to construct a scale of mathematical performance using item response theory where each student has a score on this scale representing his or her estimated ability. Variations in student ability within the Netherlands are investigated using various background variables. A total of 3,829 students across 150 schools were questioned.

Multiple Imputation

The multiple imputation technique (Rubin, 1987, Little and Rubin, 2002) can be used to handle the uncertainty regarding the ability parameter estimates.

Table 6.1. Parameter estimates of linear multilevel and MLIRT models using explanatory information at the student and school levels.

	Linear Multilevel Model					MLIRT Model				
	Empty Model		Model $\mathcal{M}_1$			Empty Model		Model $\mathcal{M}_1$		
	Mean	SD	Mean	SD	HPD	Mean	SD	Mean	SD	HPD
Fixed Effects										
γ_{00} Intercept	−.058	.052	−.312	.080	[−.474, −.164]	−.064	.060	−.326	.096	[−.525, −.146]
Student variables										
γ_{10} ISI-NV			.412	.017	[.377, .446]			.459	.020	[.420, .496]
γ_{20} SES			.217	.019	[.180, .254]			.248	.021	[.205, .290]
γ_{30} Female			−.160	.035	[−.226, −.095]			−.181	.038	[−.257, −.107]
School variables										
γ_{01} End			.464	.090	[.285, .641]			.489	.108	[.276, .699]
Random Effects										
Within schools										
σ^2_θ Residual variance	.813	.029	.578	.019	[.542, .612]	.766	.026	.472	.019	[.436, .510]
Between schools										
τ^2_{00} Intercept	.216	.039	.117	.022	[.076, .162]	.299	.051	.183	.031	[.124, .244]
Information Criteria										
−2 log-likelihood	5587.94		4852.30			6027.29		4978.83		
DIC (p_D)	5752.15(82.11)		5016.28(81.99)			6122.08(94.79)		5074.47(95.65)		

When using plausible values, the standard errors of the ability estimates can be taken into account when estimating the other multilevel model parameters.

In PISA 2003, the ability parameters were replaced by simulated values. The so-called plausible values were drawn randomly from the posterior distribution of the ability parameters given the response patterns and explanatory information. The posterior density from which plausible values were sampled is given by

$$p\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma_\theta^2\right) \propto \int p\left(\mathbf{y} \mid \boldsymbol{\theta}; \boldsymbol{\xi}\right) p\left(\boldsymbol{\theta}; \boldsymbol{\beta}, \sigma_\theta^2\right) \mathrm{d}\boldsymbol{\theta}, \tag{6.31}$$

where the item parameters and parameters of the population's ability distribution are known. A subsample of respondents from different countries, referred to as the international calibration sample, were used to estimate the international item parameters. The estimated international item parameter values were assumed to hold in each country. The explanatory variables in the posterior distribution of abilities are known as conditioning variables, and five variables were used: booklet number, gender, mother's occupation, father's occupation, and school mean mathematics score.

For each student, five plausible values for the mathematical literacy domain were drawn randomly from the posterior density in Equation (6.31). This posterior density does not depend on the missing-data indicator (in this particular case, all the ability values are missing), which indicates that the multiple imputations are appropriate under the assumption of ignorability. Together with the fact that the multiple imputations are independent samples, Schafer (1997) characterized such plausible values as Bayesianly proper based on Rubin's (1987) definition for proper multiple imputations.

Once the plausible values are obtained, the complete data can be analyzed in a second phase. An important advantage of multiple imputation inference is the temporal separation of handling the missing-data problem and analyzing the complete data. The missing-data problem is confined entirely to the first phase. One good set of plausible values can solve the missing-data problem for many future analyses. On the other hand, the two phases, imputation and analysis, are distinct, and this may lead to inconsistencies when the imputation model and the model for analysis are based on different assumptions. When the model for analysis assumes more than the imputation model, then correct inferences are made when the additional assumption of the analyst's model is true. The inferences derived from the imputed data will be valid, but the plausible values may contain additional uncertainty since the imputation model is more general than the model for analysis. When the imputation model is more restrictive, valid inferences are made when the additional restriction is true. In the case where the additional assumption of the imputation model is false, incorrect inferences are made. Plausible values that are drawn from an erroneous model can lead to erroneous conclusions.

The imputation model in Equation 6.31 has some potential weaknesses (see, e.g., Goldstein, Bonnet and Rocher, 2007). The item parameters and

population parameters are assumed to be known, and therefore the plausible values drawn do not reflect any uncertainty with respect to the corresponding estimated values. Estimated international item parameters instead of nation-specific item parameters were used in the imputation model. The difference is that the international item parameters are based on the restrictive assumption that all items function equally across nations, whereas nation-specific item parameters are assumed to function equally within that nation. It is not likely that all international item characteristics apply to each individual in the cross-national PISA survey. It may be possible that items function differently across groups when, for example, cross-national and/or cross-cultural response heterogeneity is present. A more realistic assumption is to assume that within each nation some of the item characteristics may deviate from the international item characteristics. This topic will be further discussed in Chapter 7. Finally, the abilities were assumed to be independent given the conditioning variables. However, the students were nested in schools, and the performances from students of the same school are likely to be correlated since they share common experiences and have the same teachers and teaching programs. It is likely that, given the conditioning variables, the drawn students' plausible values are still correlated.

The reported plausible values from the PISA study were used to construct five complete-data sets that were analyzed with a linear multilevel model. The posterior moments of the model parameters can be estimated from a relatively small number of draws since the complete-data posterior is based on multivariate normality given that the fraction of missing data is not too large (Little and Rubin, 2002). The complete-data inferences are combined to estimate the structural multilevel model parameters. That is, for each parameter of interest, the posterior mean and variance are estimated from the complete-data posterior distribution. The multiple imputations are used to average over the missing ability values and to approximate the posterior mean and variance of the marginal posterior distribution. This way, the posterior mean and variance of the fixed effects parameters $\boldsymbol{\gamma}$ given $M = 5$ plausible vectors $\boldsymbol{\theta}^{(m)}$ $(m = 1, \ldots, 5)$ are estimated by

$$\begin{aligned} E\left(\boldsymbol{\gamma} \mid \mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2\right) &= E\left(E\left(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}\right) \mid \mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2\right) \\ &= \int\int \boldsymbol{\gamma} p\left(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2\right) \mathrm{d}\boldsymbol{\gamma}\mathrm{d}\boldsymbol{\theta} \\ &\approx M^{-1} \sum_m \int \boldsymbol{\gamma} p\left(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}\right) \mathrm{d}\boldsymbol{\gamma} \\ &\approx M^{-1} \sum_m E\left(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}\right) = \overline{\boldsymbol{\gamma}} \end{aligned}$$

and, given $\left(\boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2\right)$,

$$\begin{aligned}\mathrm{Var}\,(\boldsymbol{\gamma} \mid \mathbf{y}) &= E\,(\mathrm{Var}\,(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y}) + \mathrm{Var}\,(E\,(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y}) \\ &\approx M^{-1} \sum_m \mathrm{Var}\left(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}\right) + \frac{1}{M-1} \sum_m (\hat{\gamma}_m - \overline{\gamma})\,(\hat{\gamma}_m - \overline{\gamma})^t \\ &\approx \overline{V}_\gamma + B,\end{aligned}$$

where $\hat{\gamma}_m = E\left(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}\right)$, $\overline{V}_\gamma$ denotes within-imputation variance (the average of the complete-data variance estimates), and B denotes the between-imputation variance (the variance of the complete-data point estimates). When M is small, an improved approximation of the posterior variance is obtained by $\overline{V}_\gamma + (1 + M^{-1})B$. The fraction of missing information is estimated by the ratio of estimated between-imputation variance to total variance; that is,

$$\frac{(1 + M^{-1})B}{\overline{V}_\gamma + (1 + M^{-1})B}.$$

The MLIRT model was used to analyze the Dutch item response data (which were also used for generating the plausible values) given background information at the student and school levels. The MLIRT model consisted of a two-parameter item response model and a structural multilevel model for the ability parameters. In the MLIRT model analysis, nation-specific item parameters were estimated simultaneously with the other MLIRT model parameters using MCMC scheme 4. In contrast to the multiple-imputation method, the uncertainty of the item parameter estimates is taken into account in estimating the other model parameters. For each vector of plausible values, the linear multilevel parameters were also estimated with MCMC using the structural multilevel part of scheme 4.

In Table 6.2, the empty structural multilevel parameter estimates are presented under the MLIRT model and the linear multilevel model. The vectors of plausible values were standardized and the MLIRT model was identified by standardizing the ability scale.[2] This makes the results under both models directly comparable. It follows that the parameter estimates and standard deviations of both empty model analyses are quite similar. The plausible values generated contain the uncertainty of the ability estimates, and valid inferences can be made from the linear empty multilevel analysis using multiple imputations. The estimated intraclass correlation coefficient is around 59%, which is the proportion of variation in ability estimates explained by the grouping of students in schools. This proportion is high and above the OECD average. Goldstein (2004) also found a huge difference in explained variation by school across countries, with many countries at 50% or more. It is not clear why such high intraclass correlation coefficients are found for the PISA data.

[2] In PISA 2003, the Dutch overall performance in mathematics was measured on a scale with mean 542 and standard deviation of around 92, whereas the metric for all participating countries had a mean of 500 and standard deviation of 100.

To investigate differences in performance between schools and the effects of student-level and school-level factors on a student's ability, several background characteristics were incorporated in the multilevel model. According to the PISA 2003 study, the following student characteristics explained variation in performance: gender, place of birth (Netherlands or foreign), language (Dutch or speaks a foreign language most of the time), and index of economic, social, and cultural status. The school's mean index of economic, social and cultural status is used as an explanatory variable for the random intercept. In Table 6.2, the results of the linear multilevel analysis using plausible values and the MLIRT analysis are reported under model $\mathcal{M}_1$.

It can be seen that the estimated standard deviations are similar. The estimated (level-3) effect, γ_{01}, is slightly larger under the MLIRT model. The plausible values were generated without taking the nested structure of the data into account, which might explain this underestimated school effect. In the MLIRT analysis, the ability parameters are re-estimated with a multilevel part that includes covariates that also accounts for the nested structure of the observed data. The posterior mean estimates of the (level-2) fixed effects are comparable. Under the linear multilevel analysis, the explanatory variables explain 6.9% of the student-level variance and 18.6% of the school-level variance, and under the MLIRT analysis, 6.4% of the student-level variance and 19.6% of the school-level variance.

From the MLIRT analysis it follows that the male students perform slightly better than the females. The native speakers also perform better than non-native speakers with a migrant background, taking account of socioeconomic differences between students and between schools. It can be concluded that students from more advantaged socioeconomic backgrounds generally perform better.

Only the difference in DICs for the MLIRT and the linear multilevel models can be compared since the deviance term does not contain the standardizing factor. The DICs are based on the log-likelihood of the multilevel model. The differences in DIC values are comparable: 279.01 for the MLIRT analysis and 282.15 for the linear multilevel analysis. This indicates that both modeling techniques lead to a comparable amount of statistical evidence for selecting model $\mathcal{M}_1$ over the empty model. The log-likelihood and DIC estimates under the linear multilevel model are the averaged estimates since single estimates were obtained for each vector of plausible values. In the MLIRT analysis, the ability parameter is integrated out using its marginal posterior distribution, and in the linear multilevel analysis the integration is approximated by taking the average over the outcomes based on five vectors of plausible values. This is a rough approximation and leads to lower DIC estimates and higher estimates of the number of effective parameters. In this case, more draws of plausible values could drastically improve the estimates. Note that under the linear multilevel analysis, the averaged log-likelihood and DIC estimates were based on the same set of plausible values, so these differences in DIC values remain interesting.

Table 6.2. PISA (Dutch) 2003: Parameter estimates of the linear multilevel model with multiple imputations and the MLIRT model using explanatory information at the student and school levels.

	Linear Multilevel Model					MLIRT Model				
	Empty Model		Model $\mathcal{M}_1$			Empty Model		Model $\mathcal{M}_1$		
	Mean	SD	Mean	SD	HPD	Mean	SD	Mean	SD	HPD
Fixed Effects										
γ_{00} Intercept	−.039	.065	.022	.063	[−.103, .142]	−.037	.065	.017	.062	[−.103, , 139]
Student variables										
γ_{10} Student is female			−.166	.024	[−.208 − .126]			−.163	.026	[−.216, −.112]
γ_{20} Student is foreign born			−.281	.050	[−.371, −.196]			−.275	.055	[−.384, −.168]
γ_{30} Student speaks foreign language most of the time			−.226	.047	[−.317 − .137]			−.226	.058	[−.342, −.116]
γ_{40} Index of economic, social, and cultural status			.143	.015	[.115, .169]			.152	.017	[.119, .184]
School variables										
γ_{01} Mean index of economic, social, and cultural status			.359	.156	[.048, .654]			.391	.156	[.084, .701]
Random Effects										
Within schools										
σ^2_θ Residual variance	.418	.019	.389	.013	[.371, .407]	.425	.013	.398	.012	[.374, .422]
Between schools										
τ^2_{00} Intercept	.609	.075	.496	.061	[.384, .620]	.606	.074	.487	.061	[.375, .610]
Information Criteria										
−2 log-likelihood		8337.44			8045.33		8451.60			8171.71
DIC (p_D)	8643.02(152.79)				8360.87(157.77)	8659.83(104.12)				8380.82(104.55)

6.6.3 School Effects in the West Bank: Covariate Error

Up to now, the MLIRT models presented have defined a nonlinear relationship between response variables and explanatory variables, where the explanatory variables were assumed to be measured without an error. However, often it is not possible to measure all relevant explanatory variables accurately. Assessing the measurement errors is important since errors in explanatory variables can bias the association between the response and explanatory variables. Measurement errors in one covariate can also bias the link between a response variable and other covariates that are measured without an error (Fuller, 1991). A comprehensive discussion of linear and nonlinear measurement error models can be found in Fuller (1987) and Carroll, Ruppert and Stefanski (1995), respectively.

Assume that a true or exact predictor is a latent variable such that its realizations cannot be observed directly. Let the latent explanatory variable be measured by items whose observations can be observed. In the measurement error literature, this operationally defined true explanatory variable is often referred to as a gold standard (Carroll et al., 1995, p. 11), which refers to the best way of measuring the true latent variable in practice. In this application, educational leadership is considered to be a latent explanatory variable at the school level. Scheerens, Glas and Thomas (2003) defined educational leadership as one of the five process indicators of school functioning, and its impact on student achievement is often investigated in school effectiveness research. The MLIRT model is extended to deal with the error in the measurement of educational leadership, which is commonly ignored in school effectiveness research, and to demonstrate the use of latent explanatory variables.

School Leadership and Math Achievement

The effect of school leadership on student math achievement was investigated using data from a school effectiveness study in the West Bank (Shalabi, 2002). A stratified sample of 119 schools ensured that all school types and all geographical districts of the West Bank were represented. The average number of students per class was 28, with a minimum of 10 and a maximum of 46 students. A total of 3,384 grade seven students were randomly selected from the selected schools. The math achievement of grade seven students was measured using a math test consisting of 50 dichotomously scored items (correct/incorrect). A test of 25 five-point Likert items was taken by teachers and school principals to measure educational leadership. The first three response categories were collapsed since a minimal response was observed in the two lowest categories.

A structural multilevel model was used to define a relationship between educational leadership ζ and math achievement θ and to account for the fact that students were nested in schools. The MLIRT model consists of two measurement models: a two-parameter normal ogive model for measuring math

achievement using the students' item responses and a graded response model for measuring educational leadership using teachers' item responses. Let $\mathbf{Z}$ denote the latent responses to the math items (Equation (4.7)) and $\mathbf{V}$ the latent responses to the educational leadership items (Equation (4.25)). In the latent response formulation, the MLIRT model can be stated as

$$\left.\begin{array}{ll} Z_{ijk} = a_k\theta_{ij} - b_k + \epsilon_{ijk} & \text{(student level)} \\ V_{jk} \;\; = \lambda_k\zeta_j + \nu_{jk} & \text{(school level)} \end{array}\right\} \text{Measurement part}$$

$$\left.\begin{array}{ll} \theta_{ij} \;\; = \beta_{0j} + e_{ij} & \text{(student level)} \\ \beta_{0j} = \gamma_{00} + \gamma_{01}\zeta_j + u_{0j} & \text{(school level)} \end{array}\right\} \text{Structural part}$$

where the error terms in the measurement part are independently standard normally distributed and the error terms in the structural multilevel part are independently normally distributed with mean zero and variance σ_θ^2 and τ^2, respectively. For the two-parameter response model, a truncated multivariate normal prior was used for the discrimination and difficulty parameters (Chapter 2). For the graded response model, the threshold parameters were assumed to be order-restricted and uniformly distributed, and the discrimination parameters were assumed to have a common positively truncated normal prior.

The MLIRT model can be identified by fixing the scale of each latent factor. The scale of the latent factors math and educational leadership were identified by restricting the mean to zero and the variance to one. Note that this MLIRT model allows for mixed response types (binary responses to the math items and ordinal responses to the educational leadership items). Fox and Glas (2003) defined an MCMC scheme for simultaneously estimating all parameters. This MCMC scheme is an extension of scheme 4, and it includes the sampling of latent explanatory variable values.

Table 6.3 gives the structural model parameter estimates of an empty MLIRT model and an MLIRT model, denoted as $\mathcal{M}_1$, with level-2 variable female (male=0, female=1) and level-3 latent variable educational leadership. It follows that about 49% of the variance in student achievement is explained by school differences. The between-school variance is relatively high for a developing country. In developing countries, around 40% of the variance in student achievement is accounted for by schools (Scheerens et al., 2003).

From the estimates of model $\mathcal{M}_1$ it follows that female students performed better than male students. There are only 24 schools with male and female students (48 schools have only male students and 47 schools have only female students). As a result, the level-1 predictor female only reduced the between-school variation and not the within-school variation.

The factor educational leadership has a significant positive effect on student math achievement after adjusting for gender differences. This is not surprising given that the defined contents of the educational leadership concept were among other things setting periodic exams, identifying talented students, and making sure that students acquire basic skills.

Table 6.3. School effectiveness in the West Bank: Parameter estimates of the MLIRT model with latent explanatory variable educational leadership.

	Empty Model		Model $\mathcal{M}_1$		
	Mean	SD	Mean	SD	HPD
Fixed Effects					
γ_{00} Intercept	.003	.066	−.079	.072	[−.225, .058]
Student variable					
γ_{10} Female			.169	.065	[.040, .292]
School variable					
γ_{01} Educational leadership			.167	.066	[.036, .294]
Random Effects					
Within schools					
σ^2_θ Residual variance	.515	.014	.514	.015	[.487, .544]
Between schools					
τ^2_{00} Intercept	.500	.069	.456	.062	[.344, .581]
Information Criteria					
−2 log-likelihood		14570.11			14329.75

6.6.4 MMSE: Individual Trajectories of Cognitive Impairment

The Mini-Mental State Examination (MMSE) is a widely used screening instrument for the detection of cognitive impairment. The test was first described by Folstein, Folstein and McHugh (1975) as a method for grading the cognitive state. Originally it was used as a brief objective assessment of five cognitive concepts (orientation, attention, registration, recall, and language). Mood or thought disorders are not assessed, and therefore it was called mini. The benefits of the MMSE are its brevity and that it covers many different domains. In general, the MMSE is a brief screening test that quantitatively assesses the level of cognitive impairment and for repeated measurements assesses cognitive changes occurring over time.

The MMSE has gained increasing popularity for measuring neurobehavioral deficits and the cognitive abilities of elderly patients. Folstein et al. (1975) suggested the presence of dementia in persons with scores of less than 23 out of 30 and at least eight years of education. Depending on the sample, a ceiling effect might be present when a patient scores perfect who is often well-educated but meets criteria for dementia. Tombaugh (1992) reviewed the psychometric properties and utilities of the MMSE. The main outcome was that the MMSE possessed moderate to high reliability coefficients, was shown to be sensitive to cognitive impairments in persons suffering from Alzheimer's disease, and reflected cognitive decline, which is typical for dementia patients. On the other hand, criticisms included (1) that the MMSE showed limited

ability to discriminate people not demented from people with mild dementia, (2) language items were too easy to identify mild language deficits, and (3) the MMSE scores were affected by age, education, and cultural background but not gender. Schulz-Larsen, Kreiner and Lomholt (2007a, 2007b) explored the properties of the MMSE using a mixture Rasch model to detect item characteristic differences across groups (Rost, 1990; Rost and von Davier, 1995). The MMSE items were expected to act differently in the two groups of elderly with and without cognitive impairments.

MMSE data from a study of the OPTIMA[3] cohort (Oxford Project to Investigate Memory and Ageing) are analyzed. The data cover 668 participants, who were questioned on different measurement occasions. On several occasions, 26 discrete MMSE item responses (scores for item 15 have been dichotomized) and background information were observed. Let observation Y_{ijk} denote the response to item k of subject i on measurement occasion j. For each occasion j, observed MMSE responses of subject i are used to estimate the cognitive impairment θ_{ij} using a two-parameter item response model. Low scores are assumed to correspond with severe cognitive impairment.

The participants in the naturalistic follow-up study can be classified as patients (who are suffering from dementia) and controls, where some convert from controls to patients. Most of the patients show increasing cognitive impairment over time, which is characteristic of dementia and is seen as a risk factor for Alzheimer's disease. In Table 6.4, the demographic information of the study participants is given, including a distribution of the averaged sum scores across measurement occasions. It can be seen that 302 participants score on average at least 24 items correct and do not suffer from cognitive impairment. Therefore, the distribution of sum scores is likely to be a mixture of a patient's and a control's sum score distributions.

Figure 6.3 presents a so-called spaghetti plot of some participants' observed sum scores. For each participant selected, the observed sum scores are plotted and connected by the follow-up time in years. The baseline time point zero corresponds to the first time a participant is confronted with the MMSE. The darker lines represent participants that show an increase in cognitive impairments and the lighter dotted lines represent participants who show no variations in cognitive impairment over time. The plot suggests that there is a general decline in sum scores for a group of patients besides the considerable individual heterogeneity and no decline for a group of controls.

This illustrates that the classification of subjects as patients or controls is important since subjects in one group do not suffer from cognitive impairment, whereas subjects from the other group suffer from mild or severe cognitive impairment. Class membership may have different consequences and antecedents in different classes, which makes it important to distinguish among subjects in the different groups. It is to be expected that the average trajectories of cog-

[3] Data were supported by the NIHR Oxford Comprehensive Biomedical Research Centre.

Table 6.4. Demographic information for the study participants.

	Participants ($N = 668$)	
Gender		
Male	329	
Female	339	
Age	**Start**	**Mean**
50–59	55	41
60–69	195	149
70–79	323	315
80–89	149	215
90–100	9	14
Average Sum Score		
24–26	302	
22–23	66	
20–21	47	
18–19	61	
15–17	66	
< 14	126	

nitive impairment in the two groups are different and that there are different individual variations around the average trajectories.

Mixture MLIRT Modeling

The response data (level 1) are nested within measurement occasions (level 2), which are nested within subjects (level 3). An item response model is defined at level 1 for the measurement of occasion-specific impaired cognitive functioning. At level 2, the measured cognitive impairments are nested within subjects. This level accounts for the longitudinal character of the observed data. At level 3, heterogeneity among subjects is captured via random effects. A (three-level) MLIRT model is defined to account for the nested structure of the data. The MLIRT model can handle the varying number of measurements across participants, and it can handle complicated factors such as measurements not measured on the same time points and measurements not uniformly distributed across subjects.

To identify a group of patients and a group of controls, to model dependencies between the response outcomes, and to allow for different individual trajectories of cognitive impairment across the groups, a latent class model (e.g., Goodman, 1974; Lazarsfeld and Henry, 1968) or mixture model (e.g., McLachlan and Peel, 2000) is defined. Therefore, the structural part of the MLIRT model equals a two-component mixture model. That is, the population distribution of cognitive impairments is defined to be a mixture of two

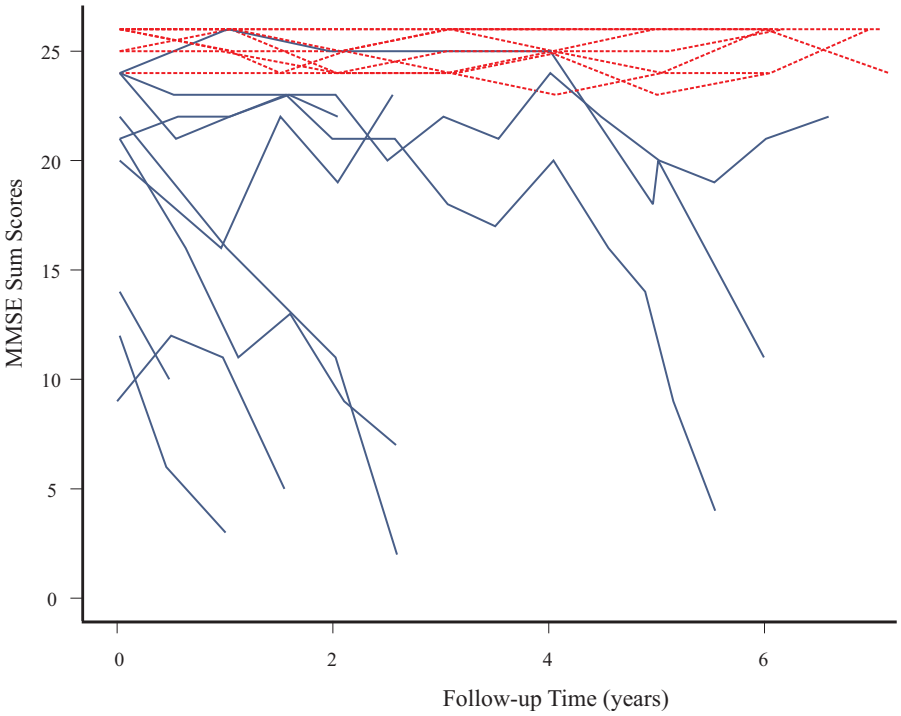

Fig. 6.3. Spaghetti plot of some participants' observed MMSE sum scores.

components presenting the population distributions of cognitive impairments of patients and controls.

Let the distribution of cognitive impairments in a population of elderly reflect the mixture of patients and controls in the population and be given by

$$p\left(\theta_{ij} \mid \boldsymbol{\Omega}\right) = \sum_{g=1}^{2} \pi_{ig} p\left(\theta_{ij} \mid \boldsymbol{\Omega}_g\right),$$

where $\boldsymbol{\Omega}_g$ are the structural multilevel parameters of group $g = 1, 2$.

Let indicator or class membership variable G_i indicate if subject i suffers from cognitive impairment ($G_i = 1$) and is labeled patient or does not ($G_i = 2$) and is labeled control. Each indicator variable G_i is distributed unconditionally as Bernoulli with success probability π_{i1}, and a conjugate beta prior is specified for the success probability. Then, the conditional posterior probability that subject i is classified as a patient equals

$$P\left(G_i = 1 \mid \mathbf{y}_i, \boldsymbol{\theta}_i, \boldsymbol{\Omega}_1, \boldsymbol{\pi}_i\right) = \frac{\pi_{i1} \prod_{j=1}^{n_i} p\left(\mathbf{y}_{ij} \mid \theta_{ij}\right) p\left(\theta_{ij} \mid \boldsymbol{\Omega}_1\right)}{\sum_{g=1,2} \pi_{ig} \prod_{j=1}^{n_i} p\left(\mathbf{y}_{ij} \mid \theta_{ij}\right) p\left(\theta_{ij} \mid \boldsymbol{\Omega}_g\right)}.$$

The introduction of the class membership variables G_i supports the computation of the posterior probability that a participant is classified as a patient

given item response data and the computation of the proportion of participants with cognitive impairment.

Diebolt and Robert (1994) introduced a Gibbs sampling algorithm for the estimation of normal mixtures with a prespecified number of mixture components. In their MCMC implementation, class membership variables G_i are considered to be missing data and in each MCMC iteration samples for the missing data are generated. The other model parameters, including the mixing proportions, are sampled given a realization of class memberships. MCMC scheme 4 can be extended with a step for sampling class memberships and a step for sampling the mixing proportions. The other sampling steps remain the same but may depend on the class membership variables.

Identifiability

The items are assumed to function in the same way for controls and patients. It is also assumed that the items function in the same way across measurement occasions. That is, the MMSE items are assumed to be time-invariant and invariant across the two latent classes. The invariance assumptions are needed to ensure that the estimated occasion-specific cognitive impairments are measured on a common scale across latent classes. Given class memberships, the latent scale is identified by fixing its mean and variance.

A finite mixture model is not identified since the distribution of the data is unchanged if the class membership labels are permuted. In the two-component mixture, the labels for patients and controls can be switched in an MCMC iteration, which leads to a nonidentified solution. To avoid any ambiguity, the mixture components can be identified by fixing the means of the components to be in nondecreasing order. This way, the mean cognitive impairment of the patients is restricted to be smaller than the mean cognitive impairment of the controls. It is also possible to identify the mixture model by specifying an order restriction on the mixing proportions. The mixture MLIRT model is identified by fixing the mean and variance of the latent scale and specifying an order restriction on the mixture component means.

OPTIMA Cohort: Modeling Individual Trajectories

Interest will be focused on the part that models the individual trajectories of cognitive impairment. First, consider a mixture MLIRT model $\mathcal{M}_0$ where the (occasion-specific) cognitive impairments are nested within subjects that are measured via a two-parameter item response model. The within-subject residual variation and the between-subject variation in cognitive impairment are assumed to be common across classes. This leads to the two-component mixture density of cognitive impairments

$$p\left(\theta_{ij} \mid \gamma_{00}, \tau_{00}^2, \sigma^2\right) = \pi_{i1}\phi\left(\gamma_{00,1} + u_{i0}, \sigma^2\right) + \pi_{i2}\phi\left(\gamma_{00,2} + u_{i0}, \sigma^2\right),$$

where $u_{i0} \sim \mathcal{N}\left(0, \tau_{00}^2\right)$. This structural mixture part enables the computation of subject-specific class probabilities and average class means of cognitive impairment.

Table 6.5 gives the parameter estimates of the two-component mixture part of $\mathcal{M}_0$. It can be seen that the mean cognitive impairment of the patients is much smaller than that of the controls. The within-subject residual variation is smaller than the between-subject variation in cognitive impairment, and around 63% of the variation in cognitive impairment is attributable to subjects.

In model $\mathcal{M}_0$, it is assumed that the between-subject variation is common across groups. However, more variation in cognitive impairment is to be expected in the patient group. The patients suffer from different degrees of cognitive impairment, whereas the controls display almost no decrease in cognitive functioning across measurement occasions, and less variation between subjects is to be expected. The follow-up times are included to model the rate of change in cognitive impairment across time. It is not likely that the rate of change is similar for all subjects. Therefore, the individual time trends are treated as random effects. The random effects of follow-up time on cognitive impairment are allowed to vary across subjects and groups. This means that the average effect of follow-up times (*Fup*) may differ across groups and that the individual variations across the averages may differ.

This is included in model $\mathcal{M}_1$, with a two-component mixture density of cognitive impairment given by

$$p\left(\theta_{ij} \mid \boldsymbol{\gamma}, \mathbf{T}, \sigma^2\right) = \pi_{i1}\phi\left(\mu_{ij,1}, \sigma^2\right) + \pi_{i2}\phi\left(\mu_{ij,2}, \sigma^2\right), \tag{6.32}$$

where

$$\begin{aligned}
\mu_{ij,g} &= \beta_{i0,g} + \beta_{i1,g} Fup_{ij}, \\
\beta_{i0,g} &= \gamma_{00,g} + u_{i0,g}, \\
\beta_{i1,g} &= \gamma_{10,g} + u_{i1,g},
\end{aligned}$$

and $\mathbf{u}_{i,g} \sim \mathcal{N}\left(0, \mathbf{T}_g\right)$, with $\mathbf{T}_g$ a diagonal matrix with elements $\tau_{00,g}^2$ and $\tau_{11,g}^2$ for $g = 1, 2$.

In Equation (6.32), a growth mixture model is presented where individual trajectories of cognitive impairment are modeled. The random intercept parameters define the starting point at the subject's first measurement occasion, and the random slope parameters define the degree of change in cognitive impairment over time. The fixed effects parameters define the average initial population level and average population trend.

In Table 6.5, the average intercept values differ, and it can be seen that the average score of the controls on the first measurement is much higher. The initial scores of cognitive impairment are comparable for the controls, but there is considerable heterogeneity in the initial scores of the patients. The average population slope is around zero for the controls. This means that their

Table 6.5. MMSE: Parameter estimates of mixture MLIRT model $\mathcal{M}_0$ and $\mathcal{M}_1$.

	Mixture MLIRT $\mathcal{M}_1$				Mixture MLIRT $\mathcal{M}_2$			
	Patients		Controls		Patients		Controls	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fixed Effects								
γ_{00} Intercept	−.998	.037	.689	.030	−.332	.037	.913	.012
Time variables								
γ_{10} Follow-up time					−.274	.013	−.007	.004
Random Effects								
Within individual								
σ^2_θ Residual variance	.133	.003	.133	.003	.043	.001	.043	.001
Between individual								
τ^2_{00} Intercept	.211	.015	.211	.015	.471	.038	.016	.003
τ^2_{11} Follow-up time					.047	.004	.002	.000

level of cognitive impairment remains constant across time. A negative average trend of −.274 was estimated for the patients. The estimated population variation for the slope equals .047. Thus, there is considerable heterogeneity in the patients' change in cognitive impairment across time.

In Figure 6.4, the estimated individual random effects are plotted for the subjects of both groups. The circles are estimates of participants (38% of the subjects) that are classified as controls with 95% reliability. The squares are the estimates of participants (53% of the subjects) that are classified as patients with 95% reliability. The triangles are the estimates of participants (9% of the subjects) that could not be classified with 95% reliability. In most cases, they are controls, with high initial values, who have converted.

It can be seen that there is almost no between-subject variation in initial scores (horizontal distance) and individual trends (vertical distance) for the controls. This group of subjects shows almost no decline over time. The patients show a lot of variation in initial scores and trends, where some have high initial scores but show severe cognitive impairment over time (lower right-hand corner). Patients with below average initial scores show mild to severe cognitive impairment over time. Patients with low scores on the first measurement occasion show severe cognitive impairment over time.

A few subjects (upper-middle part) scored below average on the first measurement occasion but improved their performance over time. In Figure 6.3, it can be seen that some subjects display an improvement in sum scores. These participants have positive individual trend effects and low starting values. They are classified as patients since their initial scores differ too much from the average initial level of the controls. Other subjects with initial scores

around zero showing a relatively small positive or negative improvement over time are also classified as patients. In a subsequent analysis, a third latent class could be introduced to capture subjects that show an improvement in their performance over time.

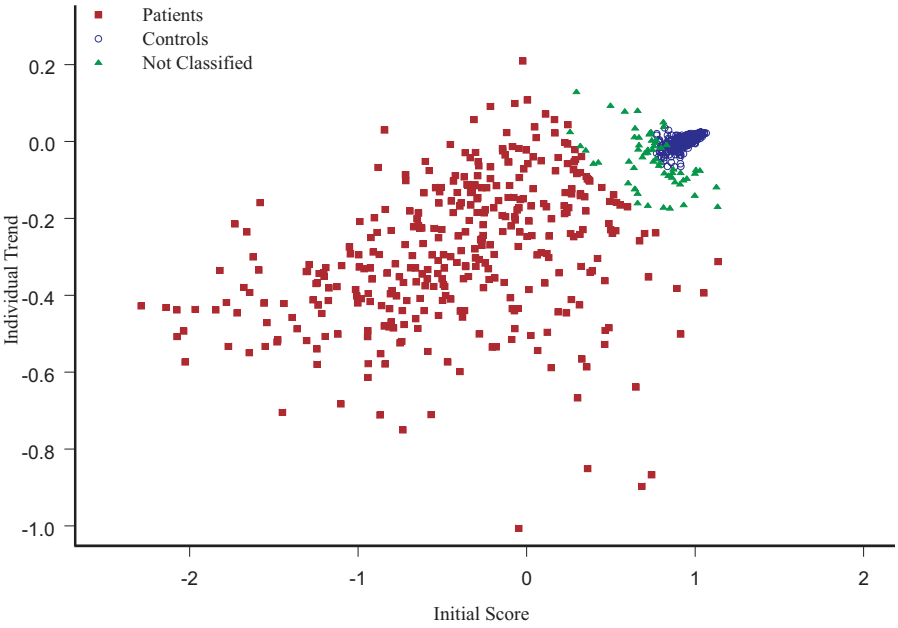

Fig. 6.4. Estimated random effects of initial levels and linear trends of cognitive impairment.

6.7 Summary and Further Reading

A grouping of respondents in clusters leads to dependencies between students due to the fact that a group of individuals share common interests, beliefs, and behaviors. The item response modeling framework is extended with a multilevel population distribution for the ability parameter to account for such a hierarchical data structure. The MLIRT model discussed can be viewed as a nonlinear mixed effects model with a linear multilevel model for the ability parameter that has a nonlinear relationship with the observed multivariate discrete response data. The class of generalized linear mixed effects models contains the Rasch version of the MLIRT model but with specific restrictions on the prior distributions.

Fox (2001, 2003) discussed MCMC algorithms for simultaneously estimating all MLIRT model parameters. Several applications of MLIRT modeling

are discussed in Fox (2004) and Fox and Glas (2003). More applications and details about the MLIRT software for the statistical package R can be found in Fox (2007). The assessment of MLIRT models is discussed in Fox (2005a).

The MLIRT model supports an improved estimate of a single ability using the fact that similar estimates exist for other individuals. This also applies for the estimation of effects within individual groups. In the educational data examples, separate equations were obtained for the males and females, where the estimated gender effect was based on the weighted information across schools. The estimation of a school-specific gender effect can be based on weighted information across schools and the information from that school. This feature of the multilevel modeling approach becomes particularly interesting when data are sparse for the males or females and when male or female students are not present in some schools (e.g., Braun, Jones, Rubin and Thayer, 1983). Besides the advantage of the borrowing strength principle, the MLIRT model allows the specification of cross-level interactions where variables at one level affect relationships at another level. In the educational examples, item discriminating effects can be evaluated for males and females.

Two other popular item response models also account for an additional nesting of item responses: the testlet model (Section 5.5.2) and the hierarchical rater model. The hierarchical rater model for multiple ratings of test items combines information from multiple raters to assess student performance. Patz, Junker, Johnson and Mariano (2002) developed a hierarchical rater model that accounts for the nested structure of the data where item responses are nested in examinees and raters (see also Verhelst and Verstralen, 2001). Conditional independence between raters' ratings is assumed by conditioning on an ideal rating, and conditional independence between ideal ratings is assumed by conditioning on the examinees' abilities.

There is extensive literature about multilevel models for continuous and discrete outcome data. A more thorough overview of multilevel modeling is given by Goldstein (2003), Hedeker and Gibbons (2006), Raudenbush and Bryk (2002), and Snijders and Bosker (1999), among others. A fully Bayesian multilevel analysis was first presented by Hill (1965) and Tiao and Tan (1965). A seminal contribution on Bayesian estimation of linear multilevel models was given by Lindley and Smith (1972). Other important work on Bayesian multilevel analysis includes Efron and Morris (1975), Dempster et al. (1981), and Gelfand and Smith (1990). Gelman and Hill (2007) demonstrate the use of WinBUGS and R for fitting multilevel models.

The MLIRT model can be extended with latent explanatory variables at different hierarchical levels as shown in Section 6.6.3. Latent explanatory variables function in the same way as manifest explanatory variables, and (cross-level) interaction effects can also be defined for latent covariates. An item response modeling framework is described for unobserved continuous covariates. A mixture modeling approach can be followed to handle discrete latent covariates (Kuha, 1997). More about latent variable modeling for handling measurement error in latent covariates can be found in, among others, Fox

and Glas (2003), Muthén (1992), Raudenbush and Bryk (2002), and Skrondal and Rabe-Hesketh (2004).

The mixture MLIRT model described in Section 6.6.4 contains a two-component mixture at level 3 for the grouping of subjects as patients or controls. The classification of multiple item observations within a measurement occasion and multiple measurement occasions within subjects is directly observed. Vermunt (2008) describes various latent class models for hierarchical data where mixtures are defined at different levels of hierarchy. Applications of (general) growth mixture models for longitudinal data are described by Muthén and Shedden (1999) and Muthén (2001), among others. Bayesian applications of mixture models can be found in Diebolt and Robert (1994), and Gelman and King (1990), and see also Gelman et al. (1995) and McLachlan and Peel (2000) and references therein.

6.8 Exercises

6.1. Assume an MLIRT model consisting of a two-parameter measurement model for the latent response data (Equation (4.38)) and an empty multilevel model for the ability parameters (Equations (6.4) and (6.5)).
(a) Show that the covariance structure of two latent responses to item k for fixed item parameters can be expressed as

$$\text{Cov}\left(Z_{ijk}, Z_{i'j'k'}\right) = \begin{cases} a_k^2\left(\sigma_\theta^2 + \tau_{00}^2\right) + \varsigma & \text{for } i = i', j = j', k = k' \\ a_k^2\tau_{00}^2 & \text{for } i \neq i', j = j', k = k' \\ 0 & \text{for } j \neq j', \end{cases}$$

where $\varsigma = 1$ or $\varsigma = \pi^2/3$ when the latent response data are normally or logistically distributed, respectively.
(b) Show that the correlation between latent responses to the same item is given by

$$\rho_I = \frac{\tau_{00}^2}{\left(\sigma_\theta^2 + \tau_{00}^2\right) + \varsigma a_k^{-2}} \quad \text{for } i \neq i', j = j', k = k',$$

and interpret the result.
(c) The covariance structure of two latent responses to different items for fixed item parameters can be expressed as

$$\text{Cov}\left(Z_{ijk}, Z_{i'j'k'}\right) = \begin{cases} a_k\left(\sigma_\theta^2 + \tau_{00}^2\right)a_{k'} + \varsigma & \text{for } i = i', j = j', k \neq k' \\ a_k\tau_{00}^2 a_{k'} & \text{for } i \neq i', j = j', k \neq k' \\ 0 & \text{for } j \neq j'. \end{cases}$$

(d) Show that the correlation between latent responses to different items is given by

$$\rho_I = \frac{\tau_{00}^2}{\left(\sigma_\theta^2 + \tau_{00}^2\right) + a_k^{-1}\varsigma a_{k'}^{-1}} \quad \text{for } i \neq i', j = j', k \neq k',$$

and interpret the result.

6.2. Consider an empty multilevel model for the ability parameter (Equations (6.4) and (6.5)).
(a) Argue that the value of τ_{00}^2 influences the relationship between the β_{0j}.
(b) Substantiate the effect of normal prior $p\left(\beta_{0j} \mid \gamma_{00}, \tau_{00}^2 \rightarrow \infty\right)$ on the expected school effect in Equation (6.19).
(c) Consider the prior $p\left(\beta_{0j}\right) \propto c$, and derive the expected school effect given the abilities. (Box and Tiao, 1973, p. 379, called this locally uniform prior a fixed effect prior.)

6.3. Consider an unrestricted MLIRT model $\mathcal{M}$. Let MLIRT model $\mathcal{M}_0$ be a restricted version in which fixed effects $\boldsymbol{\gamma}$ equal $\boldsymbol{\gamma}_0$. Assume a priori that $P\left(\mathcal{M}\right) = P\left(\mathcal{M}_0\right)$. Let $\boldsymbol{\Omega}$ denote the set of MLIRT model parameters excluding $\boldsymbol{\gamma}$.
(a) Show that the Bayes factor comparing $\mathcal{M}_0$ with $\mathcal{M}$ has the form

$$BF = \frac{\int_{\boldsymbol{\Omega}} p\left(\mathbf{y} \mid \boldsymbol{\Omega}, \mathcal{M}_0\right) p\left(\boldsymbol{\Omega} \mid \mathcal{M}_0\right) \mathrm{d}\boldsymbol{\Omega}}{\int_{\boldsymbol{\Omega}} \int_{\boldsymbol{\gamma}} p\left(\mathbf{y} \mid \boldsymbol{\Omega}, \boldsymbol{\gamma}, \mathcal{M}\right) p\left(\boldsymbol{\Omega}, \boldsymbol{\gamma} \mid \mathcal{M}\right) \mathrm{d}\boldsymbol{\gamma} \mathrm{d}\boldsymbol{\Omega}}. \tag{6.33}$$

(b) Show that the denominator of Equation (6.33) can be expressed as

$$p\left(\mathbf{y} \mid \mathcal{M}\right) = \frac{p\left(\mathbf{y} \mid \boldsymbol{\Omega}, \boldsymbol{\gamma}_0, \mathcal{M}\right) p\left(\boldsymbol{\Omega}, \boldsymbol{\gamma}_0 \mid \mathcal{M}\right)}{p\left(\boldsymbol{\Omega}, \boldsymbol{\gamma}_0 \mid \mathbf{y}, \mathcal{M}\right)}.$$

(c) Show that the numerator of Equation (6.33) can be expanded as

$$p\left(\mathbf{y} \mid \mathcal{M}_0\right) = p\left(\boldsymbol{\gamma}_0 \mid \mathbf{y}, \mathcal{M}\right) \int_{\boldsymbol{\Omega}} p\left(\mathbf{y} \mid \boldsymbol{\Omega}, \mathcal{M}_0\right) p\left(\boldsymbol{\Omega} \mid \mathcal{M}_0\right) \cdot \frac{p\left(\boldsymbol{\Omega} \mid \mathbf{y}, \boldsymbol{\gamma}_0, \mathcal{M}\right)}{p\left(\boldsymbol{\Omega}, \boldsymbol{\gamma}_0 \mid \mathbf{y}, \mathcal{M}\right)} \mathrm{d}\boldsymbol{\Omega}.$$

(d) Use the results in (b) and (c) to express the Bayes factor as

$$BF = \frac{p\left(\boldsymbol{\gamma}_0 \mid \mathbf{y}, \mathcal{M}\right)}{p\left(\boldsymbol{\gamma}_0 \mid \mathcal{M}\right)} E\left[\frac{p\left(\boldsymbol{\Omega} \mid \mathcal{M}_0\right)}{p\left(\boldsymbol{\Omega} \mid \boldsymbol{\gamma}_0, \mathcal{M}\right)} \mid \mathbf{y}\right], \tag{6.34}$$

where the expectation is with respect to $p\left(\boldsymbol{\Omega} \mid \mathbf{y}, \boldsymbol{\gamma}_0, \mathcal{M}\right)$.
(e) Show that the Bayes factor in (d) equals the ratio

$$BF = \frac{p\left(\boldsymbol{\gamma}_0 \mid \mathbf{y}, \mathcal{M}\right)}{p\left(\boldsymbol{\gamma}_0 \mid \mathcal{M}\right)} \tag{6.35}$$

when

$$p\left(\boldsymbol{\Omega} \mid \boldsymbol{\gamma} = \boldsymbol{\gamma}_0, \mathcal{M}\right) = p\left(\boldsymbol{\Omega} \mid \mathcal{M}_0\right).$$

(f) Describe a method for estimating the ratio in Equation (6.35) using MCMC scheme 4. (The special form of the Bayes factor in Equation (6.35) is from Dickey, 1971; see also Equation (3.12). The generalized form in Equation (6.34) is from Verdinelli and Wasserman, 1995.)

6.4. (continuation of Exercise 6.1) Consider the problem of making predictions with an MLIRT model.
(a) Discuss a strategy for generating predictive data $\mathbf{y}_{rep}$ from respondents in the dataset given $\mathbf{y}$ using draws from MCMC scheme 4.
(b) Discuss a strategy for generating predictive data $\mathbf{y}_{rep}$ from respondents in the dataset given $\mathbf{y}$ using a forward simulator (see Section 5.4.2).
(c) Let θ_{sj} denote the ability of an unsampled respondent s in a sampled group j. Show that

$$\begin{pmatrix} \theta_{sj} \\ \boldsymbol{\theta}_j \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_{sj}^t \mathbf{w}_j \boldsymbol{\gamma} \\ \mathbf{x}_j \mathbf{w}_j \boldsymbol{\gamma} \end{bmatrix}, \begin{bmatrix} \mathbf{x}_{sj}^t \mathbf{T} \mathbf{x}_{sj} + \sigma_\theta^2 & \mathbf{x}_{sj}^t \mathbf{T} \mathbf{x}_j^t \\ \mathbf{x}_j \mathbf{T} \mathbf{x}_{sj} & \mathbf{x}_j \mathbf{T} \mathbf{x}_j^t + \sigma_\theta^2 \mathbf{I}_{n_j} \end{bmatrix} \right).$$

(d) Describe a method for predicting θ_{sj} given estimates of the multilevel model parameters $(\boldsymbol{\theta}_j, \boldsymbol{\gamma}, \sigma_\theta^2, \mathbf{T})$. (This prediction technique can be recognized as the multilevel prediction rule of Afshartous and De Leeuw, 2005.)
(e) Discuss a method for predicting a response pattern of an unsampled respondent s in a sampled group j given item parameters and using the result in (d).

6.5. A convenient form of the posterior density of $\boldsymbol{\beta}_j$ is to be derived given $p\left(\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{T}\right)$ as the prior density (Equation (6.8)) and likelihood function $p\left(\boldsymbol{\theta}_j \mid \mathbf{x}_j, \boldsymbol{\beta}_j, \sigma_\theta^2\right)$ (Equation (6.7)) with $\sigma_\theta^2 = 1$.
(a) Show that the terms within the exponent of the posterior density can be expressed as

$$\begin{aligned} D &= \left(\boldsymbol{\beta}_j - \mathbf{w}_j \boldsymbol{\gamma}\right)^t \mathbf{T}^{-1} \left(\boldsymbol{\beta}_j - \mathbf{w}_j \boldsymbol{\gamma}\right) + \left(\boldsymbol{\theta}_j - \mathbf{x}_j \boldsymbol{\beta}_j\right)^t \left(\boldsymbol{\theta}_j - \mathbf{x}_j \boldsymbol{\beta}_j\right) \\ &= \left(\boldsymbol{\beta}_j - \mathbf{w}_j \boldsymbol{\gamma}\right)^t \mathbf{T}^{-1} \left(\boldsymbol{\beta}_j - \mathbf{w}_j \boldsymbol{\gamma}\right) + \left(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\right)^t \mathbf{x}_j^t \mathbf{x}_j \left(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\right) \\ &\quad + \left(\boldsymbol{\theta}_j - \mathbf{x}_j \hat{\boldsymbol{\beta}}_j\right)^t \left(\boldsymbol{\theta}_j - \mathbf{x}_j \hat{\boldsymbol{\beta}}_j\right), \end{aligned} \tag{6.36}$$

where $\hat{\boldsymbol{\beta}}_j = \left(\mathbf{x}_j^t \mathbf{x}_j\right)^{-1} \mathbf{x}_j^t \boldsymbol{\theta}_j$.
(b) Let the function $D\left(\boldsymbol{\beta}_j\right)$ contain the terms of (6.36) that include $\boldsymbol{\beta}_j$. Set the first derivative of $D\left(\boldsymbol{\beta}_j\right)$ to zero,

$$\frac{\mathrm{d}D\left(\boldsymbol{\beta}_j\right)}{\mathrm{d}\boldsymbol{\beta}_j} = \mathbf{0}.$$

Solve this equation and show that $\tilde{\boldsymbol{\beta}}_j$ minimizes $D\left(\boldsymbol{\beta}_j\right)$ where

$$\tilde{\boldsymbol{\beta}}_j = \left(\mathbf{T}^{-1} + \mathbf{x}^t \mathbf{x}\right)^{-1} \left(\mathbf{T}^{-1} \mathbf{w}_j \boldsymbol{\gamma} + \mathbf{x}^t \mathbf{x} \hat{\boldsymbol{\beta}}_j\right).$$

(c) Show that

$$
\begin{aligned}
D\left(\boldsymbol{\beta}_j\right) &= \left(\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j\right)^t \left(\mathbf{T}^{-1} + \mathbf{x}_j^t \mathbf{x}_j\right) \left(\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j\right) \\
&\quad + \hat{\boldsymbol{\beta}}_j^t \mathbf{x}^t \mathbf{x} \hat{\boldsymbol{\beta}}_j^t + \left(\mathbf{w}_j \boldsymbol{\gamma}\right)^t \mathbf{T}^{-1} \mathbf{w}_j \boldsymbol{\gamma} - \tilde{\boldsymbol{\beta}}_j^t \left(\mathbf{T}^{-1} + \mathbf{x}_j^t \mathbf{x}_j\right) \tilde{\boldsymbol{\beta}}_j \\
&= \left(\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j\right)^t \left(\mathbf{T}^{-1} + \mathbf{x}_j^t \mathbf{x}_j\right) \left(\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j\right) \\
&\quad + \left(\hat{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_j\right)^t \mathbf{x}_j^t \mathbf{x}_j \left(\hat{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_j\right) + \left(\mathbf{w}_j \boldsymbol{\gamma} - \tilde{\boldsymbol{\beta}}_j\right)^t \mathbf{T}^{-1} \left(\mathbf{w}_j \boldsymbol{\gamma} - \tilde{\boldsymbol{\beta}}_j\right).
\end{aligned}
$$

(d) Derive the conditional posterior density of $\boldsymbol{\beta}_j$ from (c).

6.6. Consider the conditional density of θ_{ij} as defined in Equation (6.24).
(a) Suppress the conditioning on $(\boldsymbol{\xi}, \boldsymbol{\beta}, \sigma_\theta)$ and show that the posterior probability of $\theta_l \leq \theta_{ij} \leq \theta_u$ can be expressed as

$$
\begin{aligned}
P\left(\theta_l \leq \theta_{ij} \leq \theta_u \mid \mathbf{y}_{ij}\right) = \int \Bigg[& \Phi\left(\frac{\theta_u - \mu_\theta}{\sqrt{\Omega_\theta}}\right) - \\
& \Phi\left(\frac{\theta_l - \mu_\theta}{\sqrt{\Omega_\theta}}\right) \Bigg] p\left(\mathbf{z}_{ij} \mid \mathbf{y}_{ij}\right) \mathrm{d}\mathbf{z}_{ij}, \qquad (6.37)
\end{aligned}
$$

where μ_θ and Ω_θ are defined in Equations (6.25) and (6.26), respectively.

(b) For known values of θ_u and θ_l, show how to compute the posterior probability in (6.37) using MCMC scheme 4.
(c) Let $\theta_{ij}^{(m)}$ $(m = 1, \ldots, M)$ be draws from the marginal posterior distribution of θ_{ij}. Show how to estimate the posterior probability in (6.37) using the drawn values. Does this estimate differ from the estimate of (b)?

6.7. A credible region for the multivariate parameter $\boldsymbol{\beta}$ is defined as

$$
P\left(\mathcal{C}_\beta \mid \mathbf{y}\right) = \int_{\mathcal{C}_{\boldsymbol{\beta}}} p\left(\boldsymbol{\beta} \mid \mathbf{y}\right) \mathrm{d}\boldsymbol{\beta},
$$

where $\mathcal{C}_\beta$ does not need to be an interval. Assume that $\beta_q^{(m)}$ $(m = 1, \ldots, M; q = 1, \ldots, Q)$ is an MCMC sample from the (unimodal) marginal posterior distribution.
(a) For $Q = 1$, an order statistics estimate of a $(1 - \alpha)$ credible interval is $\left(\beta_{(x)}, \beta_{(x+(1-\alpha)M)}\right)$, where $\beta_{(x)}$ is the xth smallest value in the ordered sequence of $\beta^{(m)}$. Show how to estimate an equal-tailed interval and an HPD interval using the order statistics estimator.
(b) For $Q > 1$, show how to estimate a credible region for $\boldsymbol{\beta}$ by estimating the proportion of samples that fall simultaneously in all univariate credible intervals.
(c) Argue that the bounds are conservative and that the estimated credible region is restricted to be hyper-rectangular.

6.8. Investigate the properness of a noninformative prior (see Section 2.2.1). Assume a two-parameter response model with item parameters $(\mathbf{a}, \mathbf{b})$. Let F denote a normal or logistic cumulative distribution function. A standard normal prior is defined for θ_i and an improper prior for the item parameters, $p(\mathbf{a}, \mathbf{b}) \propto \prod_{k=1}^{K} p(a_k)p(b_k) = \prod_{k=1}^{K} I(a_k > 0)$.
(a) Show that the density of the observed data given item parameters equals

$$p(\mathbf{y}_i \mid \mathbf{a}, \mathbf{b}) = \int_{-\infty}^{\infty} \prod_{k=1}^{K} F(\eta_{ik})^{y_{ik}} (1 - F(\eta_{ik}))^{1-y_{ik}} \phi(\theta_i) \, \mathrm{d}\theta_i,$$

where $\eta_{ik} = a_k\theta_i - b_k$.
(b) Derive the following upper bound and lower bound: if $Y_{i1} = 1$,

$$p(\mathbf{y}_i \mid \mathbf{a}, \mathbf{b}) \geq F(-b_1) \int_{0}^{\infty} \prod_{k=2}^{K} F(\eta_{ik})^{y_{ik}} (1 - F(\eta_{ik}))^{1-y_{ik}} \phi(\theta_i) \, \mathrm{d}\theta_i,$$

and if $Y_{i1} = 0$,

$$p(\mathbf{y}_i \mid \mathbf{a}, \mathbf{b}) \geq (1 - F(-b_1)) \int_{-\infty}^{0} \prod_{k=2}^{K} F(\eta_{ik})^{y_{ik}} (1 - F(\eta_{ik}))^{1-y_{ik}} \phi(\theta_i) \, \mathrm{d}\theta_i.$$

(c) When the responses to item 1 are all correct or all incorrect, prove that

$$\int_{0}^{\infty} \prod_{i=1}^{N} p(y_{i1} \mid a_1, b_1) \, p(a_1) \mathrm{d}a_1 \geq \infty.$$

(d) What can be said about the improperness of the posterior densities $p(\mathbf{a}, \mathbf{b} \mid \mathbf{y})$ and $p(\mathbf{a}, \mathbf{b}, \boldsymbol{\theta} \mid \mathbf{y})$?

6.9. Consider the PISA 2003 data discussed in Section 6.6.2. Use the MLIRT software (Fox, 2007) to explore relationships between background information and students' math abilities.
(a) Estimate the mean, standard deviation, and 95% HPD interval to summarize the posterior density of the intraclass correlation coefficient.
(b) Use the variable school's mean index of economic, social, and cultural status to explain variation in the mean school levels of math ability. Compute the proportional reduction in variance at the school level.
(c) Estimate the main effect of gender on math ability, and investigate whether the effect varies across schools, conditional on the school's mean index of economic, social, and cultural status.
(d) Define and estimate a cross-level interaction term to investigate whether there are gender differences in the effect of a school's mean index of economic, social, and cultural status on math ability.
(e) Define and estimate a level-2 interaction term to investigate whether there are gender differences in the effect of the index of economic, social, and cultural status on math ability.

6.9 Appendix: The Expected School Effect

In Equation (6.22), a simplified expression for the conditional expected school effect, $E\left(u_{0j} \mid \mathbf{z}_j\right)$, is given without conditioning on other model parameters for notational convenience. This expression is obtained by analytically performing an inverse problem that is contained in the general expression of the expected school effect. This inverse problem equals

$$\mathbf{D}^{-1} = \left[\left(\mathbf{J}_{n_j} \otimes \tau^2 \mathbf{a}\mathbf{a}^t\right) + \left(\mathbf{I}_{n_j} \otimes \left(\sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K\right)\right)\right]^{-1}.$$

The expression for the inverse of a Schur complement, Equation (6.14), can be used to obtain the inverted matrix. It follows that

$$\mathbf{D}^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t \mathbf{A}^{-1}}{\tau^{-2} + \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t \mathbf{A}^{-1}\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)},$$

where

$$\begin{aligned}\mathbf{A}^{-1} &= \left[\mathbf{I}_{n_j} \otimes \left(\sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K\right)\right]^{-1} \\ &= \mathbf{I}_{n_j} \otimes \left(\mathbf{I}_K - \frac{\mathbf{a}\mathbf{a}^t}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}}\right),\end{aligned}$$

again using the matrix inverse expression of Equation (6.14).

The expression for $\mathbf{D}^{-1}$ is plugged into Equation (6.21). This leads to

$$\begin{aligned}E\left(u_{0j} \mid \mathbf{z}_j\right) &= \left(\mathbf{1}_{n_j}^t \otimes \tau^2 \mathbf{a}^t\right)\mathbf{D}^{-1}\left(\mathbf{z}_j - \left(\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)\gamma_{00} - \mathbf{1}_{n_j} \otimes \mathbf{b}\right)\right) \\ &= \left(\mathbf{1}_{n_j}^t \otimes \tau^2 \mathbf{a}^t\right)\mathbf{D}^{-1}\tilde{\mathbf{z}}_j. \qquad (6.38)\end{aligned}$$

Some tedious calculations need to be done to obtain the expression in Equation (6.22). The matrix operations can be separated into three parts. The first part equals

$$\begin{aligned}\left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\tau^2\right)\mathbf{A}^{-1}\tilde{\mathbf{z}}_j &= \left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\tau^2\right)\tilde{\mathbf{z}}_j - \left(\frac{\left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\tau^2\right)\mathbf{a}\mathbf{a}^t}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}}\right)\tilde{\mathbf{z}}_j \\ &= \tau^2\left(\frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}}\right)\left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\right)\tilde{\mathbf{z}}_j,\end{aligned}$$

the second

$$\begin{aligned}&\left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\tau^2\right)\mathbf{A}^{-1}\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t\mathbf{A}^{-1}\tilde{\mathbf{z}}_j = \\ &\left(n_j\tau^2\mathbf{a}^t\mathbf{a}\right)\left(\frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t\mathbf{a}}\right)^2\left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\right)\tilde{\mathbf{z}}_j,\end{aligned}$$

and the third

$$\tau^{-2} + \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t \mathbf{A}^{-1} \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right) = \tau^{-2} + n_j \mathbf{a}^t \mathbf{a} \left(\frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t \mathbf{a}}\right).$$

The three results are plugged into Equation (6.38), which leads to

$$\begin{aligned}
E\left(u_{0j} \mid \mathbf{z}_j\right) &= \tau^2 \left(\frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t \mathbf{a}}\right) \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t \tilde{\mathbf{z}}_j - \\
&\qquad \frac{n_j \tau^2 \mathbf{a}^t \mathbf{a} \left(\frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t \mathbf{a}}\right)^2}{\tau^{-2} + n_j \mathbf{a}^t \mathbf{a} \left(\frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t \mathbf{a}}\right)} \left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\right) \tilde{\mathbf{z}}_j \\
&= \left(\frac{\tau^2 \sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t \mathbf{a}}\right) \left[1 - \frac{n_j \mathbf{a}^t \mathbf{a} \left(\frac{\tau^2 \sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t \mathbf{a}}\right)}{1 + n_j \mathbf{a}^t \mathbf{a} \left(\frac{\tau^2 \sigma_\theta^{-2}}{\sigma_\theta^{-2} + \mathbf{a}^t \mathbf{a}}\right)}\right] \left(\mathbf{1}_{n_j}^t \otimes \mathbf{a}^t\right) \tilde{\mathbf{z}}_j \\
&= \left(\frac{\frac{\tau^2}{\sigma_\theta^2 + (\mathbf{a}^t \mathbf{a})^{-1}}}{1 + \frac{n_j \tau^2}{\sigma_\theta^2 + (\mathbf{a}^t \mathbf{a})^{-1}}}\right) \left(\mathbf{1}_{n_j}^t \otimes \left(\mathbf{a}^t \mathbf{a}\right)^{-1} \mathbf{a}^t\right) \tilde{\mathbf{z}}_j \\
&= \left(\frac{n_j \tau^2}{\sigma_\theta^2 + (\mathbf{a}^t \mathbf{a})^{-1} + n_j \tau^2}\right) \left(\mathbf{1}_{n_j}^t \otimes \left(\mathbf{a}^t \mathbf{a}\right)^{-1} \mathbf{a}^t\right) \tilde{\mathbf{z}}_j / n_j \\
&= \left(\frac{n_j \tau^2}{\sigma_\theta^2 + (\mathbf{a}^t \mathbf{a})^{-1} + n_j \tau^2}\right) \hat{u}_{0j},
\end{aligned}$$

where $\hat{u}_{0j}$ is the mean least squares estimate of u_{0j} given the individual observations. This follows from Equation (6.18) when treating the within-individual and within-school error components as one normally distributed error component. That is,

$$\begin{aligned}
\hat{u}_{0j} &= \left(\left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)\right)^{-1} \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t \tilde{\mathbf{z}}_j \\
&= \left(\mathbf{1}_{n_j}^t \mathbf{1}_{n_j} \otimes \mathbf{a}^t \mathbf{a}\right)^{-1} \left(\mathbf{1}_{n_j} \otimes \mathbf{a}\right)^t \tilde{\mathbf{z}}_j \\
&= \sum_i \frac{1}{n_j} \left(\mathbf{a}^t \mathbf{a}\right)^{-1} \mathbf{a}^t \tilde{\mathbf{z}}_{ij} \\
&= \sum_i \frac{1}{n_j} \left(\mathbf{a}^t \mathbf{a}\right)^{-1} \mathbf{a}^t \left(\mathbf{z}_{ij} - (\mathbf{a}\gamma_{00} - \mathbf{b})\right) \\
&= \left(\mathbf{a}^t \mathbf{a}\right)^{-1} \mathbf{a}^t \left(\overline{\mathbf{z}}_j - (\mathbf{a}\gamma_{00} - \mathbf{b})\right).
\end{aligned}$$

6.10 Appendix: Likelihood MLIRT Model

An expression is derived for the MLIRT integrated likelihood considering a two-parameter normal ogive model and the structural multilevel model as defined in Equations (6.2) and (6.3). Integrated likelihoods, where the likelihood is marginalized with respect to the random effect parameters, of other MLIRT models can be derived in a similar way.

The idea is to obtain an expression of the likelihood for the augmented data $\mathbf{z}$ and estimate the likelihood for the observed data via MCMC. The likelihood of the observed data can be expressed as the integrated augmented data likelihood; that is,

$$p\left(\mathbf{y} \mid \boldsymbol{\gamma}, \boldsymbol{\xi}, \sigma_\theta^2, \mathbf{T}\right) = \int p\left(\mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\xi}, \sigma_\theta^2, \mathbf{T}\right) \mathrm{d}\mathbf{z}.$$

Let $\boldsymbol{\Lambda} = \left(\boldsymbol{\gamma}, \boldsymbol{\xi}, \sigma_\theta^2, \mathbf{T}\right)$. Interest is focused on the augmented data likelihood,

$$\begin{aligned} p\left(\mathbf{z} \mid \boldsymbol{\Lambda}\right) &= \frac{p\left(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{\Lambda}\right)}{p\left(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Lambda}\right)} \\ &= \frac{p\left(\mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{\xi}, \sigma_\theta^2\right)}{p\left(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)} \frac{p\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{T}\right)}{p\left(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Lambda}\right)} \\ &= p\left(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}\right) \frac{p\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{T}\right)}{p\left(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Lambda}\right)}. \end{aligned} \tag{6.39}$$

First, the conditional augmented data likelihood is derived given the random effects parameters $\boldsymbol{\beta}_j$ (the first term on the right-hand side of (6.39)). Second, the integrated augmented data likelihood is derived by dividing the conditional posterior density of $\mathbf{z}, \boldsymbol{\beta}$ given $\boldsymbol{\Lambda}$ by the conditional posterior density of $\boldsymbol{\beta}$ given $(\mathbf{z}, \boldsymbol{\Lambda})$.

The first step gives

$$\begin{aligned} p\left(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}\right) = \prod_j (2\pi)^{-Kn_j/2} \left(\frac{\Omega_\theta}{\sigma_\theta^2}\right)^{n_j/2} \exp\left(-\frac{1}{2} \sum_{i,k} \left((z_{ijk} + b_k) - a_k \tilde{\theta}_{ij}\right)^2\right) \\ \exp\left(\frac{-1}{2\sigma_\theta^2} \left(\tilde{\boldsymbol{\theta}}_j - \mathbf{x}_j \boldsymbol{\beta}_j\right)^t \left(\tilde{\boldsymbol{\theta}}_j - \mathbf{x}_j \boldsymbol{\beta}_j\right)\right), \end{aligned} \tag{6.40}$$

where $\tilde{\theta}_{ij}$ is normally distributed with mean μ_θ and variance Ω_θ according to Equations (6.25) and (6.26), respectively.

The second step requires the conditional density $p(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Lambda})$. Therefore, consider the MLIRT model presentation

$$\begin{aligned} \mathbf{Z}_j &= \boldsymbol{\theta}_j \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b} + \boldsymbol{\epsilon}_j \\ &= \left(\mathbf{x}_j \boldsymbol{\beta}_j + \mathbf{e}_j\right) \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b} + \boldsymbol{\epsilon}_j \\ &= \left(\mathbf{x}_j \left(\mathbf{w}_j \boldsymbol{\gamma} + \mathbf{u}_j\right)\right) \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b} + \mathbf{e}_j \otimes \mathbf{a} + \boldsymbol{\epsilon}_j \\ &= \underbrace{\mathbf{x}_j \mathbf{w}_j \boldsymbol{\gamma} \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b}}_{\textit{fixed part}} + \underbrace{\mathbf{x}_j \mathbf{u}_j \otimes \mathbf{a} + \mathbf{e}_j \otimes \mathbf{a} + \boldsymbol{\epsilon}_j}_{\textit{random part}}, \end{aligned} \tag{6.41}$$

where $\mathbf{Z}_j = (Z_{1j1}, \ldots, Z_{1jK}, \ldots, Z_{n_j jK})^t$ is the stacked vector of n_j individual augmented response vectors. The joint distribution of $(\mathbf{Z}_j, \boldsymbol{\beta}_j)^t$ given $\boldsymbol{\Lambda}$ is multivariate normal, and from Equation (6.41) the covariance matrix of $(\mathbf{Z}_j, \boldsymbol{\beta}_j)$ can be obtained as

$$\begin{bmatrix} \mathbf{x}_j\mathbf{T}\mathbf{x}_j^t \otimes \mathbf{a}\mathbf{a}^t + \mathbf{I}_{n_j} \otimes \left(\sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K\right) & | & \mathbf{x}_j\mathbf{T} \otimes \mathbf{a} \\ \mathbf{T}\mathbf{x}_j^t \otimes \mathbf{a}^t & | & \mathbf{T} \end{bmatrix}.$$

Subsequently, the conditional distribution of $\boldsymbol{\beta}_j \mid \mathbf{z}_j, \boldsymbol{\Lambda}$ is normal with mean

$$\tilde{\boldsymbol{\beta}}_j = E\left(\boldsymbol{\beta}_j \mid \mathbf{z}_j, \boldsymbol{\Lambda}\right) = \mathbf{w}_j\boldsymbol{\gamma} + \left(\mathbf{T}\mathbf{x}_j^t \otimes \mathbf{a}^t\right)\mathbf{D}^{-1} \cdot \left(\mathbf{z}_j - \left(\mathbf{x}_j\mathbf{w}_j\boldsymbol{\gamma} \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b}\right)\right) \tag{6.42}$$

and variance

$$\boldsymbol{\Sigma}_{\tilde{\beta}_j} = \mathrm{Var}\left(\boldsymbol{\beta}_j \mid \mathbf{z}_j, \boldsymbol{\Lambda}\right) = \mathbf{T} + \left(\mathbf{T}\mathbf{x}_j^t \otimes \mathbf{a}^t\right)\mathbf{D}^{-1}\left(\mathbf{x}_j\mathbf{T} \otimes \mathbf{a}\right), \tag{6.43}$$

where $\mathbf{D} = \mathbf{x}_j\mathbf{T}\mathbf{x}_j^t \otimes \mathbf{a}\mathbf{a}^t + \mathbf{I}_{n_j} \otimes \left(\sigma_\theta^2 \mathbf{a}\mathbf{a}^t + \mathbf{I}_K\right)$.

The likelihood of the MLIRT model is obtained by performing the second step using the conditional normal distribution of $\boldsymbol{\beta}_j \mid \mathbf{z}_j, \boldsymbol{\Lambda}$. It follows that

$$\begin{aligned} p\left(\mathbf{z}_j \mid \boldsymbol{\Lambda}\right) &= \frac{p(\mathbf{z}_j \mid \boldsymbol{\beta}_j, \boldsymbol{\Lambda}) p(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T})}{p(\boldsymbol{\beta}_j \mid \mathbf{z}_j, \boldsymbol{\Lambda})} \qquad (6.44) \\ &= (2\pi)^{-Kn_j/2} \left(\frac{\Omega_\theta}{\sigma_\theta^2}\right)^{n_j/2} |\mathbf{T}|^{-1/2} |\boldsymbol{\Sigma}_{\tilde{\beta}_j}|^{1/2} \exp\left(-S\left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\beta}}_j\right)/2\right), \end{aligned}$$

where

$$\begin{aligned} S\left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\beta}}_j\right) = \sum_{i,k} \left((z_{ijk} + b_k) - a_k\tilde{\theta}_{ij}\right)^2 + \sigma_\theta^{-2}\left(\tilde{\boldsymbol{\theta}}_j - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right)^t \left(\tilde{\boldsymbol{\theta}}_j - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right) + \\ \left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1}\left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right) \end{aligned} \tag{6.45}$$

and $\tilde{\theta}_{ij}$ is conditionally normally distributed with mean μ_θ (Equation (6.25)) and variance Ω_θ (Equation (6.26)) for $\boldsymbol{\beta}_j = \tilde{\boldsymbol{\beta}}_j$.

7

Random Item Effects Models

Cluster-specific item effects parameters are introduced that are assumed to vary over clusters of respondents. The modeling of cluster-specific item parameters relaxes the assumptions of measurement invariance. Item characteristic differences are simply allowed, and it is not necessary to classify items as being invariant or noninvariant. Tests and estimation methods are discussed for item response models with random item effects parameters.

7.1 Random Item Parameters

Thus far in this book, attention has been focused on structural models for the person parameters for which the data are strictly hierarchically structured. A typical example was given in Chapter 6, where a set of item responses belongs to one student who belongs to one specific school. The additional nesting of the responses in items leads to a cross-classified data structure. In general, an item response model describes the relationship between responses and abilities, with each observation defined by the cross-classification of persons and items. The parameters of the item response model can then vary over persons and items. Here, attention is focused on the item side of the model.

Prior beliefs about the item parameters need to be specified, and they are presented in a prior probability distribution. The prior distribution can be constructed from subjective beliefs or objective data-based information. The general prior structure given by Equation (2.3) defines an exchangeable prior model since there is often no prior knowledge about differences in item characteristics. This prior distribution already defines random item parameters.

It makes sense to handle the item parameters as random without needing sampling arguments. The prior structure, Equations (2.4) and (2.5), reduces the parameter space to two dimensions, which may improve the item parameter estimates. A prior distribution for the item parameters also makes it possible to easily integrate prior information from experts and background information, handle hierarchical item structures, and express measurement

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_7,

uncertainty. For example, the linear-logistic test model (LLTM) assumes a perfect linear decomposition of the item difficulty parameter using item covariates. This unrealistic assumption is easily adjusted by adding a random error term at the item level such that the item difficulty parameter is treated as random (e.g., Janssen, Schepers and Peres, 2004; Klein Entink, Kuhn, Hornke and Fox, 2009b).

In the psychometrics literature, other applications are described that motivate the definition of random item parameters. Albers, Does, Imbos and Janssen (1989) described an application of a Rasch model with random item and person parameters. In their longitudinal study, students are obligated to participate in a progress test four times a year for a period of six years. Each progress test consisted of items that are randomly selected from the same constant item bank. The sampling process of items induces a sampling variance that has to be taken into account. In general, in domain-referenced testing, a test is assembled for each person from an item bank. The item sampling design induces random item characteristics.

Glas and van der Linden (2003) considered the application of item cloning. In this procedure, items are generated by a computer algorithm given a parent item (e.g., item shell or item template). Although the item cloning techniques are still improving, item characteristics of cloned items show within-parent variation. The item responses are not independent conditional on the ability parameters since items from the same parent may be statistically related. The problem of conditional independence is tackled by allowing random variability between cloned item characteristics conditional on the parent item, which induces random item parameters. Another related example of random item parameters was given by Janssen, Tuerlinckx, Meulders and De Boeck (2000). They described a criterion referenced test measuring several achievement targets. They assumed that items' characteristics are correlated when the items are nested within the same criterion, leading to within-criterion dependency. A hierarchical item structure was defined for the random item parameters.

De Jong, Steenkamp and Fox (2007), and De Jong, Steenkamp, Fox and Baumgartner (2008) introduced a random item modeling approach for survey-based marketing research. They introduced random item effects to accommodate cross-national differences in item characteristics. In this case, although a fixed test is used, the item characteristics can typically be regarded as random to generalize the inferences to some population of countries. In this chapter, the natural extension to random item parameters for cross-national survey data will be pursued further.

7.1.1 Measurement Invariance

International comparative survey studies such as the Programma for International Student Assessment (PISA) and the Third International Mathematics and Science Study (TIMMS) are focused on international comparisons of student achievement. Assessing comparability of the test scores across countries,

cultures, and different educational systems is a well-known complex problem. The main issue is that the measurement instrument has to exhibit adequate cross-national equivalence. This means that the calibration of the measurement instrument remains invariant across populations (nations, countries) of examinees. Meaningful comparisons can be made when measurement invariance holds.

There are several types of invariance that have to be dealt with as a prerequisite to conducting comparisons across countries (e.g., Meredith and Millsap, 1992; Steenkamp and Baumgartner, 1998). A lack of configural invariance conveys that the latent variable being measured has some degree of differential meaningfulness across countries. The construct being measured may vary across countries. The strength of the relationship between the underlying latent variable and the items may vary across countries, which indicates a lack of metric invariance. If an item satisfies the requirement of metric invariance, scores on that item can still be biased. Cross-national differences in the means of the observed item scores can be invoked by differences in the means of the underlying construct. Scalar invariance refers to the consistency between cross-national differences in latent means and cross-national differences in observed means. A thorough discussion of these and other types of invariance in a confirmatory factor-analytic framework is given by Vandenberg and Lance (2000).

In an item response theory framework, the item response function for an invariant item is identical across populations. That is, for binary response data, the discrimination and difficulty parameter of a measurement-invariant item are assumed to be equal across populations. Mellenbergh (1989) explicitly defined a measurement-invariant item, or an unbiased item, as the property that the item's characteristic function does not depend on group membership,

$$P\left(Y_{ik} = 1 \mid G, \theta_i\right) = P\left(Y_{ik} = 1 \mid \theta_i\right), \tag{7.1}$$

where variable G characterizes the population respondent i belongs to. When Equation (7.1) does not hold, the response probability depends on group membership given the level of the latent variable. Such an item behaves differently across populations and is said to show differential item functioning (DIF). Item properties that concern the distribution of the latent variable may still vary across populations, even for a measurement-invariant item.

7.1.2 Random Item Effects Prior

Assume that an item is not invariant but observations from a group are conditionally independent given the level of the latent variable and group-specific item characteristics. The group-specific item parameters differ across groups, but the differences are considered to be random measurement errors. Therefore, item characteristics are modeled as random item effects parameters to establish conditional independence, where the random part captures random error due to differential item functioning.

In Chapter 2, an exchangeable hierarchical prior is defined that allows for a within-item correlation structure without a priori knowledge to predict how the characteristics of item k will differ from the characteristics of item k'. Then, a multivariate normal prior density for the item parameters was specified,

$$(a_k, b_k)^t \sim \mathcal{N}\left(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right) I_{\mathcal{A}_k}(a_k), \tag{7.2}$$

with prior parameters

$$\begin{aligned} \boldsymbol{\Sigma}_\xi &\sim \mathcal{IW}(\nu, \boldsymbol{\Sigma}_0), \\ \boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\xi / K_0), \end{aligned}$$

where $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$. This level of the hierarchical prior presents a model for the invariant item characteristics. It is assumed that the item parameters (a_k, b_k) are invariant across populations. They will be referred to as "international item characteristics". In present cross-national survey research, attention is focused on obtaining the values of the international item characteristics, and they are used to obtain individual and/or country scores based on the assumptions of measurement invariance.

It is not likely that all international item characteristics apply to each individual in a cross-national survey. A more realistic assumption is to assume that for each nation some of the item characteristics may deviate from the international item characteristics. That is, the nation-specific item parameter values are influenced by the international item characteristics and the nation-specific characteristics (e.g., observed responses, response behavior). Note that these kinds of differences can be explained if they are caused by, for example, specific background differences (e.g., culture or gender differences). It is also possible to allow for cross-classified differences in item characteristics when, for example, cross-national and cross-cultural response heterogeneity is present. In most cross-national research, this kind of heterogeneity is not allowed since the conventional aim is to establish measurement invariance based on the assumption that measurement instruments can be developed in such a way that they contain items that display the same statistical properties in each nation.

Subsequently, a random item effects structure is defined to allow for random error variation in the item's functioning across nations. That is,

$$\tilde{\boldsymbol{\xi}}_{kj} = \left(\tilde{a}_{kj}, \tilde{b}_{kj}\right)^t \sim \mathcal{N}\left((a_k, b_k)^t, \boldsymbol{\Sigma}_{\tilde{\xi}}\right) \tag{7.3}$$

for $j = 1, \ldots, J$. The covariance matrix may vary across items. An independent random effects structure is defined when $\boldsymbol{\Sigma}_{\tilde{\xi}}$ is a diagonal matrix with diagonal elements $\sigma^2_{a_k}$ and $\sigma^2_{b_k}$. Treating the nation-specific item parameters as random effects induces a natural covariance structure:

$$\operatorname{Cov}\left(\tilde{\boldsymbol{\xi}}_{kj}, \tilde{\boldsymbol{\xi}}_{k'j'}\right) = \begin{cases} \boldsymbol{\Sigma}_\xi + \boldsymbol{\Sigma}_{\tilde{\xi}} & k = k', j = j' \\ \boldsymbol{\Sigma}_\xi & k = k', j \neq j' \\ 0 & k \neq k'. \end{cases}$$

In general, noninvariant nation-specific item characteristics are defined but are nested within the international item characteristics.

The random item effects prior can be generalized to define random item threshold parameters for polytomous items. In the graded response model, defined in Equation (1.8) (see also Section 4.6), the threshold parameter $\kappa_{k,c}$ divided by the item discrimination parameter a_k represents the latent variable value necessary to respond above category c with probability .5. The threshold parameters can be considered invariant across countries and will be referred to as the international item threshold parameters. For each item k, an exchangeable hierarchical prior is constructed to allow for random error variation in the country-specific threshold parameters across countries. An exchangeable proper hierarchical prior is specified for the international threshold parameters. For $c = 1, \ldots, C_k - 1$,

$$p\left(\kappa_{k,c}\right) \propto I_{\mathcal{A}}\left(\kappa_{k,c}\right),$$

where $\mathcal{A} = \{\kappa_{k,c} \in \mathcal{R}, \kappa_{k,0} < \ldots < \kappa_{k,c} < \ldots < \kappa_{k,C_k}\}$ with $\kappa_{k,0} = -\infty$ and $\kappa_{k,C_k} = \infty$. Now, country-specific random threshold effects are defined as

$$\tilde{\kappa}_{kj,c} = \kappa_{k,c} + \epsilon_{\kappa_{kj,c}} \tag{7.4}$$

for $j = 1, \ldots, J$ and $c = 1, \ldots, C_{k-1}$, where $\epsilon_{\kappa_{kj,c}} \sim \mathcal{N}\left(0, \sigma^2_{\kappa_k}\right)$. The country-specific threshold effects have an order restriction,

$$-\infty = \tilde{\kappa}_{kj,0} < \ldots < \tilde{\kappa}_{kj,c} < \ldots < \tilde{\kappa}_{kj,C_k} = \infty.$$

The error term is a country-specific deviation that is independently normally distributed given the value of the international threshold parameter. The item-specific variance $\sigma^2_{\kappa_k}$ controls the amount of variation in the nation-specific threshold parameters. Note that the order of the country-specific threshold parameters does not lead directly to an order of the international threshold parameters. However, the order of the conditional expected country-specific threshold parameters orders the international threshold parameters. The residual random threshold parameters are independently normally distributed with mean zero given $\boldsymbol{\kappa}_k$. It follows that

$$E\left(\kappa_{k,1} + \epsilon_{\kappa_{kj,1}} \mid \boldsymbol{\kappa}_k\right) < E\left(\kappa_{k,2} + \epsilon_{\kappa_{kj,2}} \mid \boldsymbol{\kappa}_k\right)$$

and, subsequently, $\kappa_{k,1} < \kappa_{k,2}$.

Analogous to the introduction of random threshold effects, an exchangeable hierarchical prior for the discrimination parameters that allows for random variation across countries is defined as

$$\tilde{a}_{kj} \mid a_k \sim \mathcal{N}\left(a_k, \sigma^2_{a_k}\right), \tag{7.5}$$

$$a_k \sim \mathcal{N}\left(\mu_a, \sigma^2_a\right) I_{\mathcal{A}_k}(a_k), \tag{7.6}$$

where $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$. Note that the prior level-1 and level-2 variances only differ with respect to a single index. However, the variance terms have

completely different meanings. The prior variance at level 2 represents the a priori uncertainty about the value of a_k, and the prior variance at level 1 represents the cross-national variation in item discrimination. The level-1 variance parameter is of specific interest since it can be used to investigate the assumption of measurement invariance for item k.

In random effects modeling, interest is usually focused on the effect of grouping the observations where the level-2 variance is informative about the between-group variability. In this case, nation-specific item characteristics are clustered within items. The between-item variability (at level 2) often is not of particular interest, and items usually discriminate differently (see Section 5.3). Here, interest is focused on the between-country variability (at level 1) since it presents the variation in country-specific item parameters.

To complete the full Bayesian modeling approach, proper conjugate priors are specified for the variance and covariance prior parameters. An inverse Wishart distribution is specified for $\mathbf{\Sigma}_{\tilde{\xi}}$ with degrees of freedom $n_{\tilde{\xi}} \geq 2$ and scale matrix $\mathbf{S}_{\tilde{\xi}}$. An exchangeable inverse gamma distribution is set for $\sigma^2_{a_k}$, and a normal inverse gamma prior is set for the parameters (μ_a, σ^2_a).

7.2 A Random Item Effects Response Model

Binary Response Data

An observation Y_{ijk} refers to an answer of respondent i in nation j to item k. Taking account of random item effects, the item response model (level 1) for binary data reads as

$$P\left(Y_{ijk} = 1 \mid \theta_i, \tilde{\boldsymbol{\xi}}_{jk}\right) = \begin{cases} \Phi\left(\tilde{a}_{kj}\theta_i - \tilde{b}_{kj}\right) \\ \Psi\left(d\left(\tilde{a}_{kj}\theta_i - \tilde{b}_{kj}\right)\right), \end{cases} \tag{7.7}$$

where for the moment the nesting of respondents in countries is ignored. The random item effects parameters are assumed to be normally distributed. The logistic item response model with normally distributed random item effects leads to a complex MCMC scheme. Moreover, in contrast to the logistic item response model, it will be shown that normally distributed random item effects fit naturally into the normal ogive response model, which leads to readily interpretable parameters.

The normal ogive response model in Equation (7.7) describes the probability of a correct response given a person's ability and group-specific item characteristics. The distribution of the random item effects is introduced in Equation (7.3). Assume an independent random item effects structure. Then a combined or integrated likelihood model can be stated as

$$P\left(Y_{ijk} = 1 \mid \theta_i, \boldsymbol{\xi}_k\right) = \Phi\left(a_k\theta_i - b_k - \epsilon_{b_{kj}} + \epsilon_{a_{kj}}\theta_i\right), \tag{7.8}$$

where the random item effects $\epsilon_{b_{kj}}$ and $\epsilon_{a_{kj}}$ are normally distributed with mean zero and variance $\sigma^2_{b_k}$ and $\sigma^2_{a_k}$, respectively. Conditional on a person's ability, in addition to the mean structure, there are two sources of variation: the unexplained variation in difficulties and discriminations across countries for item k.

Observations from respondents from the same country to item k are assumed to be correlated. Let this correlation structure be modeled via the random item difficulty parameters. Let $\tilde{Z}_{ijk}$ denote a standard normally distributed underlying response variable. The conditional probability of $Y_{ijk} = 1$ given θ_i and $\boldsymbol{\xi}_k$ can be expressed as the expected conditional success probability,

$$\begin{aligned} P\left(Y_{ijk} = 1 \mid \theta_i, \boldsymbol{\xi}_k\right) &= E\left(\Phi\left(a_k\theta_i - \tilde{b}_{kj}\right)\right) \qquad (7.9)\\ &= E\left(P\left(\tilde{Z}_{ijk} \leq a_k\theta_i - \tilde{b}_{kj} \mid \tilde{b}_{kj}\right)\right)\\ &= P\left(\tilde{Z}_{ijk} \leq a_k\theta_i - b_k - \epsilon_{b_{kj}}\right)\\ &= P\left(\tilde{Z}_{ijk} + \epsilon_{b_{kj}} \leq a_k\theta_i - b_k\right)\\ &= \Phi\left(\frac{a_k\theta_i - b_k}{\sqrt{1+\sigma^2_{b_k}}}\right)\\ &= \Phi\left(a^*_k\theta_i - b^*_k\right), \qquad (7.10) \end{aligned}$$

where the third identity holds since the expected value of the conditional probability is the unconditional probability. It follows that the combined model, Equation (7.10), is again a normal ogive item response model, which motivates the use of the normal ogive formulation. Note that the international item parameters from Equation (7.10) are smaller than the true international item parameters due to ignoring cross-national variation in item difficulties since

$$\begin{aligned} a^*_k &= a_k/\sqrt{1+\sigma^2_{b_k}},\\ b^*_k &= b_k/\sqrt{1+\sigma^2_{b_k}}. \end{aligned}$$

Latent continuous responses are defined that underlie the discrete observed response data. Therefore, let Z_{ijk} denote the latent response of respondent i in country j to item k and

$$Z_{ijk} \mid Y_{ijk}, \theta_i, \tilde{\boldsymbol{\xi}}_{kj} \sim \mathcal{N}\left(\tilde{a}_{kj}\theta_i - \tilde{b}_{kj}, 1\right), \qquad (7.11)$$

where Y_{ijk} is the indicator that Z_{ijk} is positive. In this way, augmented data are defined according to Equation (4.7) given country-specific item parameters. Assume that $\theta_i \sim \mathcal{N}\left(\mu_\theta, \sigma^2_\theta\right)$. Then the combined likelihood model for the latent response data can be written as

$$\begin{aligned} Z_{ijk} &= \tilde{a}_{kj}\theta_i - \tilde{b}_{kj} + \epsilon_{ijk} \\ &= \tilde{a}_{kj}\left(\mu_\theta + \sigma_\theta\epsilon_\theta\right) - \left(b_k + \epsilon_{b_{kj}}\right) + \epsilon_{ijk} \\ &= \underbrace{a_k\mu_\theta - b_k}_{\textit{fixed part}} + \underbrace{a_k\sigma_\theta\epsilon_\theta + \epsilon_{a_{kj}}\mu_\theta + \epsilon_{a_{kj}}\sigma_\theta\epsilon_\theta - \epsilon_{b_{kj}} + \epsilon_{ijk}}_{\textit{random part}}, \end{aligned} \tag{7.12}$$

where ϵ_θ and ϵ_{ijk} are standard normally distributed. The fixed part is the general mean latent response to item k in the population. Besides the measurement error part peculiar to Z_{ijk}, the random part contains four additional random terms. Three sources of variability are expected: among persons, and among discrimination and difficulty parameters across countries for item k. The term $\epsilon_{b_{kj}}$ denotes the random effects error of the item difficulties since it is assumed that the difficulty of the item varies across countries. The term $a_k\sigma_\theta\epsilon_\theta$ contains the random effects error of the persons and accounts for the between-person variation in ability. The random effects error of the discriminations times the mean ability, $\epsilon_{a_{kj}}\mu_\theta$, accounts for the residual variation across countries in the association between the mean ability and the latent response. Finally, the term $\epsilon_{a_{kj}}\sigma_\theta\epsilon_\theta$ accounts for the residual variation across countries in the association between the between-person variation in ability and the latent response.

The normal ogive response model with random item effects partitions the total variance in the outcome into a within-country part and a between-country part. The within-country variance is further partitioned into error variance between individuals and error variance between an individual's observations. The between-country variance is induced by the random effects item parameters. The conditional covariance between two latent observations to item k can be expressed as (Exercise 7.5)

$$\begin{aligned} \operatorname{Cov}\left(Z_{ijk}, Z_{i'j'k}\right) &= \operatorname{Cov}\left(\tilde{a}_{kj}\theta_i - \tilde{b}_{kj} + \epsilon_{ijk}, \tilde{a}_{kj'}\theta_{i'} - \tilde{b}_{kj'} + \epsilon_{i'j'k}\right) \\ &= \operatorname{Cov}\left(\tilde{a}_{kj}\theta_i, \tilde{a}_{kj'}\theta_{i'}\right) + \operatorname{Cov}\left(b_{kj}, b_{kj'}\right) + \operatorname{Cov}\left(\epsilon_{ijk}, \epsilon_{i'j'k}\right) \\ &= \begin{cases} \sigma^2_{a_k}\sigma^2_\theta + \sigma^2_{a_k}\mu^2_\theta + a^2_k\sigma^2_\theta + \sigma^2_{b_k} + 1 & i = i', j = j' \\ \sigma^2_{a_k}\mu^2_\theta + \sigma^2_{b_k} & i \neq i', j = j' \\ 0 & j \neq j'. \end{cases} \end{aligned}$$

An intraclass correlation coefficient can be defined that presents the proportion of variance in the latent outcomes that is attributable to countries,

$$\rho_{\xi_k} = \frac{\sigma^2_{a_k}\mu^2_\theta + \sigma^2_{b_k}}{\sigma^2_{a_k}\sigma^2_\theta + \sigma^2_{a_k}\mu^2_\theta + a^2_k\sigma^2_\theta + \sigma^2_{b_k} + 1}. \tag{7.13}$$

Polytomous Response Data

Observed ordinal response data are stored in a matrix $\mathbf{y}$, and an observation y_{ijk} refers to an answer of respondent i in nation j to item k in category

$c = 1, \ldots, C_k$. Taking account of random item effects, Equations (7.4) and (7.5), the level-1 item response model reads as

$$P\left(Y_{ijk} = c \mid \theta_i, \tilde{\boldsymbol{\xi}}_{kj}\right) = \begin{cases} \Phi\left(\tilde{a}_{kj}\theta_i - \tilde{\kappa}_{kj,c-1}\right) - \Phi\left(\tilde{a}_{kj}\theta_i - \tilde{\kappa}_{kj,c}\right) \\ \Psi\left(\tilde{a}_{kj}\theta_i - \tilde{\kappa}_{kj,c-1}\right) - \Psi\left(\tilde{a}_{kj}\theta_i - \tilde{\kappa}_{kj,c}\right), \end{cases} \quad (7.14)$$

where $\tilde{\boldsymbol{\xi}}_{kj} = (\tilde{a}_{kj}, \tilde{\boldsymbol{\kappa}}_{kj})$. The latent (auxiliary) variable formulation for the graded response model was introduced in Chapter 4, where normally distributed latent response data were defined; see Equation (4.25).

The introduction of a normally distributed underlying variable allows for easily interpretable parameters and easy formulation of an MCMC algorithm, and it increases the flexibility for making different model adjustments such as incorporating explanatory variables at the level of items and/or persons. Therefore, attention is focused on the normal ogive or probit version of the graded response model. Here, in the same way, normally distributed augmented data are defined as

$$Z_{ijk} \mid Y_{ijk} = c, \theta_i, \tilde{\boldsymbol{\xi}}_{kj} \sim \mathcal{N}\left(\tilde{a}_{kj}\theta_i, 1\right) I\left(\tilde{\kappa}_{kj,c-1} \leq Z_{ijk} \leq \tilde{\kappa}_{kj,c}\right). \quad (7.15)$$

Subsequently, a level-1 response model in terms of a standard normally distributed underlying variable $\tilde{Z}_{ijk}$ can be defined as

$$\begin{aligned} P\left(Y_{ijk} = c \mid \theta_i, \tilde{\boldsymbol{\xi}}_{kj}\right) &= P\left(\tilde{a}_{kj}\theta_i - \tilde{\kappa}_{kj,c} \leq \tilde{Z}_{ijk} \leq \tilde{a}_{kj}\theta_i - \tilde{\kappa}_{kj,c-1}\right) \\ &= P\left(\tilde{\kappa}_{kj,c-1} - \tilde{a}_{kj}\theta_i \leq \tilde{Z}_{ijk} \leq \tilde{\kappa}_{kj,c} - \tilde{a}_{kj}\theta_i\right) \\ &= P_{ijk}(c) - P_{ijk}(c-1). \end{aligned}$$

Cumulative or threshold models such as the graded response model are often defined by modeling the conditional cumulative response probability. The probability $P_{ijk}(c)$ presents the probability of responding in or below category c given item and person parameters. The conditional cumulative response probability can be given in terms of the observable variable Y_{ijk} and the unobservable underlying variable $\tilde{Z}_{ijk}$ since

$$P\left(Y_{ijk} \leq c \mid \theta_i, \tilde{\boldsymbol{\xi}}_{kj}\right) = P\left(\tilde{Z}_{ijk} \leq \tilde{\kappa}_{kj,c} - \tilde{a}_{kj}\theta_i \mid \theta_i, \tilde{\boldsymbol{\xi}}_{kj}\right).$$

Subsequently, a linear structure on the underlying latent variable $\tilde{Z}_{ijk}$ has the form

$$\begin{aligned} \tilde{Z}_{ijk} &= \tilde{\kappa}_{kj,c} - \tilde{a}_{kj}\theta_i + \epsilon_{ijk} \\ &= \kappa_{k,c} - a_k\theta_i + \epsilon_{\kappa_{kj,c}} - \epsilon_{a_{kj}}\theta_i + \epsilon_{ijk}, \end{aligned} \quad (7.16)$$

where $\tilde{Z}_{ijk} \leq \tilde{\kappa}_{kj,c} - \tilde{a}_{kj}\theta_i$ if $Y_{ijk} \leq c$ and $\tilde{Z}_{ijk} > \tilde{\kappa}_{kj,c} - \tilde{a}_{kj}\theta_i$ if $Y_{ijk} > c$.

The combined cumulative response model in terms of the underlying continuous response data can be recognized as a linear mixed effects model with

normally distributed random effects. The term $\epsilon_{\kappa_{kj,c}}$ presents the unexplained random effects error of upper grade thresholds across countries for category c of item k. The random effect consisting of the unexplained random discrimination effect times the ability, $\epsilon_{a_{kj}}\theta_i$, accounts for the variation across countries in the association between ability and the latent responses to item k. Note that θ_i is a random effects parameter that accounts for heterogeneity among persons.

In the linear latent response model, Equation (7.16), the latent response is modeled as a function of fixed item effects, $(a_k, \kappa_{k,c})$, and random (country-specific) item effects $(\epsilon_{a_{kj}}, \epsilon_{\kappa_{kj,c}})$. It can be shown that the implied marginal response model, by integrating out the random item effects, is again a normal ogive graded response model.

Therefore, assume random country-specific shifting of thresholds but invariant discrimination parameters such that $\sigma^2_{a_k} = 0$. The expected conditional cumulative success probability, where the expectation is taken with respect to the distribution of the random threshold parameters, can be expressed as

$$\begin{aligned}
P\left(Y_{ijk} \leq c \mid \theta_i, \boldsymbol{\xi}_k\right) &= E\left(\Phi\left(a_k\theta_i - \tilde{\kappa}_{kj,c-1}\right)\right) \\
&= E\left(P\left(\tilde{Z}_{ijk} \leq a_k\theta_i - \tilde{\kappa}_{kj,c-1} \mid \tilde{\kappa}_{kj,c-1}\right)\right) \\
&= P\left(\tilde{Z}_{ijk} \leq a_k\theta_i - \kappa_{k,c-1} - \epsilon_{\kappa_{kj,c-1}}\right) \\
&= P\left(\tilde{Z}_{ijk} + \epsilon_{\kappa_{kj,c-1}} \leq a_k\theta_i - \kappa_{k,c-1}\right) \\
&= \Phi\left(\frac{a_k\theta_i - \kappa_{k,c-1}}{\sqrt{\left(1 + \sigma^2_{\kappa_k}\right)}}\right) \\
&= \Phi\left(a^*_k\theta_i - \kappa^*_{k,c-1}\right).
\end{aligned}$$

Subsequently, the expected conditional success probability can be expressed as

$$P\left(Y_{ijk} = c \mid \theta_i, \boldsymbol{\xi}_k\right) = \Phi\left(a^*_k\theta_i - \kappa^*_{k,c-1}\right) - \Phi\left(a^*_k\theta_i - \kappa^*_{k,c}\right). \tag{7.17}$$

It follows that the marginal model is again a normal ogive graded response model. The logistic version of the graded response model does not have this property which is a reason for using the normal ogive graded response model. Furthermore, the international item parameters from Equation (7.17) are typically smaller than the true parameters, and they are attenuated when ignoring cross-national variation in item thresholds ($\sigma_{\kappa_k} > 0$) since

$$\begin{aligned}
a^*_k &= a_k / \sqrt{1 + \sigma^2_{\kappa_k}}, \\
\kappa^*_{k,c} &= \kappa_{k,c} / \sqrt{1 + \sigma^2_{\kappa_k}}.
\end{aligned}$$

7.2.1 Handling the Clustering of Respondents

Up to now, the nesting of respondents in countries has been ignored. However, the clustering of respondents can be properly modeled via a structural multilevel model, as presented in Chapter 6. Therefore, consider the empty multilevel model

$$\theta_{ij} = \beta_{0j} + e_{ij}, \tag{7.18}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{7.19}$$

where θ_{ij} is modeled by a country-specific intercept β_{0j} and the within-country level-1 error term e_{ij}. The level-1 error variance can also be specified to be country-specific, which means that the errors are independent normally distributed with mean zero and variance $\sigma^2_{\theta_j}$ given the country-specific intercept. This corresponds to relaxing the assumption of the factor variance invariance. The conditional variability of the latent variable is assumed to be different across countries given β_{0j}. However, the error term at level 2 is assumed to be independent normally distributed with mean zero and variance τ^2_{00}, which means that the covariance structure is assumed to be common across countries.

7.2.2 Explaining Cross-national Variation

The responses in each nation j are conditionally independent given the person and country-specific item parameters. Therefore, the usual local independence assumption holds when conditioning on the random effects. The posterior density of the nation-specific item parameters can be expressed as

$$p\left(\tilde{\boldsymbol{\xi}}_{kj} \mid \mathbf{y}_{jk}, \boldsymbol{\theta}_j, \boldsymbol{\xi}_k, \boldsymbol{\Sigma}_{\tilde{\xi}}\right) \propto \prod_{i=1}^{n_j} p\left(y_{ijk} \mid \theta_{ij}, \tilde{\boldsymbol{\xi}}_{kj}\right) p\left(\tilde{\boldsymbol{\xi}}_{kj} \mid \boldsymbol{\xi}_k, \boldsymbol{\Sigma}_{\tilde{\xi}}\right).$$

Response observations and prior information will be used to make inferences about the nation-specific item parameters. The international item parameters determine the prior mean of the nation-specific item parameters. The posterior mean of the nation-specific item parameters is also determined by the within-country information. In a full Bayes approach, shrinkage estimates of the nation-specific item parameters are obtained where the amount of shrinkage is driven by information in the data and the level of shrinkage depends on the heterogeneity among the country-specific item parameters. There is more shrinkage towards the common (international) item characteristics when less observed country-specific information is available or when the country-specific item parameters are tightly distributed around the international item parameters.

The variation in country-specific item parameters can be explained by cross-national explanatory information. Let matrix $\mathbf{v}$ contain information that

explains heterogeneity in item characteristics across countries. The random item effects model in Equation (7.3) can be extended to handle such covariate information as

$$\tilde{a}_{kj} = a_k + \mathbf{v}_{kj}\boldsymbol{\delta}_a + \epsilon_{a_{kj}}, \tag{7.20}$$

$$\tilde{b}_{kj} = b_k + \mathbf{v}_{kj}\boldsymbol{\delta}_b + \epsilon_{b_{kj}}, \tag{7.21}$$

where the errors are multivariate normally distributed with mean zero and variance $\boldsymbol{\Sigma}_{\tilde{\xi}}$. The covariate information to explain variation in cross-national item characteristics can differ between item parameters. The same extension can be made for the random item effects model for polytomous data. Subsequently, to explain heterogeneity among respondents, within-nation and between-nation covariate information can be used, as explained in Chapter 6, Equations (6.7) and (6.8). Note that in Chapter 6 the index j refers to a school but here it refers to a country.

7.2.3 The Likelihood for the Random Item Effects Model

Let $\boldsymbol{\Omega}$ represent the common model parameters such that $\boldsymbol{\Omega} = \left(\sigma_\theta^2, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\xi}, \boldsymbol{\Sigma}_{\tilde{\xi}}\right)$. The likelihood model for the observed data contains random effects at the item and person levels. The random effects models are used to average the conditional likelihood, which leads to an unconditional likelihood of only the common parameters. For that reason, the random effects models are considered part of the likelihood. Then, the likelihood of the common parameters is given by

$$p\left(\mathbf{y} \mid \boldsymbol{\Omega}\right) = \int\int\int \left[\prod_{j=1}^{J}\left[\prod_{i=1}^{n_j}\left[\prod_{k=1}^{K} p\left(y_{ijk} \mid \theta_{ij}, \tilde{\boldsymbol{\xi}}_{kj}\right) p\left(\tilde{\boldsymbol{\xi}}_{kj} \mid \mathbf{v}_{kj}, \boldsymbol{\delta}, \boldsymbol{\xi}_k, \boldsymbol{\Sigma}_{\tilde{\xi}}\right)\right.\right.\right.$$
$$\left.\left.\left. \mathrm{d}\tilde{\boldsymbol{\xi}}_{kj}\right] p\left(\theta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j, \sigma_\theta^2\right) \mathrm{d}\theta_{ij}\right] p\left(\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{T}\right)\right] \mathrm{d}\boldsymbol{\beta}_j. \tag{7.22}$$

The likelihood of the common model parameters consists of three levels. At the first level, the observations are distributed according to an item response model with country-specific random item effects and random person effects. At the second level, the distribution of the random effects is specified. At this level, the covariate information, stored in matrices $\mathbf{x}$ and $\mathbf{v}$, explains heterogeneity among respondents in country j and among item characteristics across countries, respectively. The third level describes the distribution of the random country effects that account for the between-country variation, and covariate $\mathbf{w}$ is used to explain this variation.

The modeling framework related to the likelihood in Equation (7.22) is depicted in Figure 7.1. The ellipse describes the nonlinear item response model. The multilevel structure on the person parameters was already discussed in

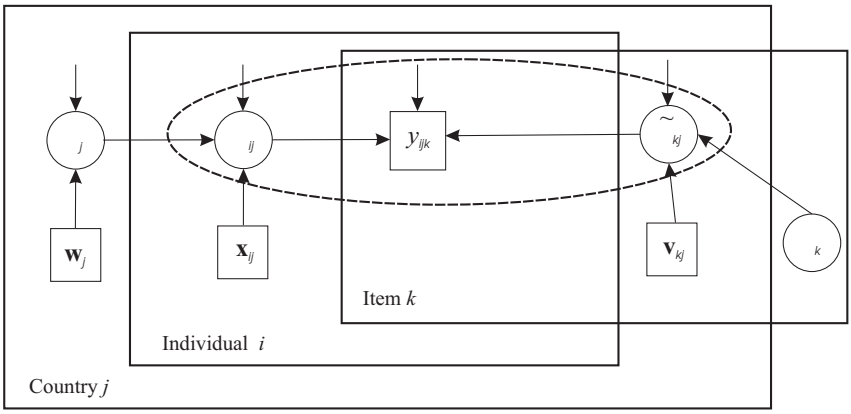

Fig. 7.1. Path diagram for a random item effects model.

Chapter 6. In this case, level-2 and level-3 random effects models describe the within-country and between-country variations of persons, respectively. The item observations are nested within the individuals, and the individuals are nested within countries. At level 2, random item effects parameters are introduced that account for the variation in item characteristics across countries. The explanatory information $\mathbf{v}_{kj}$ explains between-country variability in characteristics of item k. The hierarchical structure of the model is depicted in boxes. There is a strict hierarchical structure, where item observations are nested within individuals that are nested in countries. Another hierarchical structure of the model is the nesting of item observations in countries, leading to country-specific item characteristics. This is illustrated by allowing the item box to overlap the country box outside the individual box.

7.3 Identification: Linkage Between Countries

Meaningful comparisons across countries can be made when the individuals are measured on the same scale. When measurement invariance is present, the levels of the latent variable are measured on a common scale and cross-national differences can be meaningfully analyzed, and this will lead to valid interpretations. When the measurement scales differ across countries, cross-national differences in mean or variance levels are difficult to interpret since they do not reflect the true differences between countries.

As stated in Section 4.4, the scale of the latent variable needs to be identified. It was shown that it is possible to identify the scale via restrictions on item parameters or to fix the mean and if necessary the variance of the distribution of the latent variable. When restricting the scale in each country, the scale of the latent variable is identified, but it is not common across countries. Thus, additional restrictions are needed to link the countries.

Therefore, traditional practice in the analysis of cross-national data is to establish measurement invariance. The items satisfy the assumption of measurement invariance when their characteristics do not vary across nations. The item parameters are said to be invariant. In that case, persons with identical scores on the latent variable also have the same item scores. It is not necessary that all items exhibit measurement invariance. Comparing respondents on a common scale across countries is technically possible with at least one invariant item. This invariant item can be used as an anchor item to establish a common scale across countries. The other items are allowed to function differently but should be related to the common scale. Among others, Steenkamp and Baumgartner (1998) noted that at least two invariant items are needed when one needs to test the assumption of measurement invariance. This procedure has several limitations. First, testing items for measurement invariance is an exploratory post-hoc procedure that is prone to capitalization on chance. Second, when there are only a few invariant items, the usual tests for differential item functioning may identify invariant items as exhibiting noninvariance since the model also tries to fit the other noninvariant items (Holland and Wainer, 1993). Third, the existence of two invariant items is not realistic when the number of items is small and the number of countries is high. Then, measurement invariance may only hold for subgroups of countries.

7.3.1 Identification Without (Designated) Anchor Items

The random item effects model cannot be identified via marker or anchor items when all country-specific item parameters are modeled as random item parameters. Obviously, the random item parameters do not satisfy the assumption of measurement invariance. However, without identifying restrictions the scale of the latent variable is not identified. The object is to identify the random item effects model in such a way that a common latent scale is established across countries.

The combined random item effects model as stated in Equation (7.12) can be identified by fixing the scale of the latent variable. For example, for $\mu_\theta = 0$ and $\sigma_\theta = 1$, the model in Equation (7.12) can be recognized as a mixed effects model with random item effects and standard normally distributed random person effects; that is, random item difficulty effects and random item discrimination effects on the association between the (identified) random person effects and the latent responses.

First, assume that the nesting of respondents in countries is modeled via an empty multilevel model on the person parameters (Equations (6.4) and (6.5)), and assume invariant item discrimination. Then, the conditional success probability given country-specific difficulty parameters and latent country means is given by

$$
\begin{aligned}
P\left(Y_{ijk}=1 \mid \beta_{0j}, \tilde{b}_{kj}\right) &= E\left(\Phi\left(\theta_{ij}-\tilde{b}_{kj}\right)\right) \\
&= E\left(P\left(\tilde{Z}_{ijk} \leq \theta_{ij}-\tilde{b}_{kj} \mid \theta_{ij}, \tilde{b}_{kj}\right)\right) \\
&= P\left(\tilde{Z}_{ijk} \leq \beta_{0j}+e_{ij}-\tilde{b}_{kj} \mid \beta_{0j}, \tilde{b}_{kj}\right) \\
&= \Phi\left(\frac{\beta_{0j}-\tilde{b}_{kj}}{\sqrt{1+\sigma_{\theta}^2}}\right),
\end{aligned}
$$

where the random person effects e_{ij} are independently normally distributed and the random country effects and random item difficulty effects are independently distributed as $\beta_{0j} \sim \mathcal{N}(\gamma_{00}, \tau^2)$ and $\tilde{b}_{kj} \sim \mathcal{N}(b_k, \sigma_{b_k}^2)$, respectively. The model is not identified by fixing the mean of the latent scale by setting $\gamma_{00}=0$ or by restricting the sum of the international difficulty parameters, $\sum_k b_k = 0$. There is an indeterminacy between the country-specific latent mean and the location of the country-specific item difficulties. A common shift in the country-specific item difficulties can be compensated for by a similar shift in the country-specific latent mean. Therefore, in each country j, the location of the random item difficulties is restricted by setting $\sum_k \tilde{b}_{kj}=0$. This way, a common shift in random item difficulties in country j is not allowed. Instead, a comparable shift in the country-specific latent mean is captured. Subsequently, the location of the latent variable is identified in each country j, and therefore the general location of the common scale is identified. Note that, despite the identifying restriction, a common shift in country-specific item difficulties is possible when covariate information is available that explains it. In that case, the location of the country-specific item difficulties is restricted conditional on the item predictor effects $\boldsymbol{\delta}_b$ (see Equation (7.21)).

Second, assume random item effects parameters but still assume factor variance invariance such that $\sigma_{\theta_j}^2 = \sigma_{\theta}^2$ for each j. Condition only on the random effects for notational convenience. Then the conditional success probability can be expressed as

$$
\begin{aligned}
&P\left(Y_{ijk}=1 \mid \beta_{0j}, \tilde{b}_{kj}\right) = E\left(E\left(\Phi\left(\tilde{a}_{kj}\theta_{ij}-\tilde{b}_{kj} \mid \theta_{ij}, \tilde{\boldsymbol{\xi}}_{kj}\right)\right)\right) \\
&= E\left(P\left(\tilde{Z}_{ijk} \leq \tilde{a}_{kj}\left(\beta_{0j}+e_{ij}\right)-\tilde{b}_{kj} \mid \beta_{0j}, \tilde{\boldsymbol{\xi}}_{kj}\right)\right) \\
&= P\left(\tilde{Z}_{ijk} \leq a_k\left(\beta_{0j}+e_{ij}\right)+\epsilon_{a_{kj}}\left(\beta_{0j}+e_{ij}\right)-\tilde{b}_{kj} \mid \beta_{0j}, \tilde{b}_{kj}\right). \quad (7.23)
\end{aligned}
$$

Besides the restriction on the country-specific difficulty parameters, there is still an indeterminacy in the location of the international item discrimination parameters and the latent person errors e_{ij} (reflected in the term $a_k(\beta_{0j}+e_{ij})$). A common shift in the international discrimination parameters can be compensated for by a shift in the latent person errors. Therefore, it is necessary to fix the variance of the latent variable (for example, via rescaling; see Section 4.4.2) or restrict the international discrimination parameters by setting $\prod_k a_k = 1$.

Third, also assume factor variance noninvariance and allow country-specific variances of the latent variable. As a result, a restriction on the total variance of the latent variable leads to an unidentified model. This follows from the fact that a common shift of all item discrimination effects $\epsilon_{a_{k_j}}$ can be compensated for by an equal shift of the country-specific errors e_{ij} in country j. Note that the indeterminacy in the term $\epsilon_{a_{k_j}}(\beta_{0j} + e_{ij})$ of Equation (7.23) is apparent. This identification problem can be solved by restricting the country-specific item discriminations such that $\prod_k \tilde{a}_{kj} = 1$ for each j. In each country, a common shift in the random item discrimination effects is not allowed and will be captured by a shift in the country-specific variance of the latent variable. In the same way as for the random item difficulty effects, a common shift is allowed when such an effect can be explained by country-specific information.

The generalization to polytomous items is straightforward. For the most general case, the country-specific discrimination parameters are restricted via $\prod_k \tilde{a}_{kj} = 1$ for each j. A common shift in the country-specific threshold parameters can be compensated by a comparable negative shift in the country-specific latent mean. This indeterminacy is neutralized by disallowing a common shift in the country-specific threshold parameters and, in that case, forcing a comparable shift in the latent country mean. Let $\overline{\kappa}_{j,2} = \sum_k \tilde{\kappa}_{kj,2}/K$ denote the mean threshold value across items in country j. The rescaled threshold parameters are defined as $\tilde{\kappa}_{kj,c} - \overline{\kappa}_{j,2}$ for $k = 1, \ldots, K$, and $c = 1, \ldots, C_k - 1$, and they have a fixed location. It is easily seen that this location does not change when adding the same constant to all country-specific threshold parameters. Rescaling the random threshold parameters in each country also identifies the general location of the scale.

7.3.2 Concluding Remarks

This modeling framework has several important advantages. The person parameters can be estimated on a common scale given noninvariant items. Identification of the model does not depend on the presence of marker items. That is, measurement invariance assumptions are not needed to meaningfully compare item scores across countries. The model based on measurement invariance is nested within the model for measurement noninvariance. As a result, evidence of measurement invariance can be obtained by comparing both models. Deviations in item characteristics across countries can be considered measurement errors. Differences can be explained by explanatory information, but without the proper covariates they are considered to be measurement errors. The proposed modeling approach accounts for the (un)explained heterogeneity in response probabilities. For example, if the level of stylistic response is measured, differences between item characteristics across nations may be controlled by adjusting for it.

Shrinkage estimates of the nation-specific item parameters are based on the observations from that nation and the international item characteristics,

which are based on all respondents' observations. The shrinkage towards the international item parameters enhances the stability of the country-specific item parameter estimates.

The random item effects modeling approach reduces the number of effective model parameters in comparison with modeling nation-specific item parameters as fixed effects, which will rapidly increase the number of parameters with a growing number of nations and response options. Invariance hypotheses can be tested by evaluating different models via an information criterion. The Bayes factor can also be used for testing a null hypothesis concerning measurement invariance, which can be computed via the sampled values obtained from estimating the parameters of the most general model.

7.4 MCMC: Handling Order Restrictions

An MCMC scheme will be presented for the general random item effects model that allows measurement-noninvariant item parameters and country-specific distributional properties of the latent variable. In a fully Bayesian hierarchical modeling approach, each sampling step is characterized by a conditional independence assumption. This strategy leads to a straightforward implementation of the different sampling steps. However, specific attention has to be paid to the steps for drawing international and nation-specific threshold parameters, which are complicated due to order restrictions.

7.4.1 Sampling Threshold Values via an M-H Algorithm

Order restrictions complicate a direct sampling procedure and an M-H procedure is proposed to obtain samples from the full conditional. In the same way as in Chapter 4, adaptive M-H steps are defined to avoid specifying any tuning parameters. An M-H sampling procedure is discussed for the nation-specific and international threshold parameters and the threshold variance parameter $\sigma^2_{\kappa_k}$ (Exercise 7.6).

Nation-Specific Threshold Parameters

The adaptive proposal distribution is based on the information of the last iteration, denoted as m, and item k's candidate threshold values should follow the order restriction. Therefore, generate candidate nation-specific threshold values from

$$\tilde{\kappa}^*_{kj,c} \sim \mathcal{N}\left(\tilde{\kappa}^{(m)}_{kj,c}, \sigma^2_{mh}\right) I\left(\tilde{\kappa}^*_{kj,c-1} < \tilde{\kappa}^*_{kj,c} < \tilde{\kappa}^{(m)}_{kj,c+1}\right) \tag{7.24}$$

for $c = 1, \ldots, C_k - 1$, where $\tilde{\kappa}^{(m)}_{kj,c}$ is the value of $\tilde{\kappa}_{kj,c}$ in iteration m. The variance of the proposal distribution is adjusted to improve the efficiency of the MCMC algorithm as described in Section 4.2.

The conditional posterior density of threshold parameters $\tilde{\boldsymbol{\kappa}}_{kj}$ is given by

$$p\left(\tilde{\boldsymbol{\kappa}}_{kj} \mid \theta_{ij}, \boldsymbol{\kappa}_k, \sigma^2_{\kappa_k}, \mathbf{y}\right) \propto \prod_{i|j} p\left(y_{ijk} \mid \theta_{ij}, \tilde{\boldsymbol{\kappa}}_{kj}\right) \prod_{c=1}^{C_k-1} p\left(\tilde{\kappa}_{kj,c} \mid \kappa_{k,c}, \sigma^2_{\kappa_k}\right),$$

where

$$p\left(\tilde{\kappa}_{kj,c} \mid \kappa_{k,c}, \sigma^2_{\kappa_k}\right) = \frac{\phi\left(\tilde{\kappa}_{kj,c}; \kappa_{k,c}, \sigma^2_{\kappa_k}\right)}{\int_{\tilde{\kappa}_{kj,c-1}}^{\tilde{\kappa}_{kj,c+1}} \phi\left(\tilde{\kappa}_{kj,c}; \kappa_{k,c}, \sigma^2_{\kappa_k}\right)}. \tag{7.25}$$

Note that the part of the normalizing constant that is attributable to the truncation cannot be ignored since it includes the threshold parameters of interest. An acceptance ratio can be defined to evaluate a proposed candidate $\tilde{\boldsymbol{\kappa}}^*_{kj}$ with the current state $\tilde{\boldsymbol{\kappa}}^{(m)}_{kj}$. The acceptance ratio R is defined as

$$R = \frac{p\left(\tilde{\boldsymbol{\kappa}}^*_{kj} \mid \theta_{ij}, \boldsymbol{\kappa}_k, \sigma^2_{\kappa_k}, \mathbf{y}\right)}{p\left(\tilde{\boldsymbol{\kappa}}^{(m)}_{kj} \mid \theta_{ij}, \boldsymbol{\kappa}_k, \sigma^2_{\kappa_k}, \mathbf{y}\right)} \prod_{c=1}^{C_k-1} \frac{\Phi\left(\frac{\tilde{\kappa}^{(m)}_{kj,c+1} - \tilde{\kappa}^{(m)}_{kj,c}}{\sigma_{mh}}\right) - \Phi\left(\frac{\tilde{\kappa}^*_{kj,c-1} - \tilde{\kappa}^{(m)}_{kj,c}}{\sigma_{mh}}\right)}{\Phi\left(\frac{\tilde{\kappa}^*_{kj,c+1} - \tilde{\kappa}^*_{kj,c}}{\sigma_{mh}}\right) - \Phi\left(\frac{\tilde{\kappa}^{(m)}_{kj,c-1} - \tilde{\kappa}^*_{kj,c}}{\sigma_{mh}}\right)}.$$

At iteration $m+1$, the transition $\tilde{\boldsymbol{\kappa}}^{(m+1)}_{kj} = \tilde{\boldsymbol{\kappa}}^*_{kj}$ is made with probability $\min(1, R)$. The second term of the acceptance ratio accounts for the difference in the normalizing constant of the proposal densities.

International Threshold Parameters

In the same way, an M-H step can be defined for the international threshold parameters. The conditional posterior density of the international threshold parameter is defined by

$$p\left(\kappa_{k,c} \mid \tilde{\boldsymbol{\kappa}}_{k,c}, \sigma^2_{\kappa_k}\right) \propto \prod_{j} p\left(\tilde{\kappa}_{kj,c} \mid \kappa_{k,c}, \sigma^2_{\kappa_k}\right),$$

which can be recognized as the product of truncated normal prior densities according to Equation (7.25). Define a proposal distribution as

$$\kappa^*_{k,c} \sim \mathcal{N}\left(\kappa^{(m)}_{k,c}, \sigma^2_{mh}\right) I\left(\kappa^*_{k,c-1} < \kappa^*_{k,c} < \kappa^{(m)}_{k,c+1}\right)$$

for $c = 1, \ldots, C_k - 1$. Now, an acceptance ratio can be defined to evaluate the posterior densities of the states κ^*_k and $\kappa^{(m)}_k$ while taking account of differences in the proposal densities.

7.4.2 Sampling Threshold Values via Gibbs Sampling

The nation-specific threshold parameters have a truncated normal prior according to Equation (7.4). The object is to define a corresponding set of threshold parameters that have a nontruncated normal prior. This new set of threshold parameters has to comply with the prior model for the nation-specific threshold parameters. This is achieved by restricting the cumulative prior probabilities of the transformed set of parameters to equal the cumulative prior probabilities of the nation-specific threshold parameters. Then, the transformed set of threshold parameters can be used to sample directly from the full conditionals of the other prior parameters.

Consider the normal distribution of a nation-specific threshold parameter truncated to the interval $\mathcal{A} = (\tilde{\kappa}_{kj,c-1}, \tilde{\kappa}_{kj,c+1})$

$$\tilde{\kappa}_{kj,c} = \nu_{kj,c} I_{\mathcal{A}}\left(\nu_{kj,c}\right),$$

where

$$I_{\mathcal{A}}\left(\nu_{kj,c}\right) = \begin{cases} 1 \ \nu_{kj,c} \in \mathcal{A} \\ 0 \text{ otherwise,} \end{cases}$$

where $\nu_{kj,c} \sim \mathcal{N}\left(\kappa_{k,c}, \sigma^2_{\kappa_k}\right)$. Now, parameter $\nu_{kj,c}$ can be recognized as the nontruncated version of $\tilde{\kappa}_{kj,c}$. Define the cumulative normal distribution function $F(\nu_{kj,c}) = \Phi\left((\nu_{kj,c} - \kappa_{k,c})/\sigma_{\kappa_k}\right)$.

The cumulative distribution function of $\tilde{\kappa}_{kj,c}$ follows from the cumulative distribution function of $\nu_{kj,c}$. That is, the cumulative probability p of values less than or equal to $\tilde{\kappa}_{kj,c}$ can be expressed as

$$p = G\left(\tilde{\kappa}_{kj,c}\right) = \frac{F\left(\tilde{\kappa}_{kj,c}\right) - F\left(\tilde{\kappa}_{kj,c-1}\right)}{F\left(\tilde{\kappa}_{kj,c+1}\right) - F\left(\tilde{\kappa}_{kj,c-1}\right)}. \tag{7.26}$$

In compliance with the prior model for $\tilde{\kappa}_{kj,c}$, the value of $\nu_{kj,c}$ that satisfies the equation $G\left(\tilde{\kappa}_{kj,c}\right) = F\left(\nu_{kj,c}\right)$ is considered to be the corresponding unique value of $\nu_{kj,c}$, where F is the nontruncated version of G. This corresponding value of $\nu_{kj,c}$ can be computed via the inverse cumulative distribution function method. Given probability p from Equation (7.26), the value of $\nu_{kj,c}$ is given by

$$\begin{aligned} \nu_{kj,c} &= F^{-1}\left(G\left(\tilde{\kappa}_{kj,c}\right)\right) = F^{-1}\left(p\right) = \kappa_{k,c} + \sigma_{\kappa_k} \Phi^{-1}\left(p\right) \\ &= \kappa_{k,c} + \sigma_{\kappa_k} \Phi^{-1}\left(\frac{F\left(\tilde{\kappa}_{kj,c}\right) - F\left(\tilde{\kappa}_{kj,c-1}\right)}{F\left(\tilde{\kappa}_{kj,c+1}\right) - F\left(\tilde{\kappa}_{kj,c-1}\right)}\right). \end{aligned} \tag{7.27}$$

This procedure can be repeated for each item k, j, and c.

Now, consider the factorization

$$\begin{aligned} p\left(\boldsymbol{\nu}_{k,c}, \kappa_{k,c}, \sigma^2_{\kappa_k} \mid \tilde{\boldsymbol{\kappa}}_{k,c}\right) &\propto p\left(\tilde{\boldsymbol{\kappa}}_{k,c} \mid \boldsymbol{\nu}_{k,c}, \kappa_{k,c}, \sigma^2_{\kappa_k}\right) p\left(\boldsymbol{\nu}_{k,c}, \kappa_{k,c}, \sigma^2_{\kappa_k}\right) \\ &\propto p\left(\tilde{\boldsymbol{\kappa}}_{k,c} \mid \boldsymbol{\nu}_{k,c}\right) p\left(\boldsymbol{\nu}_{k,c}, \kappa_{k,c}, \sigma^2_{\kappa_k}\right) \\ &\propto p\left(\boldsymbol{\nu}_{k,c} \mid \kappa_{k,c}, \sigma^2_{\kappa_k}\right) p\left(\kappa_{k,c}\right) p\left(\sigma^2_{\kappa_k}\right), \end{aligned}$$

where the conditional density $p\left(\tilde{\boldsymbol{\kappa}}_{k,c} \mid \boldsymbol{\nu}_{k,c}\right)$ equals one due to the relationship in Equation (7.27). As a result, inferences for $\kappa_{k,c}$ and $\sigma^2_{\kappa_k}$ can be based on $\boldsymbol{\nu}_k$. That is, the nontruncated normally distributed threshold parameters $\boldsymbol{\nu}_k$ can be used to sample international threshold parameters and a threshold variance parameter directly (see MCMC scheme 5).

7.4.3 Simultaneous Estimation via MCMC

MCMC Scheme 5

A1) Binary response data

1. *Sample augmented data* $\mathbf{z}^{(m+1)}$ *according to Equation* (7.11).
2. *For each* k *and* j, *sample* $\tilde{\boldsymbol{\xi}}_{kj}$ *from the conditional distribution*

$$\tilde{\boldsymbol{\xi}}_{kj} \mid \mathbf{z}_{kj}^{(m+1)}, \boldsymbol{\theta}_j^{(m)}, \boldsymbol{\xi}_k^{(m)}, \boldsymbol{\Sigma}_{\tilde{\xi}}^{(m)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\tilde{\xi}}^*, \boldsymbol{\Omega}_{\tilde{\xi}}\right),$$

where

$$\boldsymbol{\Omega}_{\tilde{\xi}}^{-1} = \mathbf{H}^t\mathbf{H} + \boldsymbol{\Sigma}_{\tilde{\xi}}^{-1}, \tag{7.28}$$

$$\boldsymbol{\mu}_{\tilde{\xi}}^* = \boldsymbol{\Omega}_{\tilde{\xi}}\left(\mathbf{H}^t\mathbf{z}_{kj} + \boldsymbol{\Sigma}_{\tilde{\xi}}^{-1}\boldsymbol{\xi}_k\right), \tag{7.29}$$

and $\mathbf{H} = \left(\boldsymbol{\theta}_j, -\mathbf{1}_{n_j}\right)$.

3. *For each* k, *sample* $\boldsymbol{\xi}_k$ *from the conditional distribution*

$$\boldsymbol{\xi}_k \mid \tilde{\boldsymbol{\xi}}_k^{(m+1)}, \boldsymbol{\Sigma}_{\tilde{\xi}}^{(m)}, \boldsymbol{\mu}_{\xi}^{(m)}, \boldsymbol{\Sigma}_{\xi}^{(m)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\xi}^*, \boldsymbol{\Omega}_{\xi}\right) I_{\mathcal{A}_k}\left(a_k\right),$$

where

$$\boldsymbol{\Omega}_{\xi}^{-1} = J\boldsymbol{\Sigma}_{\tilde{\xi}}^{-1} + \boldsymbol{\Sigma}_{\xi}^{-1},$$

$$\boldsymbol{\mu}_{\xi}^* = \boldsymbol{\Omega}_{\xi}\left(J\boldsymbol{\Sigma}_{\tilde{\xi}}^{-1}\hat{\boldsymbol{\xi}}_k + \boldsymbol{\Sigma}_{\xi}^{-1}\boldsymbol{\mu}_{\xi}\right),$$

with $\hat{\boldsymbol{\xi}}_k = \sum_j \tilde{\boldsymbol{\xi}}_{kj}/J$.

4. *Sample* $\boldsymbol{\Sigma}_{\tilde{\xi}}^{(m+1)}$ *given* $\tilde{\boldsymbol{\xi}}_k^{(m+1)}$ *and* $\boldsymbol{\xi}_k^{(m+1)}$ *from an inverse Wishart distribution with* $J + n_0$ *degrees of freedom and scale matrix*

$$\sum_j \left(\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\xi}_k\right)\left(\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\xi}_k\right)^t + \mathbf{S}_0.$$

5. *Sample* $\boldsymbol{\mu}_{\xi}^{(m+1)}, \boldsymbol{\Sigma}_{\xi}^{(m+1)}$ *via M-H accounting for the positivity restriction on the international item discriminations.*

A2) Polytomous response data

1. *Sample augmented data* $\mathbf{z}^{(m+1)}$ *according to Equation* (7.15).

2. *For each k and j, sample $\tilde{a}_{kj}^{(m+1)}$ from the conditional distribution*

$$\tilde{a}_{kj} \mid \mathbf{z}_{jk}^{(m+1)}, \boldsymbol{\theta}_j^{(m)}, a_k^{(m)}, \sigma_{a_k}^{2(m)} \sim \mathcal{N}\left(\mu_{\tilde{a}}^*, \Omega_{\tilde{a}}\right),$$

where

$$\begin{aligned} \Omega_{\tilde{a}}^{-1} &= \boldsymbol{\theta}_j^t \boldsymbol{\theta}_j + \sigma_{a_k}^{-2}, \\ \mu_{\tilde{a}}^* &= \Omega_{\tilde{a}} \left(\boldsymbol{\theta}_j^t \mathbf{z}_{jk} + a_k \sigma_{a_k}^{-2}\right). \end{aligned}$$

3. *For each k, sample discrimination parameters $a_k^{(m+1)}$ from the conditional distribution*

$$a_k \mid \tilde{\mathbf{a}}_k^{(m+1)}, \sigma_{a_k}^{2(m)}, \mu_a, \sigma_a^2 \sim \mathcal{N}\left(\mu_a^*, \Omega_a\right) I_{\mathcal{A}_k}\left(a_k\right),$$

where

$$\begin{aligned} \Omega_a^{-1} &= J\sigma_{a_k}^{-2} + \sigma_a^{-2} \\ \mu_a^* &= \Omega_a \left(\sum_j \tilde{a}_{kj}/\sigma_{a_k}^2 + \mu_a/\sigma_a^2 \right). \end{aligned}$$

4. *For each k, sample $\sigma_{a_k}^{2(m+1)}$ from an inverse gamma density with shape parameter g_1 and scale parameter $\sum_j \left(\tilde{a}_{kj} - a_k\right)^2/2 + g_2$.*
5. *For each k and j, sample $\tilde{\kappa}_{kj,c}^*$ for $c = 1, \ldots, C_k - 1$ from the proposal distribution; see Equation (7.24). The candidate nation-specific threshold parameters are evaluated for each item-country combination. Sample $U_{kj} \sim \mathcal{U}(0,1)$, and set $\tilde{\boldsymbol{\kappa}}_{kj}^{(m+1)} = \tilde{\boldsymbol{\kappa}}_{kj}^*$ when $U_{kj} \leq \min(1, R)$ where the acceptance ratio R is defined below Equation (7.25).*
6. *Compute $\boldsymbol{\nu}^{(m+1)}$ given $\tilde{\boldsymbol{\kappa}}^{(m+1)}$ according to Equation (7.27). For each k and $c = 1, \ldots, C_k - 1$, sample $\kappa_{k,c}^{(m+1)}$ from the truncated conditional posterior distribution,*

$$\kappa_{k,c} \mid \boldsymbol{\nu}_{k,c}^{(m+1)}, \sigma_{\kappa_k}^{2(m)} \sim \mathcal{N}\left(\overline{\nu}_{k,c}, \sigma_{\kappa_k}^2/J\right) I_{\mathcal{A}}\left(\kappa_{k,c}\right),$$

where $\overline{\nu}_{k,c} = \sum_j \nu_{kj,c}/J$ and $\mathcal{A} = (\kappa_{k,c-1}, \kappa_{k,c+1})$.
7. *Sample $\sigma_{\kappa_k}^{2(m+1)}$ from an inverse gamma density with shape parameter $g_1 + J(C_k - 1)/2$ and scale parameter*

$$\sum_j \sum_c \left(\nu_{kj,c} - \kappa_{k,c}\right)^2 \Big/ 2 + g_2.$$

B) Sample multilevel parameter values

1. *For each i and j, sample $\theta_{ij}^{(m+1)}$ given $\mathbf{z}_{ij}^{(m+1)}$, $\boldsymbol{\xi}^{(m+1)}$, $\boldsymbol{\beta}_j^{(m)}$, and $\sigma_{\theta_j}^{2(m)}$ according to step 4 in MCMC scheme 4.*

2. *For each j, sample $\boldsymbol{\beta}_j^{(m+1)}$ given $\boldsymbol{\theta}_j^{(m+1)}$, $\sigma_{\theta_j}^{2(m)}$, $\mathbf{T}^{(m)}$, and $\boldsymbol{\gamma}^{(m)}$ according to step 5 in MCMC scheme 4.*
3. *Sample $\boldsymbol{\gamma}_j^{(m+1)}$ given $\boldsymbol{\beta}_j^{(m+1)}$, $\mathbf{T}^{(m)}$, and σ_γ according to step 6 in MCMC scheme 4.*
4. *For each j, sample $\sigma_{\theta_j}^{2(m+1)}$ given $\boldsymbol{\theta}_j^{(m+1)}$ and $\boldsymbol{\beta}_j^{(m+1)}$ from an inverse gamma density with shape parameter $g_1 + n_j/2$ and scale parameter $g_2 + (\boldsymbol{\theta}_j - \mathbf{x}_j\boldsymbol{\beta}_j)^t (\boldsymbol{\theta}_j - \mathbf{x}_j\boldsymbol{\beta}_j)/2$.*
5. *Sample $\mathbf{T}^{(m+1)}$ given $\boldsymbol{\beta}^{(m+1)}$ and $\boldsymbol{\gamma}^{(m+1)}$ according to step 8 in MCMC scheme 4.*

7.5 Tests for Invariance

Measurement invariance and factor variance invariance can be tested via an information criterion. Therefore, an expression is needed for the likelihood of $\boldsymbol{\Lambda} = \left(\boldsymbol{\gamma}, \boldsymbol{\xi}, \sigma_\theta^2, \mathbf{T}, \boldsymbol{\Sigma}_{\tilde{\xi}}\right)$, the parameters of interest. An expression is derived for the random item effects likelihood for binary data. The likelihood for polytomous response data can be derived in a similar way. In correspondence with the computation of the MLIRT likelihood in Section 6.10, an expression for the augmented likelihood is derived and via MCMC the likelihood for the observed data is estimated. Express the likelihood of the observed data as the integrated augmented data likelihood:

$$\begin{aligned} p\left(\mathbf{y} \mid \boldsymbol{\gamma}, \boldsymbol{\xi}, \sigma_\theta^2, \mathbf{T}, \boldsymbol{\Sigma}_{\tilde{\xi}}\right) &= \int p\left(\mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\xi}, \sigma_\theta^2, \mathbf{T}, \boldsymbol{\Sigma}_{\tilde{\xi}}\right) \mathrm{d}\mathbf{z} \\ &= \int \frac{p\left(\mathbf{z}, \tilde{\boldsymbol{\xi}}^* \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)}{p\left(\tilde{\boldsymbol{\xi}}^* \mid \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)} \frac{p\left(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{\Lambda}\right)}{p\left(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Lambda}\right)} \mathrm{d}\mathbf{z}. \end{aligned} \tag{7.30}$$

For notational convenience, assume that $\tilde{\boldsymbol{\xi}}_{kj}$ is normally distributed with mean $\boldsymbol{\mu}_{\tilde{\theta}}^*$ from Equation (7.29) and variance $\boldsymbol{\Omega}_{\tilde{\xi}}$ from Equation (7.28). Then, for $\tilde{\boldsymbol{\xi}}_{kj}^* = \tilde{\boldsymbol{\xi}}_{kj}$, the augmented data likelihood, under the integral sign in Equation (7.30), can be expressed as

$$\begin{aligned} p\left(\mathbf{z} \mid \boldsymbol{\Lambda}\right) = \prod_j &(2\pi)^{-Kn_j/2} \left|\boldsymbol{\Sigma}_{\tilde{\xi}}\right|^{K/2} \left(\Omega_\theta/\sigma_\theta^2\right)^{n_j/2} |\mathbf{T}|^{-1/2} \cdot \\ &\left|\boldsymbol{\Sigma}_{\tilde{\beta}_j}\right|^{1/2} \exp\left(-S\left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\beta}}_j\right)/2\right), \end{aligned}$$

where

$$\begin{aligned} S\left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\beta}}_j\right) = &\sum_{i,k} \left((z_{ijk} + \tilde{b}_{kj}) - \tilde{a}_{kj}\tilde{\theta}_{ij}\right)^2 + \sum_k \left(\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\xi}_k\right)^t \boldsymbol{\Sigma}_{\tilde{\xi}}^{-1} \left(\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\xi}_k\right) \\ + &\sigma_{\theta_j}^{-2} \left(\tilde{\boldsymbol{\theta}}_j - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right)^t \left(\tilde{\boldsymbol{\theta}}_j - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right) + \left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)^t \mathbf{T}^{-1} \left(\tilde{\boldsymbol{\beta}}_j - \mathbf{w}_j\boldsymbol{\gamma}\right), \end{aligned}$$

using the expression for the augmented data likelihood in Equation (6.44). Note that the residual sum of squares is partitioned. The second term on the right-hand side presents the contribution of cross-national differences in item characteristics to the total residual sum of squares. It can be seen that the residual sum of squares of the MLIRT model in Equation (6.45), is nested within the residual sum of squares of the random item effects model. It follows that the MLIRT model can be compared with the random item effects model via a DIC or BIC to test for full measurement invariance. When the MLIRT model is preferred, the items exhibit measurement invariance. Other restricted models can be defined to test for specific measurement invariance assumptions.

Invariance tests based on HPD regions can be constructed from the posterior distributions of the most general random item effects model. This approach has the advantage that assumptions of measurement invariance and factor variance invariance can be tested by estimating the general random item effects model. To illustrate this approach, consider the conditional distribution of the nation-specific latent variable variances,

$$\begin{aligned} p\left(\sigma^2_{\theta_1}, \ldots, \sigma^2_{\theta_J} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) &\propto \prod_j \left(\sigma^2_{\theta_j}\right)^{-((g_1+n_j)/2+1)} \exp\left(-\frac{s_j^2 + g_2}{2\sigma^2_{\theta_j}}\right) \\ &\propto \prod_j \left(\sigma^2_{\theta_j}\right)^{-\left(n'_j/2+1\right)} \exp\left(-\frac{n'_j s'^2_j}{2\sigma^2_{\theta_j}}\right), \end{aligned}$$

where

$$s'^2_j = \frac{\sum_{i|j} \left(\theta_{ij} - \mathbf{x}^t_{ij}\boldsymbol{\beta}_j\right)^2 + g_2}{n_j + g_1},$$

using an inverse gamma prior with shape and scale parameters $g_1/2$ and $g_2/2$, respectively. Define $(J-1)$ linear independent contrasts $\Delta_j = \log\sigma^2_{\theta_j} - \log\sigma^2_{\theta_J}$. The point $\boldsymbol{\Delta}_0 = \mathbf{0}$ corresponds to the event that $\sigma^2_{\theta_1} = \ldots = \sigma^2_{\theta_J}$. The point $\boldsymbol{\Delta}_0 = \mathbf{0}$ is included in a $(1-\alpha)$ HPD region if and only if

$$P\left(p\left(\boldsymbol{\Delta} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) > p\left(\boldsymbol{\Delta}_0 \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) < 1 - \alpha.$$

Box and Tiao (1973, pp. 133–136) showed that, for $n_j \to \infty$, this probability statement is approximately equal to the probability statement

$$P\left(\chi^2_{J-1} \leq -\sum_{j=1}^{J} n_j \left(\log s'^2_j - \log \overline{s}'^2\right)\right) < 1 - \alpha, \tag{7.31}$$

where $\overline{s}'^2$ is the weighted average variance across nations. The marginal posterior probability that the point $\boldsymbol{\Delta}_0$ is included in the HPD region is computed by integrating out the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ using MCMC samples from their joint posterior. That is, the conditional posterior probability in Equation

(7.31) is evaluated in each MCMC iteration and the averaged posterior probability is considered to be an estimate of the marginal posterior probability.

To test for measurement invariance, consider the conditional posterior distribution of the nation-specific item parameters,

$$p\left(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z}_{kj}, \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_{\tilde{\xi}}, \boldsymbol{\xi}_k\right) \propto \exp\left(-\frac{1}{2}\sum_j \left(\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\mu}^*_{\tilde{\xi}}\right)^t \boldsymbol{\Omega}^{-1}_{\tilde{\xi}} \left(\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\mu}^*_{\tilde{\xi}}\right)\right),$$

where $\boldsymbol{\mu}^*_{\tilde{\xi}}$ and $\boldsymbol{\Omega}^{-1}_{\tilde{\xi}}$ are defined in Equations (7.29) and (7.28), respectively. The term within the exponent is chi-square distributed with $2J$ degrees of freedom. It follows that the event that $\tilde{\boldsymbol{\xi}}_{kj} = \boldsymbol{\xi}_k$ for each j, corresponding to the assumption of measurement invariance, is included in the $(1-\alpha)$ HPD region if and only if

$$P\left(\chi^2_{2J} \leq \sum_j \left(\boldsymbol{\xi}_k - \boldsymbol{\mu}^*_{\tilde{\xi}}\right)^t \boldsymbol{\Omega}^{-1}_{\tilde{\xi}} \left(\boldsymbol{\xi}_k - \boldsymbol{\mu}^*_{\tilde{\xi}}\right)\right) < 1-\alpha.$$

In the same way as above, the marginal posterior probability is estimated via MCMC. The procedure can be extended to test measurement invariances of a set of items simultaneously, to test scalar and metric invariances separately, or to test measurement-invariant properties of polytomous scored items.

7.6 International Comparisons of Student Achievement

To illustrate the random item effects modeling approach, data from the Programma of International Student Assessment (PISA) survey of mathematics performance were analyzed. In 2003, four subject domains were tested (mathematics, reading, science, and problem solving). There were 13 clusters, including seven mathematics clusters and two clusters in each of the other domains. According to the test design, 13 test booklets were defined that consisted of four clusters. Each cluster appeared exactly once in each of the four possible positions within a test booklet. The test items were distributed across 13 test booklets in a rotated test design (a balanced incomplete test design) such that each test item appeared in four test booklets. This type of test design ensured a wide coverage of content while at the same time keeping the individual testing burden low. Sampled students were randomly assigned one of the test booklets. Not all test booklets assessed the same domains, however, mathematical literacy was assessed in all test booklets.

Mathematics was the major domain, and 85 items were used to assess student achievement across seven clusters and 13 test booklets. Each test booklet was designed to be of approximately equal difficulty and equivalent content coverage. Due to the balanced design, item parameter estimates are not influenced by a booklet effect. The different locations of domains within each

booklet were expected to lead to booklet influences. Therefore, in PISA 2003, an international test booklet effect was incorporated into the measurement model to correct for the effect of item location. Although a booklet effect parameter can be incorporated in the random item effects model, the response data from eight mathematics items from booklet 1 were analyzed to avoid any booklet effects. A total of 9,796 students were sampled at random from 40 participating countries (excluding Liechtenstein, with 28 students).

The main problem consists of ensuring comparability of test scores across countries, cultures, and educational systems. In the present modeling approach, item and person effects are estimated simultaneously, allowing for between-country differences of item characteristics and student achievement. Therefore, consider the random item effects model

$$\begin{aligned} P\left(Y_{ijk}=1 \mid \theta_{ij}, \tilde{\boldsymbol{\xi}}_k\right) &= \Phi\left(\tilde{a}_{kj}\theta_{ij}-\tilde{b}_{kj}\right), \qquad (7.32)\\ \left(\tilde{a}_{kj}, \tilde{b}_{kj}\right)^t &= (a_k, b_k)^t + \left(\epsilon_{a_{kj}}, \epsilon_{b_{kj}}\right)^t, \\ \theta_{ij} &= \beta_{0j} + e_{ij}, \end{aligned}$$

where $e_{ij} \sim \mathcal{N}\left(0, \sigma^2_{\theta_j}\right)$, $\beta_{0j} \sim \mathcal{N}\left(\gamma_{00}, \tau^2\right)$, and $\left(\epsilon_{a_{kj}}, \epsilon_{b_{kj}}\right)^t \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\xi}}\right)$. This model will be referred to as model $\mathcal{M}_4$ and allows for measurement-noninvariant items and cross-national heterogeneity in latent means and variances. Model $\mathcal{M}_4$ is identified by restricting the sum of the nation-specific item difficulties to zero in each country and by restricting the product of nation-specific item discriminations to one in each country.

In Table 7.1, three models are considered that are nested within the general model $\mathcal{M}_4$ (Equation (7.32)). Model $\mathcal{M}_0$ is an MLIRT model with an empty structural multilevel model. Model $\mathcal{M}_0$ considers that all items exhibit measurement invariance. Cross-national differences in latent means are allowed but differences in latent variances are not. The model was identified by restricting the mean and variance of the latent scale to be zero and one, respectively. Model $\mathcal{M}_1$ allows for country-specific item parameters but disregards cross-national heterogeneity in the latent variable and assumes a standard normal prior for it. Model $\mathcal{M}_1$ is identified since the scale of the latent variable is restricted with mean zero and variance one. Model $\mathcal{M}_2$ allows for nation-specific item parameters and cross-national differences in latent means. The model was identified by restricting the sum of the nation-specific item difficulties to be zero in each country and the variance of the latent variable to be one.

In model $\mathcal{M}_1$, the cross-national heterogeneity in the latent variable is ignored but cross-national differences in item characteristics are allowed. It can be seen that the estimated international item parameters resemble the estimated international item parameters of model $\mathcal{M}_0$. However, there are large cross-national variations in item discrimination and difficulty. The estimated item-specific posterior standard deviations are given under the labels σ_{a_k} and σ_{b_k}. For example, the estimated country-specific difficulty parameters of the

Table 7.1. PISA 2003: Exploring cross-national item variation.

	Invariant		Noninvariant Items						
	Model $\mathcal{M}_0$		Model $\mathcal{M}_1$			Model $\mathcal{M}_2$			
Item	Mean	SD	Mean	SD	σ_{a_k}	Mean	SD	σ_{a_k}	$p_0(a_k)$
Discrimination Parameter									
1	.808	.027	.727	.032	.077	.817	.037	.089	.860
2	1.060	.035	1.076	.064	.284	1.099	.055	.230	.999
3	.729	.024	.611	.029	.068	.717	.029	.066	.567
4	.688	.023	.617	.027	.082	.694	.032	.115	.971
5	.555	.020	.533	.027	.091	.582	.029	.117	.997
6	.367	.025	.343	.036	.102	.402	.044	.160	.999
7	.694	.026	.606	.032	.074	.691	.035	.097	.926
8	.660	.025	.637	.035	.093	.684	.034	.112	.971
Item	Mean	SD	Mean	SD	σ_{b_k}	Mean	SD	σ_{b_k}	$p_0(b_k)$
Difficulty Parameter									
1	−.586	.016	−.581	.066	.384	−.590	.030	.133	.999
2	.185	.016	.172	.075	.452	.194	.036	.115	.963
3	−.040	.014	−.040	.061	.378	−.039	.030	.106	.992
4	−.359	.015	−.377	.058	.342	−.359	.026	.105	.976
5	−.018	.013	−.026	.040	.242	−.024	.022	.082	.727
6	−1.510	.023	−1.549	.046	.241	−1.549	.027	.100	.909
7	−.780	.017	−.785	.058	.343	−.785	.026	.098	.954
8	−.942	.018	−.966	.045	.262	−.957	.022	.082	.756
Structural Part									
Fixed									
γ_{00}	.010	.083	.000	-		.734	.083		
Random									
σ^2_θ	.791	.013	1.000	-		.791	.014		
τ^2_{00}	.269	.063				.270	.063		
Information Criteria									
−2log L	95727.5		102322.1			94681.0			
DIC (p_D)	100395.1(4667)					98960.2(4279)			

first three items are for Turkey .104, .643, and .572, for the United Kingdom −.742, .153, and −.189, and for the Netherlands −.691, .011, and −.275. Model $\mathcal{M}_0$ treats all items as measurement-invariant but captures the heterogeneity across respondents, and it follows that around 25% of the variation in estimated ability is explained by the grouping of students into countries.

The variation in item difficulties is unrealistically large since it includes the cross-national variation in latent means, which is not explicitly modeled. This

follows directly from the estimates of model $\mathcal{M}_2$. The reported estimates of the international item parameters were transformed to a standard scale to allow a direct comparison with the international item parameter estimates of the other models. The estimated within-country and between-country variabilities in student achievement correspond to the estimated values of model $\mathcal{M}_0$. Under model $\mathcal{M}_2$, conditional on the comparable cross-national variation in the latent means, a substantial amount of cross-national variation in item difficulties was found. For model $\mathcal{M}_2$, the country-specific difficulty estimates of items 1–3 of Turkey, .087, .809, and .544, the United Kingdom, −.116, 1.029, and .396, and the Netherlands, .033, 1.052, and .405, differ substantially.

The joint hypothesis of full measurement invariance is tested by comparing the estimated DIC value of model $\mathcal{M}_2$ with that of model $\mathcal{M}_0$. It follows that full measurement invariance is not supported. Under the label $p_0(a_k)$, the posterior probability content of the HPD interval is given, which just includes the event $\tilde{a}_{kj} = a_k$ for $j = 1, \ldots, J$ for each item k. Thus, with 95% confidence it is concluded that the discriminations of items 1, 3, and 7 are invariant. In the same way, invariance of item difficulty does hold for items 5, 6, and 8 with 95% confidence.

The estimated cross-national variations in item discrimination of models $\mathcal{M}_1$ and $\mathcal{M}_2$ are comparable since both models assume factor variance invariance. However, cross-national variations in item discrimination may include cross-national variations in the factor variance. Model $\mathcal{M}_3$ assumes noninvariant factor means and variances and measurement-invariant items. In Table 7.2, the international item parameter estimates of models $\mathcal{M}_3$ and $\mathcal{M}_4$ are given. In the same way as for the item parameter estimates of model $\mathcal{M}_2$, the international item parameter estimates of model $\mathcal{M}_4$ were transformed to a standard scale.

The estimated international item parameters of model $\mathcal{M}_3$ and their standard deviations are comparable with the estimated values of model $\mathcal{M}_0$. However, the population distributions of achievement differ since model $\mathcal{M}_3$ does not assume factor variance invariance. In Table 7.2, the general within-country factor variance, denoted as $\overline{\sigma}^2_\theta$, is presented, which is the mean within-country variation. It follows that this mean level corresponds to the estimated variance component of model $\mathcal{M}_0$. Under model $\mathcal{M}_3$, the estimated between-country variation in mean achievement is slightly lower, which leads to a general intraclass correlation coefficient of .21. A country-specific intraclass correlation coefficient can be computed given the country-specific residual factor variance via $\hat{\tau}^2_{00}/(\hat{\tau}^2_{00} + \hat{\sigma}^2_{\theta_j})$. The intraclass correlation coefficient varies across countries from .160 (Turkey) to .296 (Iceland). The estimated intraclass correlation coefficients for the Netherlands and the United Kingdom are .194 and .180, respectively. The range of estimated intraclass correlation coefficients shows that substantial differences are found over countries in their effects on student achievement. From the estimated DIC values of models $\mathcal{M}_0$ and $\mathcal{M}_3$, it can also be concluded that the assumption of invariant factor variance does not hold when assuming measurement-invariant items.

Table 7.2. PISA 2003: Exploring cross-national item variation and factor variation.

	Invariant Items		Noninvariant Items			
	Model $\mathcal{M}_3$		Model $\mathcal{M}_4$			
Item	Mean	SD	Mean	SD	σ_{a_k}	$p_0(a_k)$
Discrimination Parameter						
1	.803	.027	.732	.045	.085	.436
2	1.068	.034	1.020	.125	.123	.820
3	.729	.025	.631	.038	.070	.177
4	.689	.023	.628	.036	.075	.299
5	.556	.020	.529	.032	.075	.391
6	.366	.025	.328	.063	.056	.093
7	.692	.026	.621	.039	.072	.218
8	.662	.025	.611	.043	.071	.258
Item	Mean	SD	Mean	SD	σ_{b_k}	$p_0(b_k)$
Difficulty Parameter						
1	−.586	.016	−.588	.032	.158	.999
2	.182	.016	.185	.092	.167	.999
3	−.042	.014	−.004	.042	.118	.999
4	−.359	.015	−.367	.027	.126	.999
5	−.018	.013	−.023	.034	.105	.988
6	−1.510	.023	−1.519	.086	.127	.999
7	−.780	.017	−.785	.034	.116	.999
8	−.942	.018	−.956	.039	.092	.932
Structural Part						
Fixed						
γ_{00}	.010	.073	.496	.051		
Random						
$\overline{\sigma}^2_\theta$	.797	.023	.370	.011		
τ^2_{00}	.218	.056	.104	.025		
Information Criteria						
−2log L	94331.2		87277.4			
DIC (p_D)	99984.3(5653.1)		91593.5(4316.1)			

In comparison with model $\mathcal{M}_3$, the estimated international item parameters are comparable, but the estimated posterior standard deviations are higher for model $\mathcal{M}_4$. This additional uncertainty is caused by the cross-national variation in the item parameters. For model $\mathcal{M}_4$, the estimated cross-national variation in discrimination parameters is slightly smaller than that of model $\mathcal{M}_2$. This reduction is caused by allowing noninvariant factor variances. A country-specific common shift in variation in item discrimination is

not allowed and is recognized as a shift in the country-specific factor variance. The difference in estimated DIC values indicates that the hypothesis of factor variance invariance is rejected given noninvariant items.

From the posterior probability contents of the HPD intervals that just include the international item parameter values it follows that all item discriminations are invariant given noninvariant factor variances. Furthermore, with 95% confidence, it is concluded that only the difficulty of item 8 is invariant.

The general model $\mathcal{M}_4$ allowed country-specific differences but retained a common measurement scale. It was shown that country differences are present in the item characteristics and in the students' population distribution. In PISA 2003, items are assumed to be invariant, but the present analysis shows that full measurement invariance is not supported by the data. Ignoring the multilevel structure in the population distribution leads to large cross-national variation in item characteristics. That is, the corresponding estimated item characteristic variation contains the between-country variation in students' abilities, which complicates the statistical inferences and parameter interpretations. The different sources of variation are recognized in model $\mathcal{M}_4$, which makes it possible to detect measurement and structural differences without making unrealistic simplifying assumptions.

Making meaningful comparisons across countries is a complex issue that requires a flexible model that allows for different levels of variation, including cross-national differences. In this light, Goldstein (2004) argued that a multilevel modeling approach is necessary and that country differences are included rather than eliminated in favor of a common measurement scale. This emphasizes the usefulness of the random item effects approach that supports a simultaneous multilevel analysis of response data from an international comparative survey study, accounting for cross-national differences and leading to a common measurement scale.

7.7 Discussion

In many test settings, a test is considered to be fixed and is used for measuring students' abilities. From this point of view, inferences are made with respect to the test that is used without the objective of generalizing the inferences to some population of tests. The test or item parameters are treated as fixed effects. A generalization from the sample of respondents to the population of respondents is preferred, and the person parameters are typically treated as random effects parameters. In this chapter, it is assumed that the item characteristics of the test vary over clusters of respondents. The item parameters are modeled as random item effects such that inferences about the item characteristics can be generalized to the population of clusters.

The MLIRT model introduced in Chapter 6 is extended with a random item effects structure. The random item effects vary over a grouping structure,

which can be induced by a grouping of respondents (e.g., countries, schools) or a grouping of items (e.g., item cloning, item bank). The grouping structure of the student population may differ from the grouping structure of the hierarchical item population. In that case, a cross-classified random effects model can handle the more complex data structure in which responses are cross-classified by two grouping structures (e.g., Raudenbush and Bryk, 2002, Chapter 12).

In this chapter, a typical grouping structure is defined such that there are two sources of between-country variation. At the level of persons, the respondents are nested in countries and there is variation across countries in latent means and variances. At the level of items, there is cross-country variation in item characteristics and the country-specific characteristics are nested in the international item characteristics. The doubly nested structure leads to a complex identification problem that is solved by preventing common shifts of the country-specific item characteristics. Background information can be used to explain the between-country variability.

The MCMC algorithm can also handle polytomous response data, leading to random threshold effects. De Jong et al. (2007) present an application of consumers' susceptibility to normative influence given cross-national polytomous response data. The model allows for cross-national random threshold effects besides cross-national differences in scale usage. Fahrmeir and Tutz (2001) and Tutz and Hennevogl (1996) extended the ordinal cumulative model with random threshold effects that may vary over clusters or subjects. In that case, the preference for response categories can vary across individuals. De Jong and Steenkamp (2009) incorporated a mixture distribution for the random item effects parameters to allow latent class differences. In this approach, unexplained cross-national heterogeneity in random item characteristics is partly captured by latent item class differences. Soares, Gonçalves and Gamerman (2009) used a mixture modeling approach to classify an item as measurement-invariant (anchor item) or that exhibits differential item functioning (DIF item). The invariant as well as the noninvariant item characteristics are all identified and estimated simultaneously with the other model parameters.

7.8 Exercises

7.1. Rasch (1960) assumed that comparisons between persons are invariant over items and comparisons between items are invariant over persons (e.g., Embretson and Reise, 2000).
(a) For the one-parameter logistic response model, show that the difference in log odds for an item k of two persons indexed i and i' can be expressed as

$$\begin{aligned} f\left(P\left(Y_{ik}\right), P\left(Y_{i'k}\right)\right) &= \ln \frac{P\left(Y_{ik}\right)}{1-P\left(Y_{ik}\right)} - \ln \frac{P\left(Y_{i'k}\right)}{1-P\left(Y_{i'k}\right)} \\ &= \left(\theta_i - b_k\right) - \left(\theta_{i'} - b_k\right), \end{aligned}$$

where $P(Y_{ik}) = P(Y_{ik} = 1 \mid \theta_i, b_k)$.
(b) Argue that the comparison in (a) between the two persons is independent of the item k that is used (invariant-person comparison).
(c) Show that the difference in log odds for a person i of two items indexed k and k' can be expressed as

$$f(P(Y_{ik}), P(Y_{ik'})) = (\theta_i - b_k) - (\theta_i - b_{k'}).$$

(d) Argue that the comparison between the two items in (c) is independent of the person i that is used to compare them (invariant-item comparison).

7.2. (continuation of Exercise 7.1) Consider the logistic one-parameter response model for hierarchically structured response data.
(a) Assume that the respondents are nested in groups. Show that the difference in log odds of two persons to item k can be expressed as

$$f(P(Y_{ijk}), P(Y_{i'j'k})) = (\theta_{ij} - b_k) - (\theta_{i'j'} - b_k),$$

where $\theta_{ij} \sim \mathcal{N}(\beta_{0j}, \sigma^2_\theta)$.
(b) Explain that the difference in (a) is independent of the item and that it consists of a between-group as well as a between-individual difference in log odds.
(c) Assume random item difficulty parameters for the noninvariant items. Show in the same way that the difference in log odds of two noninvariant items for a person represents a between-group as well as a between-individual part.

7.3. To improve understanding of the identification issue, assume that response data from one respondent per country were observed. A random item effects model for the latent response data can be stated as

$$Z_{jk} = \beta_{0j} - \tilde{b}_{kj} + \epsilon_{jk},$$

where $\beta_{0j} \sim \mathcal{N}(\gamma_{00}, \tau^2)$, $\tilde{b}_{kj} \sim \mathcal{N}(b_k, \sigma^2_{b_k})$, and $\epsilon_{jk} \sim \mathcal{N}(0, 1)$.
(a) Show that a common shift in the item difficulties of country j can be counterbalanced by a common negative shift in the country mean β_{0j}.
(b) Is it possible to identify this random item effects model by fixing one latent country mean, which would be the standard country?
(c) Explain how the model can be identified via an anchor item.

7.4. Consider the random item effects model in Equation (7.7), but assume fixed item discriminations, $\tilde{a}_{kj} = 1$, for each j and k. Let $\theta_i \sim \mathcal{N}(0, 1)$ and $\tilde{b}_{kj}$ be normally distributed with mean zero and standard deviation $\sigma_{b_k} = .15$. Use Figure 7.2 to answer the following questions.
(a) For $\theta_i = 0$, explain and determine the vertical distance between the item characteristic curves when b_h and b_l are the upper and lower limits of the 95% density interval of difficulty parameter $\tilde{b}_{kj}$.

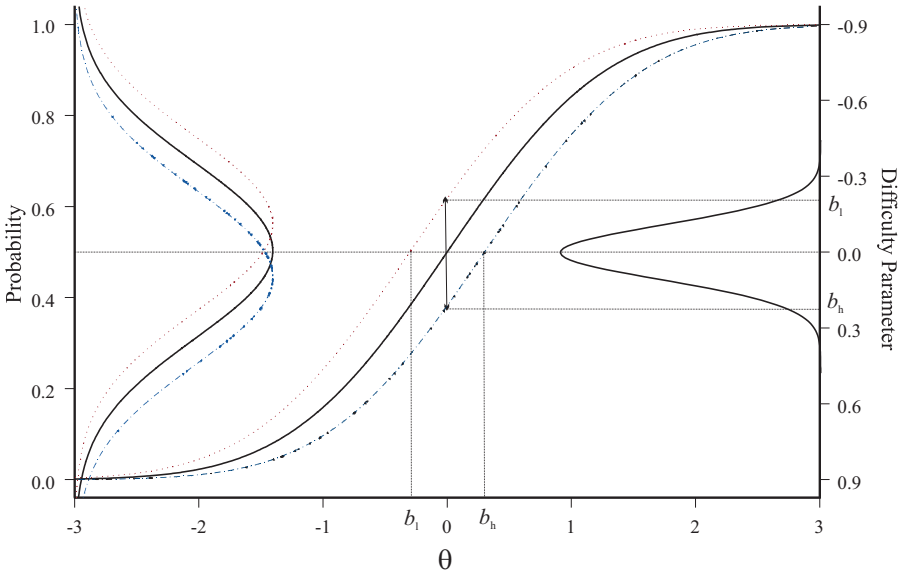

Fig. 7.2. Item characteristic curves for different difficulty values from the random effects distribution (right vertical axis). The probability of a correct response (left vertical axis) is given as a function of the ability parameter (horizontal axis) for item difficulty values from the random difficulty distribution (right vertical axis).

(b) For a success probability of .5, explain and determine the horizontal difference in item characteristic curves.
(c) For a constant ability level, does a linear change in item difficulty lead to a linear change in the probability of success?
(d) The corresponding densities of probabilities are graphed to the right of the left vertical axis. Explain the shift in means of the densities.
(e) A population average item characteristic curve can be defined as the expected success probability as a function of the ability, where the expectation is taken over the density of the random difficulty parameter; see Equation (7.10). Why will the average success probabilities be shrunk towards .5? Outline the population average characteristic curve.

7.5. Consider the combined random item effects model for binary response data as stated in Equation (7.12).
(a) Show that the total variance of Z_{ijk} equals

$$\mathrm{Var}\,(Z_{ijk}) = \sigma^2_{a_k}\sigma^2_\theta + \sigma^2_{a_k}\mu^2_\theta + a^2_k\sigma^2_\theta + \sigma^2_{b_k} + 1.$$

Use the iterated conditional variance identity to handle the product of two normally distributed variables:

$$\mathrm{Var}\,(\tilde{a}_{kj}\theta_i) = E\,(\mathrm{Var}\,(\tilde{a}_{kj}\theta_i \mid \theta_i)) + \mathrm{Var}\,(E\,(\tilde{a}_{kj}\theta_i \mid \theta_i))\,.$$

(b) Derive the intraclass correlation coefficient given in Equation (7.13).
(c) The combined random item effects model is extended with independent normally distributed priors for the (international) item parameters, $a_k \sim \mathcal{N}\left(\mu_a, \sigma_a^2\right)$ and $b_k \sim \mathcal{N}\left(\mu_b, \sigma_b^2\right)$. Show that the total variance of Z_{ijk} equals

$$\mathrm{Var}\left(Z_{ijk}\right) = \sigma_a^2\sigma_\theta^2 + \sigma_a^2\mu_\theta^2 + \mu_a^2\sigma_\theta^2 + \sigma_{a_k}^2\sigma_\theta^2 + \sigma_{a_k}^2\mu_\theta^2 + \sigma_{b_k}^2 + \sigma_b^2 + 1.$$

(d) Derive three different kinds of intraclass correlation coefficients: (1) the correlation between latent responses to item k of different respondents from the same country, (2) the correlation between latent responses to item k from different respondents from different counties, and (3) the correlation between latent responses to different items from the same respondent.

7.6. Implement an M-H step for the variance parameter $\sigma_{\kappa_k}^2$ using an inverse gamma prior. The conditional posterior density of the threshold variance parameter $\sigma_{\kappa_k}^2$ can be stated as

$$p\left(\sigma_{\kappa_k}^2 \mid \tilde{\boldsymbol{\kappa}}_k, \boldsymbol{\kappa}_k\right) \propto \prod_j p\left(\tilde{\boldsymbol{\kappa}}_{kj} \mid \boldsymbol{\kappa}_k, \sigma_{\kappa_k}^2\right) p\left(\sigma_{\kappa_k}^2\right).$$

(a) Show that the full conditional posterior distribution does not reduce to standard form due to order constraints.
(b) Define a suitable proposal distribution for generating candidates.
(c) Define the corresponding acceptance ratio, and summarize the M-H step in an algorithmic form.

8

Response Time Item Response Models

Response times and responses can be collected via computer adaptive testing or computer-assisted questioning. Inferences about test takers and test items can therefore be based on the response time and response accuracy information. Response times and responses are used to measure a respondent's speed of working and ability using a multivariate hierarchical item response model. A multivariate multilevel structural population model is defined for the person parameters to explain individual and group differences given background information. An application is presented that illustrates novel features of the model.

8.1 Mixed Multivariate Response Data

Nowadays, response times (RTs) are easily collected via computer adaptive testing or computer-assisted questioning. The RTs can be a valuable source of information on test takers and test items. The RT information can help to improve routine operations in testing such as item calibration, test design, detection of cheating, and adaptive item selection.

The collection of multiple item responses and RTs leads to a set of mixed multivariate response data since the individual item responses are often observed on an ordinal scale, whereas the RTs are observed on a continuous scale. The observed responses are imperfect indicators of a respondent's ability. When measuring a construct such as ability, attention is focused on the accuracy of the test results. The observed RTs are indicators of a respondent's speed of working, and speed is considered to be a different construct. As a result, mixed responses are used to measure the two constructs ability and speed.

Although response speed and response accuracy measure different constructs (Schnipke and Scrams, 2002, and references therein), the reaction-time research in psychology indicates that there is a relationship between response

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_8,

speed and response accuracy (Luce, 1986). This relationship is often characterized as a speed–accuracy trade-off. A person can decide to work faster, but this will lead to a lower accuracy. The trade-off is considered to be a within-person relationship: a respondent controls the speed of working and accepts the related level of accuracy. It will be assumed that each respondent chooses a fixed level of speed, which is related to a fixed accuracy.

A hierarchical measurement model was proposed by van der Linden (2007) to model RTs and dichotomous responses simultaneously that accounts for different levels of dependency. The different stages of the model capture the dependency structure of observations nested within persons at the observational level and the relationship between speed and ability at the individual level. Klein Entink, Fox and van der Linden (2009a), and Fox, Klein Entink and van der Linden (2007) extended the model for measuring accuracy and speed (1) to allow time-discriminating items, (2) to handle individual and/or group characteristics, and (3) to handle the nesting of individuals in groups. This extension has a multivariate multilevel structural population model for the ability and the speed parameters that can be considered a multivariate extension of the structural part of the MLIRT model of Chapter 6. In this chapter, the complete modeling framework will be discussed, and an extension is made to handle polytomous response data.

The RTs and responses have been modeled from different viewpoints. Scheiblechner (1979) and Maris (1993) explicitly modeled RTs separately from the responses. They both focused on uncomplicated cognitive tasks, where it is assumed that most items would be solved correctly were there enough time, and excluded accuracy scores. Among others, Roskam (1997) and Verhelst, Verstralen and Jansen (1997) developed a regular item response model with time parameters added to model RTs and responses. Thissen (1983) developed a timed-testing model for RTs that contains a response accuracy term. This leads to a log-linear relationship between response speed and response accuracy. Schnipke and Scrams (2002) give an overview of RT models for test items, and van der Linden (2007) discusses the main differences between a multilevel modeling perspective for RTs and single-level RT models for test items.

8.2 Measurement Models for Ability and Speed

An item-based test is used to measure ability as an underlying construct that cannot be observed directly. In general, item response models (as discussed in Chapter 4) are adopted to make inferences from the multivariate item response data since they describe the probability of a correct response to an item as a function of ability and item characteristics. The notation remains the same: the observation of item k ($k = 1, \ldots, K$) of person i ($i = 1, \ldots, n_j$) in group j ($j = 1, \ldots, J$) is denoted as y_{ijk}.

Various families of distributions have been employed in psychometric applications to model RT data; e.g., Poisson, gamma, Weibull, and lognormal distributions (see Maris, 1993; Schnipke and Scrams, 1997; Thissen, 1983; van Breukelen, 2005; van der Linden, 2006). Here, the log of the RTs is assumed to be normally distributed. Note that RT distributions are skewed to the right since RTs have a natural lower bound at zero. In Section 8.5, it will be shown that RTs combined with latent continuous item responses can be treated in a multivariate model, which simplifies the statistical inferences.

Let T_{ijk} denote the log-RT of person i in group j on item k. Then, T_{ijk} is normally distributed, with a mean depending on the speed at which the person works denoted as ζ_{ij}. Speed is assumed to be the underlying construct that is measured via the item RTs. It is assumed that persons work with a constant speed during a test. A higher speed of working leads to lower RTs, which is recognized by defining a negative linear relationship between the log-RTs and the speed parameter.

Items have different time intensities, which means that they differ in the time that is required to obtain their solution at a constant level of speed. The time intensity of item k is given by λ_k. A higher λ_k indicates that item k is expected to consume more time. In Figure 8.1, the item characteristic curves (ICCs) of two dichotomous items are given, where the left y-axis presents the probability of a correct response as a function of the ability parameter (x-axis). The items have different levels of difficulty since their success probabilities differ at a constant level of ability. An increase in ability leads to an increase in the success probability. The increase is the same for both items since the discrimination parameters are the same. Now, let the person parameter on the x-axis denote the speed of working. For both items, the expected RT is given (on the right y-axis) as a function of the speed parameter. The so-called RT characteristic curve (RTCC) shows that the RTs decrease with speed. It can be seen that at a constant speed the RTs of both items differ, which indicates that the items have different time intensities. The upper RTCC corresponds to the most time-intensive item. Furthermore, an increase in the speed leads to a similar decrease in RTs.

The effect of changing the level of speed may vary across items. Items may differ with respect to the decrease in RT when increasing the level of speed. Therefore, an item characteristic time-discrimination parameter φ_k is introduced. In Figure 8.2, ICCs and RTCCs of two items are given with different discrimination and time-discrimination parameters, respectively. It can be seen that the difference in expected RTs between persons working at different speed levels is less for the lower time-discriminating item. The same holds for the ICCs. The less discriminating item shows a smaller difference in success probabilities for persons with different abilities. The difficulty and time-intensity parameters are equal for both items ($b = 0, \lambda = 4$).

The model for measuring a person's speed level from the RTs resembles the two-parameter item response model for measuring ability. This measurement model also has two item characteristic parameters, the time intensity and

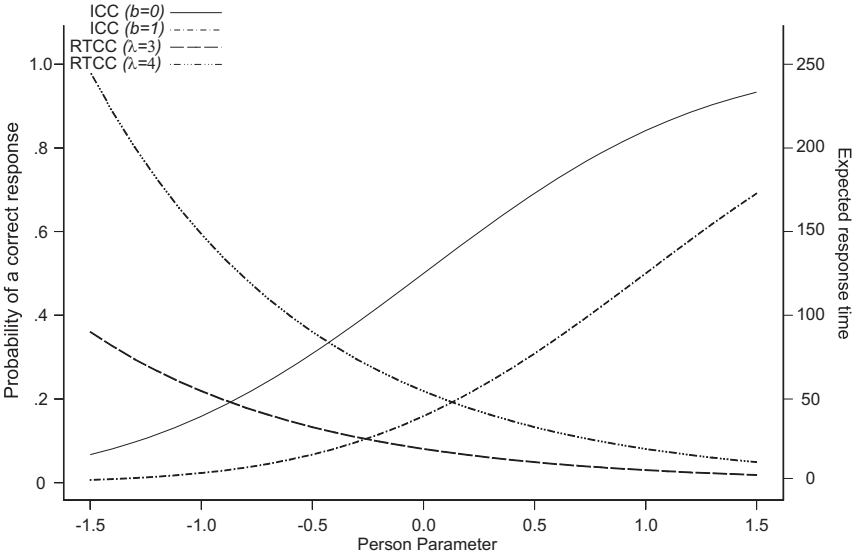

Fig. 8.1. Item characteristic curves (ICCs) and response time characteristic curves (RTCCs) for two items with equal discrimination parameters.

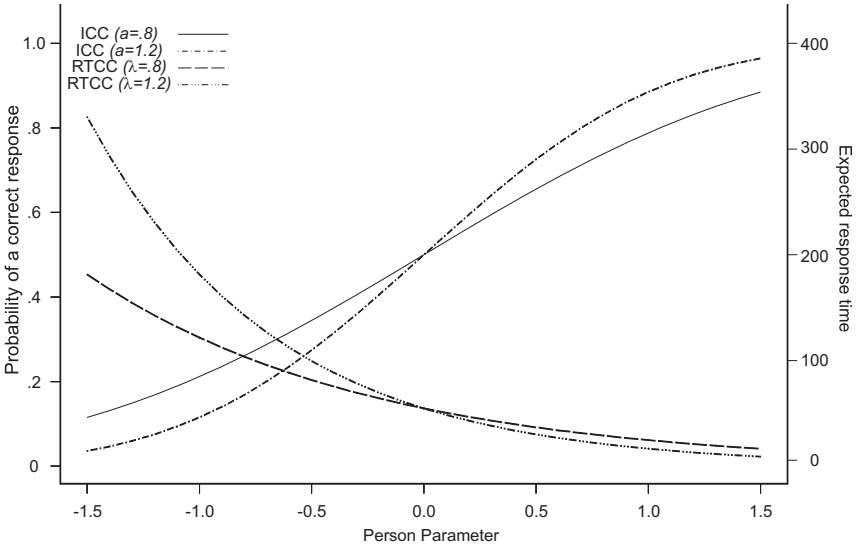

Fig. 8.2. Item characteristic curves (ICCs) and response time characteristic curves (RTCCs) for two items with equal difficulty and time-intensity parameters.

the time-discrimination parameter, and a (unidimensional) underlying latent variable. The latent variable is also nonlinearly related to the observations. The correspondence is closest for continuous observed responses; for these,

see, for instance, Mellenbergh (1994b) and Shi and Lee (1998). The model is referred to as the item response time model and is given by

$$T_{ijk} = \lambda_k - \varphi_k \zeta_{ij} + \epsilon_{\zeta_{ijk}}, \tag{8.1}$$

where $\epsilon_{\zeta_{ijk}} \sim \mathcal{N}\left(0, \omega_k^2\right)$. Notice that the interpretation of the model parameters in Equation (8.1) results in a different location of the minus sign compared with the two-parameter item response model. In Figures 8.1 and 8.2, the expected RTs are plotted, which are nonlinearly (exponentially) related to the speed parameter. The expected log-RTs are linearly related to the speed parameter; see Equation (8.1). Then, the corresponding log-RT characteristic curve is a straight line.

A person's observed RTs are assumed to be independent given the level of speed. This conditional independence assumption can be stated as

$$p\left(\mathbf{t}_{ij} \mid \zeta_{ij}\right) = \prod_k p\left(t_{ijk} \mid \zeta_{ij}\right). \tag{8.2}$$

In a similar way, the item responses of a person are conditionally independent given the level of ability; see Equation (1.1). The two conditional independence assumptions induce a third conditional independence assumption. A person's item response and item response time are conditionally independent given the level of ability and speed. As a result, the following factorizations can be made

$$\begin{aligned} p\left(\mathbf{y}_{ij}, \mathbf{t}_{ij} \mid \theta_{ij}, \zeta_{ij}\right) &= \prod_k p\left(y_{ijk}, t_{ijk} \mid \theta_{ij}, \zeta_{ij}\right) \\ &= \prod_k p\left(y_{ijk} \mid \theta_{ij}\right) p\left(t_{ijk} \mid \zeta_{ij}\right) \\ &= p\left(\mathbf{y}_{ij} \mid \theta_{ij}\right) \prod_k p\left(t_{ijk} \mid \zeta_{ij}\right) \\ &= p\left(\mathbf{y}_{ij} \mid \theta_{ij}\right) p\left(\mathbf{t}_{ij} \mid \zeta_{ij}\right). \end{aligned} \tag{8.3}$$

In the second step, the induced conditional independence assumption is used. In the third and fourth steps, the conditional independence assumptions of the item response model and the item response time model are used, respectively.

8.3 Joint Modeling of Responses and Response Times

The total variation in observed RTs and responses can be partitioned into variations due to (i) response variation in item RTs and item responses, (ii) the sampling of persons and items, and (iii) the sampling of groups. The different sources that contribute to the variation between item responses and item RTs are modeled via random effects. The three observed levels of hierarchy lead to different stages of the model.

Response variation is modeled via an item response model and an item response time model, and they define level 1 of the joint model. The measurement models define a probabilistic relationship between the RTs and responses and the underlying person parameters. The normal ogive item response model is used since the underlying continuous responses (Equation (4.7)) with the corresponding log-RTs are multivariate normally distributed. Then, a multivariate normal joint measurement model can be constructed for the ability and speed parameters, which will simplify the statistical inferences.

8.3.1 A Structural Multivariate Multilevel Model

At level 2, the ability and speed parameters are considered as outcome variables of a multivariate multilevel model. They are modeled to have a multivariate normal distribution. This allows the specification of a within-person correlation structure for the ability and speed parameters. Between-person differences in ability and speed can be modeled by covariates $\mathbf{x}$. Subsequently, group differences between the latent outcome variables are explained as a function of group-level covariates $\mathbf{w}$ at a third level. The higher-level regression structure for the latent person parameters makes it possible to partition their total variance into within-group and between-group components. This way, inferences can be made about the person parameters for different groups simultaneously given the explanatory variables.

The latent outcomes (speed and ability) of respondent i in group j are linearly related to $\mathbf{x}_{ij} = \left(\mathbf{x}_{1ij}^t \oplus \mathbf{x}_{2ij}^t\right)$. The regression effects, $\boldsymbol{\beta}_j = \text{vec}\left(\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}\right)$ are allowed to vary across groups where the matrix operation vec creates a column vector by stacking all columns under each other. Now, the structural multivariate multilevel model can be expressed as

$$[\theta_{ij}, \zeta_{ij}] = \left[\mathbf{x}_{1ij}^t, \mathbf{x}_{2ij}^t\right][\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}] + \left[e_{\theta_{ij}}, e_{\zeta_{ij}}\right], \tag{8.4}$$

$$[\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}] = \left[\mathbf{w}_{1j}^t, \mathbf{w}_{2j}^t\right][\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2] + [\mathbf{u}_{1j}, \mathbf{u}_{2j}]. \tag{8.5}$$

The regression coefficients are specified as random but can be restricted to be common to all groups. The error terms at each level are assumed to be multivariate normally distributed, and the level-2 error terms are assumed to be independent of the level-3 error terms.

The structural multivariate outcome variables can be stacked in a vector. Let $\boldsymbol{\Omega}_{ij} = (\theta_{ij}, \zeta_{ij})^t$. It follows that

$$\boldsymbol{\Omega}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}_j + \mathbf{e}_{ij}, \tag{8.6}$$

$$\boldsymbol{\beta}_j = \mathbf{w}_j\boldsymbol{\gamma} + \mathbf{u}_j, \tag{8.7}$$

where $\mathbf{w}_j = (\mathbf{w}_{1j} \oplus \mathbf{w}_{2j})$ and $\mathbf{e}_{ij} = \left(e_{\theta_{ij}}, e_{\zeta_{ij}}\right)^t$ is multivariate normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_P$. The covariance matrix $\boldsymbol{\Sigma}_P$ equals

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix}.$$

The covariance parameter ρ represents the within-person dependency between speed and ability, and the person parameters speed and ability are independent for $\rho = 0$. This independency is conditional on the structure of the item parameters since such a dependency can also be modeled via a covariance structure on the item parameters (Klein Entink et al., 2009b). When ρ is positive, persons who work faster on average are expected to have above-average abilities. The group-level error term, $\mathbf{u}_j$, is assumed to be multivariate normally distributed with mean zero and covariance matrix $\mathbf{V}$. More stable covariance parameter estimates can be obtained by restricting this covariance matrix to be block-diagonal with diagonal matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. In this case, the random effects in the regression of $\boldsymbol{\theta}$ on $\mathbf{x}_1$ are allowed to correlate but are independent of those in the regression of $\boldsymbol{\zeta}$ on $\mathbf{x}_2$.

Let $\boldsymbol{\Omega}_j$ be a single vector that contains the vector of individual abilities and the vector of speed levels stacked on top of each other. Then, the structural multilevel model for group j becomes

$$\begin{aligned}\boldsymbol{\Omega}_j &= \mathbf{x}_j\boldsymbol{\beta}_j + \mathbf{e}_j \\ &= \mathbf{x}_j\mathbf{w}_j\boldsymbol{\gamma} + \mathbf{x}_j\mathbf{u}_j + \mathbf{e}_j,\end{aligned} \tag{8.8}$$

where $\mathbf{x}_j = (\mathbf{x}_{1j} \oplus \mathbf{x}_{2j})$ and $\mathbf{e}_j \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_P \otimes \mathbf{I}_{n_j}\right)$. Note that the structural multilevel part in Equation (8.8) is closely related to the structural multilevel part of the MLIRT model (Equation (6.9)) but the latter assumes independence between the latent outcomes given the random group effects.

Marginalizing over the random regression effects in Equation (8.8), the conditional covariance structure of $\boldsymbol{\Omega}_j$ becomes

$$\text{Cov}\left(\boldsymbol{\Omega}_j \mid \mathbf{x}_j, \mathbf{w}_j\right) = \mathbf{x}_j\mathbf{V}\mathbf{x}_j^t + \boldsymbol{\Sigma}_P \otimes \mathbf{I}_{n_j}. \tag{8.9}$$

The covariance structure consists of a component that deals with the between-group heterogeneity and the second component represents the within-person covariance structure. This within-person structure is assumed to be common across individuals and groups.

The structural parts in Equations (8.6) and (8.7) are referred to as level-2 and level-3 models, respectively. The structural component of the model allows a simultaneous regression analysis of all person parameters on explanatory variables at the individual and group levels while taking into account the dependencies between the individuals within each group. As a result, among other things, conclusions can be drawn as to the size of the effects of the explanatory variables on the test takers' ability and speed as well as the correlation between these person parameters. Note that hypotheses on these effects can be tested simultaneously.

The model can be used for various analyses. First, the analysis might focus on the item parameters; more specifically, the relationships between the characteristics of the items in the domain covered by the test. For example, the correlation between the time intensity and difficulty parameters of the items can be investigated. Second, the analysis could be about the structural

relationships between explanatory information at the individual and/or group levels and the test takers' ability and speed. For example, the variance components of the structural model might help in exploring the partitioning of the variance of the speed parameters across the different levels of analysis. Third, interest might be in the random effects in the model (e.g., to identify atypical individuals or groups with respect to their ability or speed).

8.3.2 The RTIRT Likelihood Model

The likelihood part of the model consists of three stages. At level 1, the simultaneous measurement framework for responses and RTs is defined. At level 2, a structural multivariate multilevel model is defined for both underlying latent variables. The structural multivariate multilevel model consists of two stages. The three stages will be referred to as the likelihood part of the RTIRT (response time item response theory) model. A more comprehensive name could also stress the integration of both measurement models for measuring speed and ability with the structural multivariate multilevel model that explains differences at the different levels of ability and speed.

In Figure 8.3, a path diagram of the likelihood part of the RTIRT model for responses and RTs is given. The two ellipses represent the measurement models for the multivariate outcome variables θ_{ij} and ζ_{ij}. Note that the measurement models have different functional forms. In this case, a lognormal model is used for the continuous RTs (upper ellipse) and a normal ogive model is used for the discrete responses (lower ellipse). The item parameters of both measurement models, $\mathbf{\Lambda}_{1k}$ and $\mathbf{\Lambda}_{2k}$, are allowed to correlate to capture within-item correlations.

Three boxes are plotted, where the box indexed as items is plotted in the box individual, which is plotted in the box group to illustrate the nested structure of the data. The individual and group boxes constitute the structural multivariate multilevel model for $\mathbf{\Omega}_{ij}$. In this structural part of the model, the latent variables are allowed to correlate to model unexplained within-individual correlation between ability and speed. It follows that it is possible to construct an underlying within-item and/or within-person correlation structure to deal with the correlations between RTs and responses.

There are three levels of uncertainty: (1) at the observation level, (2) at the individual level, and (3) at the group level. In this multivariate case, the measurement models quantify measurement error corresponding to the nonlinear relationship between responses and the individual abilities, and RTs and the individual speed levels. The explanatory information $\mathbf{x}_{ij}$ and $\mathbf{w}_j$ at the individual and group levels explains simultaneously variability in the ability and speed levels within and between groups.

The likelihood of the RTIRT model can be obtained by integrating over the random effects. It follows that

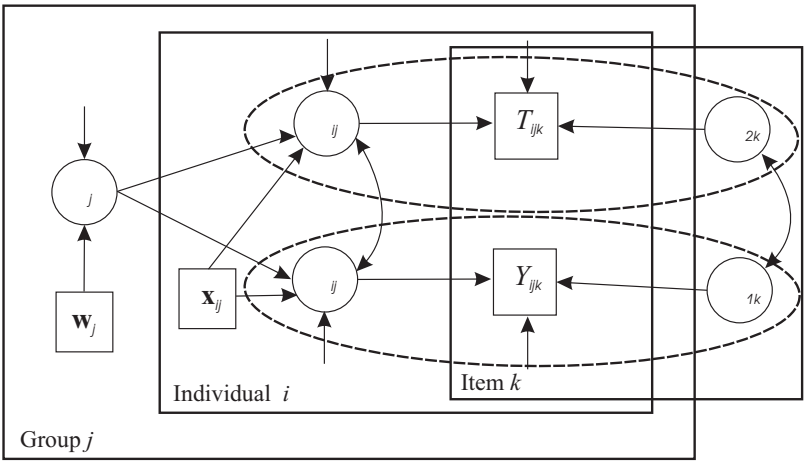

Fig. 8.3. Path diagram for the RTIRT model.

$$p(\mathbf{y}, \mathbf{t} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}_P, \mathbf{V}) =$$
$$\prod_j \int \left[\int \int \prod_{i|j} p(\mathbf{y}_{ij} \mid \theta_{ij}, \mathbf{a}, \mathbf{b})\, p(\mathbf{t}_{ij} \mid \zeta_{ij}, \boldsymbol{\varphi}, \boldsymbol{\lambda}) \cdot \right.$$
$$\left. p(\theta_{ij}, \zeta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j, \boldsymbol{\Sigma}_P)\, \mathrm{d}\theta_{ij} \mathrm{d}\zeta_{ij} \right] p(\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{V})\, \mathrm{d}\boldsymbol{\beta}_j. \qquad (8.10)$$

At the observation level, the product of distributions of the observations follows from the conditional independence assumptions; see Equation (8.3). This represents the measurement part of the model. The other levels describe the population distributions of the person parameters and random group effects.

8.4 RTIRT Model Prior Specifications

Prior distributions are to be specified for the fixed effect parameters $\boldsymbol{\gamma}$, the item parameters, and the covariance parameters $(\boldsymbol{\Sigma}_P, \mathbf{V})$. A multivariate normal prior distribution with mean $\boldsymbol{\gamma}_0$ and covariance matrix $\boldsymbol{\Sigma}_\gamma$ is a conjugate prior for the fixed effects parameters. Attention is focused on a prior for $\boldsymbol{\Sigma}_P$ with a model identifying restriction. Furthermore, an item prior structure is defined for the measurement models that allows for within-item correlations.

8.4.1 Multivariate Prior Model for the Item Parameters

The exchangeable prior distribution for the item parameters of both measurement models is specified such that, for each item k, the vector $\boldsymbol{\Lambda}_k = (a_k, b_k, \varphi_k, \lambda_k)$ is assumed to follow a multivariate normal distribution with

mean vector $\boldsymbol{\mu}_I = (\mu_a, \mu_b, \mu_\varphi, \mu_\lambda)$ and covariance matrix $\boldsymbol{\Sigma}_I$. The assumption introduces a common within-item correlation structure. For example, it may be expected that easy items require less time to be solved than more difficult items. If so, the time intensity parameter correlates positively with the item difficulty parameter. The guessing parameter of the response model has no analogous parameter in the item response time model (since there is no guessing aspect for the RTs). Therefore, it does not serve a purpose to include it in this multivariate model, and an independent prior for the guessing parameter is specified.

8.4.2 Prior for $\boldsymbol{\Sigma}_P$ with Identifying Restrictions

The model is identified by fixing the metric of the two latent person parameters ability and speed. The metric of the ability parameter is identified by fixing the mean and the variance. This is done by fixing the general mean to zero, $\mu_\theta = 0$, and the variance, σ^2_θ, to one. The RTs have a natural unit. However, there is an indeterminacy in the mean (variance) that is determined via the item intensity (time-discrimination) parameters and the mean (variance) of the latent speed parameter. The variance of the metric is identified via the restriction $\prod \varphi_k = 1$. The mean of the metric is identified by fixing the general mean to zero, $\mu_\zeta = 0$.

Within an MCMC algorithm, both general means are easily restricted to zero by transforming in each iteration the sampled vector of ability and speed values in such a way that each transformed sample has a mean of zero. Subsequently, a prior is chosen such that $\sigma^2_\theta = 1$ with probability one, and the covariance matrix of the latent person parameters $(\theta_{ij}, \zeta_{ij})^t$ equals

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} 1 & \rho \\ \rho & \sigma^2_\zeta \end{bmatrix}. \tag{8.11}$$

Note that a multivariate probit model is identified by fixing the diagonal elements of the covariance matrix (Chib and Greenberg, 1995) but that, because of the special nature of the RTs, in the current case only one element of $\boldsymbol{\Sigma}_P$ has to be fixed.

In general, there are two issues when restricting a covariance matrix. First, defining proper priors for a restricted covariance matrix is rather difficult. For example, for the conjugate inverse Wishart prior, there is no choice of parameter values that reflects a restriction on the variance of the ability parameter such as that above. For the multinomial probit model, McCulloch, Polson and Rossi (2000) tackled this problem by specifying proper diffuse priors for the unidentified parameters and reporting the marginal posterior distributions of the identified parameters. However, it is hard to specify prior beliefs about unidentified parameters. Second, for a Gibbs sampler, sampling from a restricted covariance matrix requires extra attention. Chib and Greenberg (1995) defined individual priors on the free covariance parameters but, as a

result, the augmented data had to be sampled from a special truncated region and the values of the free covariance parameter could only be sampled using an M-H step. However, such steps involve the specification of an effective proposal density with tuning parameters that can only be fixed through a cumbersome process.

A general approach for sampling from a restricted covariance matrix can be found in Browne (2006), but it also is based on an M-H algorithm. For completeness, there is an alternative approach. Barnard, McCullogh and Meng (2000) formulated a prior directly for the identified parameters. In order to do so, they factored the covariance matrix into a diagonal matrix with standard deviations and a correlation matrix and specified an informative prior for the latter. This prior was then incorporated into a Griddy-Gibbs sampler. However, such algorithms can be slow and require the choice of a grid size and boundaries. Boscardin and Zhang (2004) followed a comparable approach but used a parameter-extended M-H algorithm for sampling values from the conditional distribution of the correlation matrix.

In the present approach, the conditional distribution of ζ_{ij} given θ_{ij} has a normal density,

$$\zeta_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j, \rho, \sigma_\zeta^2 \sim \mathcal{N}\left(\mathbf{x}_{2ij}^t \boldsymbol{\beta}_{2j} + \rho(\theta_{ij} - \mathbf{x}_{1ij}^t \boldsymbol{\beta}_{1j}), \tilde{\sigma}_\zeta^2\right),$$

where $\tilde{\sigma}_\zeta^2 = \sigma_\zeta^2 - \rho^2$. Parameter ρ can be viewed as the regression parameter in a normal regression problem of ζ_{ij} on θ_{ij} with variance $\tilde{\sigma}_\zeta^2$. Specifying conjugate priors for these parameters,

$$\begin{aligned} \rho &\sim \mathcal{N}\left(\rho_0, \sigma_\rho^2\right), \\ \tilde{\sigma}_\zeta^2 &\sim \mathcal{IG}\left(g_1, g_2\right), \end{aligned}$$

the full conditional posterior densities become

$$\rho \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\beta} \sim \mathcal{N}\left(\Delta\left(\rho_0 \sigma_\rho^{-2} + \tilde{\sigma}_\zeta^{-2}\left(\boldsymbol{\theta} - \mathbf{x}_1\boldsymbol{\beta}_1\right)^t\left(\boldsymbol{\zeta} - \mathbf{x}_2\boldsymbol{\beta}_2\right)\right), \Delta\right), \quad (8.12)$$

$$\tilde{\sigma}_\zeta^2 \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\beta}, \rho \sim \mathcal{IG}\left(g_1 + N/2, g_2 + \Xi^t\Xi/2\right), \quad (8.13)$$

where

$$\begin{aligned} \Delta^{-1} &= \tilde{\sigma}_\zeta^{-2}\left(\boldsymbol{\theta} - \mathbf{x}_1\boldsymbol{\beta}_1\right)^t\left(\boldsymbol{\theta} - \mathbf{x}_1\boldsymbol{\beta}_1\right) + \sigma_\rho^{-2}, \\ \Xi &= \left(\boldsymbol{\zeta} - \mathbf{x}_2\boldsymbol{\beta}_2\right) - \rho\left(\boldsymbol{\theta} - \mathbf{x}_1\boldsymbol{\beta}_1\right). \end{aligned}$$

When implementing an MCMC sampler, random draws of the elements of covariance matrix $\boldsymbol{\Sigma}_P$ in Equation (8.11) are constructed from the samples drawn from (8.12) and (8.13).

The (symmetric) covariance matrix constructed from the sampled values is positive definite. The determinant of $\boldsymbol{\Sigma}_P$ equals $|\boldsymbol{\Sigma}_P| = \sigma_\zeta^2 - \rho^2 = \tilde{\sigma}_\zeta^2$ and $\tilde{\sigma}_\zeta^2 > 0$, and it follows that the determinant $|\boldsymbol{\Sigma}_P| > 0$. The latter is sufficient to guarantee that symmetric matrix $\boldsymbol{\Sigma}_P$ is positive definite and that it has an inverse.

The key element of the present approach is the specification of a proper prior distribution for the covariance matrix with one fixed diagonal element and the construction of random draws from the matrix of the corresponding conditional posterior distribution. The draws will show more autocorrelation due to this new parameterization. This implies that more MCMC iterations are needed to cover the support of the posterior distribution adequately, a measure that only involves a (linear) increase in the running time of the sampler. On the other hand, convergence of the algorithm is easily established without having to specify any tuning parameter.

8.5 Exploring the Multivariate Normal Structure

A linear relationship is established between the new augmented variable Z_{ijk} and the ability parameter via data augmentation. For binary response data, the augmentation step is defined in Equation (4.7), and for polytomous response data, it is defined in Equation (4.25). Note that an additional sampling step is introduced when using the three-parameter item response model. The normally distributed augmented response data and the log-RTs are considered to be outcome variables. The dependent level-1 measurements are linearly related to the latent person parameters and can be analyzed jointly as multivariate data.

Statistical inferences can be made from the complete data using the factorization

$$p\big(\mathbf{y},\mathbf{t},\mathbf{z},\mathbf{s} \mid \boldsymbol{\Lambda},\mathbf{c},\boldsymbol{\gamma},\boldsymbol{\Sigma}_P,\mathbf{V}\big) = p\big(\mathbf{y} \mid \mathbf{z},\mathbf{s}\big)p(\mathbf{s}|\mathbf{c})p\big(\mathbf{z},\mathbf{t} \mid \boldsymbol{\Lambda},\boldsymbol{\gamma},\boldsymbol{\Sigma}_P,\mathbf{V}\big) \quad (8.14)$$

where $\mathbf{s}$ denotes the augmented data for knowing or not knowing the correct answer. Interest is focused on exploring the relationships between ability and speed. Therefore, the term on the far right-hand side of (8.14) will be explored in more detail. This likelihood can be taken to be that of a normal multivariate multilevel model,

$$p\big(\mathbf{z},\mathbf{t} \mid \boldsymbol{\Lambda},\boldsymbol{\gamma},\boldsymbol{\Sigma}_P,\mathbf{V}\big) = \int\int\int p\big(\mathbf{z} \mid \boldsymbol{\theta},\mathbf{a},\mathbf{b}\big)p\big(\mathbf{t} \mid \boldsymbol{\zeta},\boldsymbol{\varphi},\boldsymbol{\lambda}\big) \\ p\big(\boldsymbol{\zeta},\boldsymbol{\theta} \mid \boldsymbol{\beta},\boldsymbol{\Sigma}_P\big)p\big(\boldsymbol{\beta} \mid \boldsymbol{\gamma},\mathbf{V}\big)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\zeta}\mathrm{d}\boldsymbol{\beta}. \quad (8.15)$$

The first two factors in this decomposition occur because of the independence of the responses and RTs given the latent person parameters. The last two factors represent the structural multivariate multilevel part of the model.

The augmented response data and log-RTs are stacked in one vector to explore simultaneously the within-item and between-item relations for ability and speed. For binary response data, the vector of augmented data $\mathbf{Z}_{ij} = (Z_{ij1},\ldots,Z_{ijK})^t$ plus the vector of difficulty parameters, $\mathbf{b}$, and the similar vector of log-RTs $\mathbf{T}_{ij} = (T_{ij1},\ldots,T_{ijK})^t$ minus the vector of time intensity

parameters, $\boldsymbol{\lambda}$, are stacked in a vector $\mathbf{Z}^*_{ij}$. Then, both measurement models can be presented as a linear regression structure,

$$\begin{aligned}\mathbf{Z}^*_{ij} &= (\mathbf{a} \oplus -\boldsymbol{\varphi})\,(\theta_{ij}, \zeta_{ij})^t + \boldsymbol{\epsilon}_{ij} \\ &= \mathbf{H}_P \boldsymbol{\Omega}_{ij} + \boldsymbol{\epsilon}_{ij},\end{aligned} \tag{8.16}$$

where $\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}_K \oplus \mathbf{I}_K\boldsymbol{\omega}\big)$. A similar linear relationship can be found for polytomous data using the augmentation step in Equation (4.25), and in that case, $\mathbf{Z}^*_{ij} = \text{vec}\,(\mathbf{Z}_{ij}, \mathbf{T}_{ij} - \boldsymbol{\lambda})$.

Inferences from this multivariate model are simplified by taking advantage of some of the properties of the multivariate normal distribution. For example, assume for a moment that the item parameters are known, and define $\mathbf{Z}^*_{ij} = \text{vec}(\tilde{\mathbf{Z}}_{ij}, \tilde{\mathbf{T}}_{ij}) = \text{vec}(\mathbf{Z}_{ij} + \mathbf{b}, \mathbf{T}_{ij} - \boldsymbol{\lambda})$. Level 1 and level 2 of the model can then be represented by the following multivariate structure:

$$\begin{bmatrix} \theta_{ij} \\ \zeta_{ij} \\ \tilde{\mathbf{Z}}_{ij} \\ \tilde{\mathbf{T}}_{ij} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_{1ij}^t\boldsymbol{\beta}_{1j} \\ \mathbf{x}_{2ij}^t\boldsymbol{\beta}_{2j} \\ \mathbf{a}\theta_{ij} \\ -\boldsymbol{\varphi}\zeta_{ij} \end{bmatrix}, \left[\begin{array}{cc|c|c} \sigma^2_\theta & \rho & \sigma^2_\theta\mathbf{a}^t & -\rho\boldsymbol{\varphi}^t \\ \rho & \sigma^2_\zeta & \rho\mathbf{a}^t & -\sigma^2_\zeta\boldsymbol{\varphi}^t \\ \hline \mathbf{a}\sigma^2_\theta & \mathbf{a}\rho & \mathbf{a}\sigma^2_\theta\mathbf{a}^t + \mathbf{I}_K & -\mathbf{a}\rho\boldsymbol{\varphi}^t \\ \hline -\boldsymbol{\varphi}\rho & -\boldsymbol{\varphi}\sigma^2_\zeta & -\boldsymbol{\varphi}\rho\mathbf{a}^t & \boldsymbol{\varphi}\sigma^2_\zeta\boldsymbol{\varphi}^t + \boldsymbol{\omega}\mathbf{I}_K \end{array}\right] \right). \tag{8.17}$$

This representation provides insight into the complex correlational structure hidden in the data and entails several possible inferences. It also helps us to derive some of the conditional posterior distributions for the MCMC sampling algorithm (e.g., the conditional posterior distributions of the latent person parameters given the augmented data). For a general treatment of the derivation of conditional distributions from multivariate normal distributions, see, for instance, Searle et al. (1992).

Parameter ρ, which controls the covariance between the θs and ζs, plays an important role in the model. It can be considered the bridge between the separate measurement models for ability and speed. Therefore, its role within the hierarchical structure will be explored in more detail.

The conditional covariance between the latent response and log-RT on item k is given by

$$\begin{aligned}\text{Cov}\,(Z_{ijk}, T_{ijk} \mid \rho, \boldsymbol{\Lambda}_k) &= \text{Cov}\left(a_k\theta_{ij} - b_k + \epsilon_{\theta_{ijk}}, -\varphi_k\zeta_{ij} + \lambda_k + \epsilon_{\zeta_{ijk}}\right) \\ &= \text{Cov}\left(a_k\theta_{ij} + \epsilon_{\theta_{ijk}}, -\varphi_k\zeta_{ij} + \epsilon_{\zeta_{ijk}}\right) \\ &= \text{Cov}\,(a_k\theta_{ij}, -\varphi_k\zeta_{ij}) \\ &= -a_k\,\text{Cov}\,(\theta_{ij}, \zeta_{ij})\,\varphi_k \\ &= -a_k\rho\varphi_k,\end{aligned}$$

due to independence between the residuals as well as between the residuals and the person parameters. Since a_k and φ_k are positive, the latent response and log-RT correlate negatively when ρ is positive. So, in spite of conditional independence between the responses and log-RTs given the person parameters, their correlation is negative.

Due to properties of the multivariate normal distribution, the conditional distribution of θ_{ij} given ζ_{ij} is also normal:

$$\theta_{ij} \mid \zeta_{ij}, \boldsymbol{\beta}_j, \sigma^2_\theta, \sigma^2_\zeta, \rho \sim \mathcal{N}\left(\mathbf{x}^t_{1ij}\boldsymbol{\beta}_{1j} + \rho\sigma_\zeta^{-2}(\zeta_{ij} - \mathbf{x}^t_{2ij}\boldsymbol{\beta}_{2j}), \sigma^2_\theta - \rho^2\sigma_\zeta^{-2}\right).$$

A greater covariance ρ between the person parameters gives a greater reduction of the conditional variance of θ_{ij} given ζ_{ij}. The expression also shows that the amount of information about θ_{ij} in ζ_{ij} depends both on the precision of measuring the speed parameter and its correlation with the ability parameter. It can be seen that, for $\rho > 0$, the prior expected value of θ_{ij} is assumed to be higher when working at a higher speed.

From Equation (8.17), it also follows that for $\rho = 0$ the conditional posterior expectation of θ_{ij} given ζ_{ij} and the response data reduce to

$$E\big(\theta_{ij} \mid \boldsymbol{\beta}_{1j}, \tilde{\mathbf{z}}_{ij}, \sigma^2_\theta, \mathbf{a}, \mathbf{b}\big) = \big(\mathbf{a}^t\mathbf{a} + \sigma_\theta^{-2}\big)^{-1}\big(\mathbf{a}^t\tilde{\mathbf{z}}_{ij} + \sigma_\theta^{-2}\mathbf{x}^t_{ij}\boldsymbol{\beta}_{1j}\big). \qquad (8.18)$$

This expression can be recognized as the precision-weighted mean of the predictions of θ_{ij} from the (augmented) response data and from the linear regression of θ on $\mathbf{x}$; see Equation (6.12).

In Equation (8.17), in addition to the responses and RTs, the random test takers were the only extra source of heterogeneity. But another level of heterogeneity was added in (8.8), where the test takers were assumed to be nested within groups and the regression effects were allowed to vary randomly across them. Because of these random effects and correlations, the marginal covariances between the measurements change.

Several comments can be made with respect to the structural multivariate multilevel part of the model:

- In Equation (8.17), a special structure (compound symmetry) for the covariance matrix of the residuals at the level of individuals was shown to exist. This structure may lead to more efficient inferences. For a general discussion of possible parameterizations and estimation methods for multivariate random effects structures, see, for instance, Harville (1977), Rabe-Hesketh and Skrondal (2001), and Reinsel (1983).
- Linear multivariate three-level structures for continuous responses are discussed by Goldstein (2003) and Snijders and Bosker (1999), among others. As already indicated, the covariance structure of the level-3 random regression effects is assumed to be block-diagonal. This means that the parameters in the regression of $\boldsymbol{\theta}$ on $\mathbf{x}$ are conditionally independent of those in the regression of $\boldsymbol{\zeta}$ on $\mathbf{x}$. It is possible to allow these parameters to correlate, but this option is unattractive when the dimension of the covariance matrix becomes large. Typically, the covariance matrix is then poorly estimated (Laird and Ware, 1982).
- For the same reason, the covariance matrix in the multivariate normal prior of the fixed effects in Equation (8.5) is assumed to be block-diagonal. The Bayesian approach allows us to specify different levels of prior information about this matrix.

8.6 Model Selection Using the DIC

As shown in Section 3.2.3, a well-known criterion of model selection based on a deviance fit measure is the Bayesian information criterion (BIC; Schwarz, 1978). This criterion depends on the effective number of parameters in the model as a measure of model complexity, which is often difficult to calculate for hierarchical models. Although the nominal number of parameters follows directly from the likelihood, the prior distribution imposes additional restrictions on the parameter space and reduces its effective dimension.

A DIC can be formulated for choosing between models that differ in the fixed and/or random parts of the multivariate model without having to specify the number of model parameters. Interest is focused on the likelihood of the structural parameters in the model. Therefore, the multivariate random effects model in Equation (8.16) is considered. Using the factorization in Equation (8.14), the standardized deviance is

$$D(\boldsymbol{\Omega}) = \sum_{ij} \big(\mathbf{z}^*_{ij} - \mathbf{H}_P \boldsymbol{\Omega}_{ij}\big)^t \mathbf{C}^{-1} \big(\mathbf{z}^*_{ij} - \mathbf{H}_P \boldsymbol{\Omega}_{ij}\big), \tag{8.19}$$

where $\mathbf{C} = (\mathbf{I}_K \oplus \mathbf{I}_K \boldsymbol{\omega})$. When comparing models, it can be assumed without loss of generality that $p\big(\mathbf{y}, \mathbf{t}\big) = 1$ for all models. This deviance term is based on the complete data. Subsequently, the posterior expectation of the DIC over the augmented data will be taken. Then, the DIC of interest is defined as

$$\begin{aligned} \text{DIC} &= \int \big[\text{DIC} \mid \mathbf{z}\big] p(\mathbf{z}|\mathbf{y}) d\mathbf{z} \\ &= \int \big[D(\bar{\boldsymbol{\Omega}}) + 2p_D\big] p(\mathbf{z} \mid \mathbf{y}) d\mathbf{z} \\ &= E_{\mathbf{z}}\big[D(\bar{\boldsymbol{\Omega}}) + 2p_D \mid \mathbf{y}\big], \end{aligned} \tag{8.20}$$

where $\bar{\boldsymbol{\Omega}}$ equals the posterior mean and p_D is the effective number of parameters given the augmented data. The latter can be shown to be equal to the mean deviance minus the deviance of the mean. A similar procedure was proposed for mixture models by DeIorio and Robert (2002).

A DIC will be derived for the complete-data likelihood with the random effects integrated out. The specific interest in the underlying structure of $\boldsymbol{\Omega}$ becomes apparent in this expression. The corresponding penalty term in the DIC reflects the effective number of parameters related to the multivariate structure on the person parameters. Furthermore, this DIC has the advantage that it only requires estimates of fixed effects and variance parameters and not estimates of any random effects parameters. The variances, covariances, and item parameters are considered nuisance parameters, and their values are assumed to be known. In Appendix 8.12, it is shown that the expression for p_D equals

$$
\begin{aligned}
p_D &= E_{\boldsymbol{\Omega}}\big[D(\boldsymbol{\Omega}) \mid \mathbf{z}^*\big] - D(E(\boldsymbol{\Omega} \mid \mathbf{z}^*)) \\
&= \sum_{ij} \mathrm{tr}\Big[\big(\mathbf{H}_P \mathbf{x}_{ij} \mathbf{w}_j \boldsymbol{\Sigma}_\gamma \mathbf{w}_j^t \mathbf{x}_{ij}^t \mathbf{H}_P^t + \mathbf{H}_P \mathbf{x}_{ij} \mathbf{V} \mathbf{x}_{ij}^t \mathbf{H}_P^t \\
&\qquad + \mathbf{H}_P \boldsymbol{\Sigma}_P \mathbf{H}_P^t + \mathbf{C}\big)^{-1} \big(\mathbf{H}_P \mathbf{x}_{ij} \mathbf{w}_j \boldsymbol{\Sigma}_\gamma \mathbf{w}_j^t \mathbf{x}_{ij}^t \mathbf{H}_P^t \\
&\qquad + \mathbf{H}_P \mathbf{x}_{ij} \mathbf{V} \mathbf{x}_{ij}^t \mathbf{H}_P^t + \mathbf{H}_P \boldsymbol{\Sigma}_P \mathbf{H}_P^t\big)\Big], \qquad (8.21)
\end{aligned}
$$

where $\mathrm{tr}(\cdot)$ denotes the trace function (i.e., the sum of the diagonal elements). The expectation is taken with respect to the posterior distribution of $\boldsymbol{\Omega}$ given the fixed effects and covariance parameters.

The first term of the DIC, $D(E(\boldsymbol{\Omega}_{ij} \mid \mathbf{z}_{ij}^*))$, requires an expression of the marginal expectation of $\boldsymbol{\Omega}_{ij}$ given the complete data; that is,

$$
\begin{aligned}
E\left(\boldsymbol{\Omega}_{ij} \mid \mathbf{z}_{ij}^*\right) &= E\left(E\left(\boldsymbol{\Omega}_{ij} \mid \mathbf{z}_{ij}^*, \boldsymbol{\beta}_j\right) \mid \mathbf{z}_{ij}^*\right) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{e}}^{-1} + \boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\right)^{-1} \left(\boldsymbol{\Sigma}_{\mathbf{e}}^{-1} \hat{\boldsymbol{\Omega}}_{ij} + \boldsymbol{\Sigma}_{\mathbf{u}}^{-1} \mathbf{x}_{ij} \mathbf{w}_j \boldsymbol{\gamma}\right), \qquad (8.22)
\end{aligned}
$$

where $\boldsymbol{\Sigma}_{\mathbf{u}}^{-1} = \left(\mathbf{x}_{ij} \mathbf{V} \mathbf{x}_{ij}^t + \boldsymbol{\Sigma}_P\right)^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}^{-1} = \mathbf{H}_P^t \left(\mathbf{I}_K \oplus \mathbf{I}_K \boldsymbol{\omega}\right)^{-1} \mathbf{H}_P$ (see Exercise (8.6)).

The terms in Equations (8.21) and (8.22) can be estimated as a by-product of the MCMC algorithm. DICs of nested models are computed by restricting one or more variance parameters in (8.21) to zero. Usually the variance parameters are unknown. Then the DIC has to be integrated over their marginal distribution, too. In fact, the correct Bayesian approach would be to integrate the joint posterior over the nuisance parameters to obtain the marginal posterior of interest. However, this approach is not possible since no closed-form expression of the DIC can be obtained for this marginal posterior. The DIC does not account for the unknown variances, and the term in (8.21) reflects the effective number of parameters of the model without the additional variability in the posterior because of the unknown covariance parameters. The more general case with unknown covariance parameters is complex, and no simple correction seems available. Vaida and Blanchard (2005) showed that, for a mixture model, the correction for unknown covariance parameters is negligible asymptotically. So, it seems safe to assume that their effect on the estimate of the effective number of parameters only becomes apparent when the covariance parameters are estimated less precisely.

8.7 Model Fit via Residual Analysis

The fit of the measurement models to response and RT data can be assessed through residual analysis. The residual analysis for responses was described in Section 5.2, where it was based on Bayesian latent residuals. The observed RTs are continuous-valued, which simplifies a residual analysis.

The actual observation t_{ijk} is evaluated under the posterior predictive density. That is, the probability of observing a value smaller than t_{ijk} can be estimated by

$$P\left(T_{ijk} < t_{ijk} \mid \mathbf{y}, \mathbf{t}\right) \approx \sum_m \Phi\left(t_{ijk} \mid \zeta_{ij}^{(m)}, \varphi_k^{(m)}, \lambda_k^{(m)}\right) \Big/ M \tag{8.23}$$

given $m = 1, \ldots, M$ iterations of the MCMC algorithm. If the item response time model holds, the marginal cumulative posterior probabilities follow a uniform distribution according to the probability integral transformation theorem (see Exercise 3.6). By comparing the estimated moments with the true moments of the uniform distribution, it is checked whether equally spaced intervals contain equal numbers of estimated probabilities. The implicit smoothing in the empirical cumulative distributions can also be plotted against the identity line. Cumulative posterior probabilities that show departures from uniformity indicate that corresponding observations are unlikely under the model.

In a similar way, the augmented responses are evaluated under the posterior predictive density. The probability of observing a latent response under the model equals

$$P\left(Z_{ijk} < z_{ijk} \mid \mathbf{y}\right) \approx \sum_m \Phi\left(z_{ijk}^{(m)} \mid a_k^{(m)}, b_k^{(m)}, \theta_{ij}^{(m)}\right) \Big/ M. \tag{8.24}$$

The fit of the two-parameter item response model for the augmented data can be evaluated by checking whether these cumulative posterior probabilities follow a uniform distribution.

In general, the (posterior) predictive tests of Chapter 5 can also be used to evaluate the fit of the RTIRT model. Specific posterior predictive tests can be developed to test typical assumptions of the model. In Klein Entink et al. (2009b), several posterior predictive tests were used to test the fit of the RTIRT model. Exercise 8.2 is focused on evaluating the fit of the item response time model.

8.8 Simultaneous Estimation of RTIRT

The MCMC scheme for simultaneously estimating all parameters consists of a data augmentation step as described in Section 8.5. In the augmentation step, continuous-valued responses are generated to obtain a multivariate normal model for the augmented responses and log-RTs. This multivariate structure was explored in Section 8.5. A similar multivariate structure can be constructed for the item parameters.

Therefore, let $\mathbf{Z}_k$ and the vector of log-RTs ($\mathbf{T}_k$) to item k be stacked in a vector $\mathbf{Z}_k^*$. Define a covariate matrix $\mathbf{H}_I = (\boldsymbol{\theta}, -\mathbf{1}_N) \oplus (-\boldsymbol{\zeta}, \mathbf{1}_N)$. Then a linear regression structure for the item parameters can be presented as

$$\begin{aligned}\mathbf{Z}_k^* &= \mathbf{H}_I\left(a_k, b_k, \varphi_k, \lambda_k\right)^t + \boldsymbol{\epsilon}_k \\ &= \mathbf{H}_I\boldsymbol{\Lambda}_k + \boldsymbol{\epsilon}_k,\end{aligned} \tag{8.25}$$

where $\boldsymbol{\epsilon}_k \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_N \oplus \mathbf{I}_N\omega_k\right)$. This multivariate model for the complete data facilitates the sampling of the item parameters from their conditional distributions.

MCMC Scheme 6 (RTIRT)

A1) Sample the item parameters

1. *Binary response data. The item parameters are the coefficients of the regression of* $\mathbf{Z}_k^*$ *on* $\mathbf{H}_I$ *in Equation* (8.25)*. Combined with the prior in Section 8.4.1, the conditional posterior density is a multivariate normal with parameters*

$$\begin{aligned}E\left(\boldsymbol{\Lambda}_k \mid \mathbf{z}_k^*, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_I\right) &= \left(\boldsymbol{\Sigma}_I^{-1} + \mathbf{H}_I^t\mathbf{C}_I^{-1}\mathbf{H}_I\right)^{-1}\left(\mathbf{H}_I^t\mathbf{C}_I^{-1}\mathbf{z}_k^* + \boldsymbol{\mu}_I\boldsymbol{\Sigma}_I^{-1}\right), \\ \mathrm{Var}\left(\boldsymbol{\Lambda}_k \mid \mathbf{z}_k^*, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_I\right) &= \left(\boldsymbol{\Sigma}_I^{-1} + \mathbf{H}_I^t\mathbf{C}_I^{-1}\mathbf{H}_I\right)^{-1},\end{aligned}$$

   respectively, where $\mathbf{C}_I = \mathbf{I}_N \oplus \mathbf{I}_N\omega_k$.
2. *Polytomous response data; see Exercise 8.4.*

A2) Sample the person parameters

The multivariate distribution of the person parameters and the augmented data equals

$$\begin{bmatrix}\boldsymbol{\Omega}_{ij} \\ \mathbf{z}_{ij}^*\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_{ij}\boldsymbol{\beta}_j \\ \mathbf{H}_P\mathbf{x}_{ij}\boldsymbol{\beta}_j\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_P & \boldsymbol{\Sigma}_P\mathbf{H}_P^t \\ \mathbf{H}_P\boldsymbol{\Sigma}_P & \mathbf{H}_P\boldsymbol{\Sigma}_P\mathbf{H}_P^t + \mathbf{C}_P\end{bmatrix}\right),$$

where $\mathbf{C}_P = \mathbf{I}_K \oplus \mathbf{I}_K\boldsymbol{\omega}$ *and* $\mathbf{H}_P = (\mathbf{a} \oplus -\boldsymbol{\varphi})$*. The conditional posterior of the person parameters is a multivariate normal,*

$$\boldsymbol{\Omega}_{ij} \mid \mathbf{z}_{ij}^{*(m+1)}, \boldsymbol{\Sigma}_P^{(m)}, \boldsymbol{\beta}_j^{(m)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\Omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\Omega}}\right),$$

with

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{\Omega}} &= \mathbf{x}_{ij}\boldsymbol{\beta}_j + \mathbf{H}_P\boldsymbol{\Sigma}_P(\mathbf{H}_P\boldsymbol{\Sigma}_P\mathbf{H}_P^t + \mathbf{C}_P)^{-1}\left(\mathbf{z}_{ij}^* - \mathbf{H}_P\mathbf{x}_{ij}\boldsymbol{\beta}_j\right), \\ \boldsymbol{\Sigma}_{\boldsymbol{\Omega}} &= \boldsymbol{\Sigma}_P - \boldsymbol{\Sigma}_P\mathbf{H}_P^t\left(\mathbf{H}_P\boldsymbol{\Sigma}_P\mathbf{H}_P^t + \mathbf{C}_P\right)^{-1}\mathbf{H}_P\boldsymbol{\Sigma}_P.\end{aligned}$$

A3) Sample the remaining measurement parameters

1. *For binary response data, the hyperparameters* $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ *are sampled from a normal inverse Wishart density. Given values for* $(K_0, \boldsymbol{\mu}_0)$ *and* $(\nu, \boldsymbol{\Sigma}_\mathbf{0})$*, sample* $\boldsymbol{\mu}_I^{(m+1)}$ *and* $\boldsymbol{\Sigma}_I^{(m+1)}$ *from*

$$\begin{aligned}\boldsymbol{\mu}_I \mid \boldsymbol{\Sigma}_I^{(m)}, \boldsymbol{\Lambda}^{(m+1)} &\sim \mathcal{N}\left(\frac{K_0}{K_0+K}\boldsymbol{\mu}_0 + \frac{K}{K_0+K}\overline{\boldsymbol{\Lambda}}, \boldsymbol{\Sigma}_I/(K+K_0)\right) \\ \boldsymbol{\Sigma}_I \mid \boldsymbol{\Lambda}^{(m+1)} &\sim \mathcal{IW}\left(K+\nu, \boldsymbol{\Sigma}^*\right),\end{aligned}$$

   where

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0 + K\mathbf{S} + \frac{K_0 K}{K_0 + K}\left(\overline{\boldsymbol{\Lambda}} - \boldsymbol{\mu}_0\right)\left(\overline{\boldsymbol{\Lambda}} - \boldsymbol{\mu}_0\right)^t,$$
$$\mathbf{S} = K^{-1}\sum_k \left(\boldsymbol{\Lambda}_k - \overline{\boldsymbol{\Lambda}}\right)\left(\boldsymbol{\Lambda}_k - \overline{\boldsymbol{\Lambda}}\right)^t,$$
$$\overline{\boldsymbol{\Lambda}} = K^{-1}\sum_k \boldsymbol{\Lambda}_k.$$

In the case of polytomous response data, see Exercise 8.4.

2. *For each item k, sample the residual variance $\omega_k^{2(m+1)}$ given $\varphi_k^{(m+1)}$, $\lambda_k^{(m+1)}$, $\boldsymbol{\zeta}^{(m+1)}$, and $\mathbf{t}$ from an inverse gamma distribution with parameter $g_1 + N/2$ and scale parameter*

$$g_2 + \frac{1}{2}\sum_{i,j}\left(t_{ijk} - (\lambda_k - \varphi_k\zeta_{ij})\right)^2.$$

B) Sampling of structural model parameters

1. *For each j, sample $\boldsymbol{\beta}_j^{(m+1)}$ given $\boldsymbol{\Omega}_j^{(m+1)}$, $\boldsymbol{\Sigma}_P^{(m)}$ and $\mathbf{V}^{(m)}, \boldsymbol{\gamma}^{(m)}$ from a multivariate normal with mean*

$$\boldsymbol{\mu}_\beta = \mathbf{w}_j\boldsymbol{\gamma} + \mathbf{x}_j\mathbf{V}\left(\mathbf{x}_j\mathbf{V}\mathbf{x}_j^t + \boldsymbol{\Sigma}_P\right)^{-1}\left(\boldsymbol{\Omega}_j - \mathbf{x}_j\mathbf{w}_j\boldsymbol{\gamma}\right)$$

and variance

$$\boldsymbol{\Sigma}_\beta = \mathbf{V} - \mathbf{V}\mathbf{x}_j^t\left(\mathbf{x}_j\mathbf{V}\mathbf{x}_j^t + \boldsymbol{\Sigma}_P\right)^{-1}\mathbf{x}_j\mathbf{V}.$$

2. *Assume that $\boldsymbol{\gamma}$ is a priori multivariate normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_\gamma$. Sample fixed coefficients $\boldsymbol{\gamma}^{(m+1)}$ from the full conditional*

$$\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(m+1)}, \mathbf{V}^{(m)}, \boldsymbol{\Sigma}_\gamma \sim \mathcal{N}\left(\boldsymbol{\mu}_\gamma, \boldsymbol{\Xi}_\gamma\right),$$

where

$$\boldsymbol{\mu}_\gamma = \boldsymbol{\Xi}_\gamma\sum_j \mathbf{w}_j^t\mathbf{V}^{-1}\boldsymbol{\beta}_j,$$
$$\boldsymbol{\Xi}_\gamma = \left(\sum_j \mathbf{w}_j^t\mathbf{V}^{-1}\mathbf{w}_j + \boldsymbol{\Sigma}_\gamma^{-1}\right)^{-1}.$$

3. *Sample $\mathbf{V}^{(m+1)}$ from an inverse Wishart density with $n_0 + J$ degrees of freedom and scale matrix*

$$\sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)\left(\boldsymbol{\beta}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)^t + \mathbf{V}_0,$$

where $n_0 \geq 2$.

4. *Sample* $\boldsymbol{\Sigma}_P^{(m+1)}$ *given* $\boldsymbol{\Omega}^{(m+1)}$ *and* $\boldsymbol{\beta}^{(m+1)}$ *as described in Section 8.4.2.*

This scheme is easily extended to handle a three-parameter measurement model. Then, an additional augmentation step is introduced according to Equations (4.20) and (4.21). Subsequently, guessing parameters are sampled according to Equation (4.22).

8.9 Natural World Assessment Test

A computerized version of the Natural World Assessment Test (NAW-8) was taken by 388 second year students who answered 65 items. The test takers were required to participate in the educational assessment. This low-stakes test is used to assess the quantitative and scientific reasoning proficiencies of college students. Besides the test results, the test taker's SAT score, gender (GE), a self-reported measure of citizenship (CS), and a self-reported measure of test importance (TI) were available. Citizenship was a measure of a test taker's willingness to help the university collect its assessment data. Test importance was a self-reported measure of how important the test was for the test taker.

In this example, the relationship between ability and speed is explored and the extent to which individual background information explains variability in ability and speed of working is evaluated. The three-parameter item response model is used as the measurement model for the item responses.

The RTIRT model was estimated with 20,000 iterations of the Gibbs sampler, and the first 10,000 iterations were discarded as the burn-in. Several posterior predictive checks were carried out to evaluate the fit of the model. The odds ratio (Equation (5.31)) was used to test for violations of local independence. This test indicated that less than 4% of the possible item combinations showed a significant between-item dependency. Furthermore, the replicated response patterns under the posterior distribution matched the observed data quite well using the discrepancy measure in Equation (5.33). From the posterior residual check (Equation (8.23)), the item response time model described the data well.

The model with fixed time-discrimination parameters ($\boldsymbol{\varphi} = 1$) was compared with the model with unrestricted time-discrimination parameters ($\boldsymbol{\varphi} \neq 1$). To compare the models, the DIC was computed with $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ and $\mathbf{V}$ equal to zero for both models. The estimated DIC's were 85,780 and 84,831 for the restricted and the unrestricted RTIRT models, respectively. The estimated time-discrimination parameters varied from .25 to 1.65. It was concluded that the items discriminated differently between test takers of different working speeds. This means that the rate of change in log-RT relative to changes in working speed varied over items.

Speed tests are characterized by the fact that the time limit is strict. The goal is to measure how quickly test takers can answer items. On speed

tests, some respondents may not finish all the items. Others may rapidly mark answers without thorough thought. Therefore, the time limit provokes different types of response behaviors that often violate the RTIRT's stationary speed assumption. A test taker running out of time will show a deviation of this assumption towards the end of the test. Such a deviation can be detected by estimating the outlying probabilities of the standardized residuals $\epsilon_{\zeta_{ijk}}/\omega_k$. The probability that the standardized residual exceeds some prespecified value q equals

$$P\left(\left|\epsilon_{\zeta_{ijk}}/\omega_k\right| > q \mid t_{ijk}, \varphi_k, \lambda_k, \zeta_{ij}\right) = 2\Phi\left(-q; \lambda_k - \varphi_k \zeta_{ijk}\right). \quad (8.26)$$

For $q = 2$, less than 5% of the residuals have a high posterior probability of being greater than two standard deviations. Therefore, there were no indications of speededness for this test.

Table 8.1 gives the estimated covariance components and correlations between the measurement parameters at level 1. The correlation between the person parameters was estimated to be $-.81$. This strong negative dependence indicates that higher-ability candidates are working at a lower speed. This might indicate that students who took their time were serious about the test and that students who worked fast were not doing their best to solve the items. Although care has to be taken to draw strong conclusions, the response times might reveal that the measurement instrument is not valid since some students were not taking the test seriously.

The within-item covariance estimates contain information about the common covariance item structure. The estimated item characteristic correlations are given under the heading "Cor". The difficulty of an item is positively correlated with the time intensity of an item. This is in line with the common assumption that the more difficult tasks require more time to be solved. The more difficult items and the more time-intensive items, which are positively correlated, have higher time-discrimination parameters. Thus, the more time-discriminating items can be recognized as the more difficult (time-intensive) items. The more difficult items discriminate less between test takers with different abilities. It follows that the more difficult test items can better discriminate test takers with different speed levels than test takers with different ability levels.

The individual background information was used to explain simultaneously variability in the ability and speed levels. The full RTIRT model contains all covariate information for explaining variability in both latent variables. In Table 8.2, the parameter estimates of the full RTIRT model are given. It follows that male students perform equally well on the test with respect to accuracy and speed. Students that were more willing to help the university in collecting assessment data (citizenship) were not working faster or more accurately. The SAT scores explain a significant amount of variation in the ability levels. However, the SAT scores did not explain a significant amount of variability in the speed levels. This means that the student's speed of working is not related to the SAT score. Finally, a positive relationship of test importance with

Table 8.1. NWA-8: Covariance components and correlation estimates.

Variance Components		Mean	SD	Cor.
Person Covariance Matrix $\mathbf{\Sigma}_P$				
(Ability)	σ^2_θ	1.00	-	1.00
	ρ	−.37	.02	−.81
(Speed)	σ^2_ζ	.21	.02	1.00
Item Covariance Matrix $\mathbf{\Sigma}_I$				
(Discrimination)	Σ_{11}	.03	.01	1.00
	Σ_{12}	−.03	.02	−.29
	Σ_{13}	.03	.01	.58
	Σ_{14}	.01	.02	.14
(Difficulty)	Σ_{22}	.39	.08	1.00
	Σ_{23}	.06	.03	.35
	Σ_{24}	.07	.05	.20
(Time discrimination)	Σ_{33}	.08	.02	1.00
	Σ_{34}	.08	.03	.47
(Time intensity)	Σ_{44}	.34	.06	1.00

ability was expected. Students who found the test important were expected to try harder and spend more time per item to receive a higher grade. The significant negative relationship of test importance with speed supports this hypothesis; that is, students who found the test important worked with less speed than students who did not. Test importance is also positively correlated with ability.

8.10 Discussion

A multivariate hierarchical item response measurement framework was developed for measuring ability and speed of working given item responses and RTs, respectively. The RTs and responses are assumed to be conditionally independent given the levels of speed and ability. The latent variables ability and speed may correlate, which means that speed of working can influence the accuracy of the results. This trade-off between speed and accuracy often describes a negative correlation between the speed and accuracy levels at which a person can operate. At a higher level of the RTIRT model, a multivariate multilevel structural model is defined for the person variables. At this stage, the latent correlation structure is defined and explanatory information can be incorporated into the model to explain between-individual (within-group) and between-group differences in speed and ability levels. Klein Entink et al. (2009b) showed how to model dependencies between item parameters by modeling a population model for the item parameters. In this population model,

Table 8.2. NWA-8: Explaining variance in ability and speed levels.

	Full RTIRT			Restricted RTIRT		
	Mean	SD	HPD	Mean	SD	HPD
Fixed Effects						
Ability						
γ_{00} Intercept	.00	-		.00	-	
γ_{01} SAT score	.58	.06	[.46, .70]	.60	.07	[0.47, 0.73]
γ_{02} Test importance	.65	.06	[.52, .78]	.61	.07	[0.49, 0.74]
γ_{03} Male	.14	.12	[−.09, .37]			
γ_{04} Citizenship	−.08	.06	[−.20, .04]			
Speed						
γ_{10} Intercept	.00	-		.00	-	
γ_{11} SAT score	−.02	.02	[−.07, .02]			
γ_{12} Test importance	−.22	.02	[−.27, −.17]	−.22	.02	[−.27, −.18]
γ_{13} Male	−.02	.05	[−.11, .08]			
γ_{14} Citizenship	−.01	.02	[−.06, .04]			
Residual Variance						
σ^2_θ Ability	1.00	-		1.00	-	
σ^2_ζ Speed	.21	.02	[.18, .24]	.21	.02	[.18, .24]
ρ Covariance	−.37	.02	[−.42, −.33]	−.37	.02	[−.42, −.33]
Information Criteria						
−2log-likelihood			51020.32			51001.09
DIC (p_D)			52570.3(775.0)			52551.1(775.0)

content-specific information about items can be related to the item difficulty and intensity.

The modeling framework discussed for analyzing log-RTs and responses belongs to the class of multivariate nonlinear mixed effects models. This class of models has not received much attention in the literature. For the class of multivariate linear mixed effects models, Reinsel (1982) derived closed-form estimates given a balanced design and completely observed data. Schafer and Yucel (2002) discussed the multivariate linear case where multiple responses are ignorably missing. Pinheiro and Bates (2000, Chapter 7) discussed multilevel nonlinear mixed effects models, with multiple nested factors, for univariate data including computational methods for maximum likelihood estimation that are implemented in the *nlme* library of S+ and R (R Development Core Team, 2010). A multivariate extension of the multilevel nonlinear mixed effects models was proposed by Hall and Clutter (2004). They iteratively fitted a sequence of linear mixed effects models derived from first-order Taylor expansions to the nonlinear model. In this approach, the true likelihood is considered to be the first-order approximate likelihood, and approximate maximum likeli-

hood estimates are obtained. Standard errors, likelihood ratio tests, and model selection criteria are based on the approximate likelihood. The MCMC algorithm developed can simultaneously estimate the model parameters without having to approximate the joint posterior distribution. Extensions to handle ignorable missing item responses and unbalanced designs can be made in a straightforward way.

For a long time, experimental and cognitive psychologists have used RTs to test theories. The pioneering work of the Dutch psychologist F. C. Donders (Donders, 1868) on measurement of mental processing times dates from the middle of the 19th century. Recently, RTs have been used in test theory (and survey research), especially since they are easily collected in computer adaptive (or assisted) testing. The introduction of RTs into test theory led to various applications. Attention has been focused on simultaneously measuring speed and ability (van der Linden, 2007) and using background information as collateral information (Klein Entink et al., 2009a). RTs can also be used to diagnose response problems (e.g., van der Linden and Krimpen-Stoop, 2003). Items that induce guessing behavior can be detected from the RTs when, for example, time-intensive (difficult) items take less time and are followed by incorrect answers than less time-intensive (easy) items followed by correct answers.

In test construction, item selection is based on the information functions of the items. For example, in a computer adaptive test (CAT), the selection of a new item is based on the estimated ability of the person given the observed response data and the information functions of the items that were not presented. The object is to minimize the set of items that are needed to measure a test taker's ability up to a specified accuracy. The RTs provide an additional source of data that can be used to improve the accuracy of the ability estimate (Klein Entink, 2009). Subsequently, the use of RTs as an additional source of information can improve the (minimal) subset of items that is obtained via a selection criterion based on response data. Furthermore, a different selection criterion may include an item's time characteristics such as intensity and time discrimination. The selected subset of items based on this criterion can reduce the length of the test in time in comparison with a random selection of items (van der Linden, 2008).

8.11 Exercises

8.1. Consider 12 RTs and item responses of 356 students to rule-based items for measuring speed of working and figural reasoning ability, respectively (Klein Entink et al., 2009b). The log-RTs are modeled by the response time model in Equation (8.1) with time discriminations equal to one.
(a) Define priors for the parameters of the response time model, and complete the implementation given in Listing 8.1.

Listing 8.1. WinBUGS code: Response time measurement model.

```
model{
    for (i in 1:N) {
      for (k in 1:K) {
                T[i,k] ~ dnorm(lambda[k] - zeta[i], precomega[k])
                T[i,k] <- log(RT[i,k])
       }
       zeta[i] ~ dnorm(0, preczeta)
   }

   for (k in 1:K) {
      lambda[k] ~ dnorm(mu, preclambda)
   }
```

(b) Estimate the model using WinBUGS. Check that the posterior mean estimates of the time-intensity parameters correspond closely with the observed item means.
(c) Estimate the intraclass correlation coefficient $\frac{\sigma_\zeta^2}{\sigma_\zeta^2+\omega_k^2}$ per item, and explain the results.
(d) Verify that the marginal probability of observing a log-RT less than t_{ik} on item k equals

$$\begin{aligned} P\left(T_{ik} < t_{ik}\right) &= E\left(P\left(\lambda_k - \zeta_i + \epsilon_{\zeta_{ik}} < t_{ik}\right) \mid \zeta_i\right) \\ &= \Phi\left(\frac{t_{ik} - \lambda_k}{\sqrt{\sigma_\zeta^2 + \omega_k^2}}\right). \end{aligned}$$

(e) For each item, compute the marginal probability of observing a log-RT less than the average time intensity.

8.2. To evaluate the fit of the response time model, item- and person-fit statistics based on Bayesian residuals can be defined.
(a) Find the posterior density of a Bayesian residual defined as

$$\epsilon_{\zeta_{ik}} = T_{ik} - \left(\lambda_k - \zeta_i\right).$$

(b) Explain that a person-fit statistic is defined by the function $Q_{p,i}$,

$$Q_{p,i}\left(\mathbf{T}_i\right) = \sum_k \left(\frac{\epsilon_{\zeta_{ik}}}{\omega_k}\right)^2,$$

with corresponding tail-area probability

$$p_0\left(Q_{p,i}\right) = P\left(\chi_K^2 > Q_{p,i}\left(\mathbf{T}_i\right)\right).$$

(c) In the same way, define an item-fit statistic and the tail-area probability based on the Bayesian residuals defined in (a).
(d) Use the result from Exercise 8.1(d) to define the tail-area probability

$$P\left(Z > \frac{T_{ik} - \lambda_k}{\sqrt{\sigma_\zeta^2 + \omega_k^2}}\right),$$

where Z is standard normally distributed, and explain that it can be used to investigate the compatibility of the model with the observed log-RTs.
(e) (continuation of Exercise 8.1) Evaluate the fit of the response time model by evaluating the tail-area probabilities in WinBUGS.

8.3. (continuation of Exercise 8.1) Students are grouped by the item response (correct/incorrect). Attention is focused on differential item response time functioning for the groups of respondents.
(a) Complete Listing 8.2 with priors for the parameters. Investigate whether the time-intensity parameters differ across the groups.

Listing 8.2. WinBUGS code: Differential response time measurement model.

```
model{
    for (i in 1:N) {
      for (k in 1:K) {
         T[i,k] ~ dnorm(lambda[class[i,k],k] - zeta[i],precomega[k])
         T[i,k] <- log(RT[i,k])
         class[i,k] <- Y[i,k] + 1
       }
       zeta[i] ~ dnorm(0,preczeta)
   }

   for (k in 1:K) {
      lambda[1,k] ~ dnorm(mu1,preclambda1)
      lambda[2,k] ~ dnorm(mu2,preclambda2)
   }
}
```

(b) Define a univariate normal mixture density for the speed of working at item k,

$$p(\zeta_{i_k}) = I(Y_{ik} = 0)\phi\left(\zeta_{i_k}; 0, \sigma_\zeta^2\right) + I(Y_{ik} = 1)\phi\left(\zeta_{i_k}; \mu_k, \sigma_\zeta^2\right).$$

Explain that the mixture density allows individual variation in speed of working across items.
(c) Explain that the response time model in Exercise 8.1 with a prior mixture density for the speed parameter is identified.
(d) Implement the mixture density in Listing 8.1, and investigate whether the speed of working differs across items and across correct and incorrect answers.

8.4. Assume a normal ogive graded response model (Equation (4.25)) for the observed responses. Let $\boldsymbol{\kappa}_k$ denote the set of threshold parameters for item k with the order restriction

$$-\infty < \kappa_{k1} \leq \kappa_{k2} \leq \ldots \leq \kappa_{k,C_k-1} < \infty, \tag{8.27}$$

where there are C_k categories.

(a) State the advantages and disadvantages of a multivariate normal prior distribution for the item parameters $(a_k, \boldsymbol{\kappa}_k, \varphi_k, \lambda_k)$ with mean $\boldsymbol{\mu}_I$ and covariance matrix $\boldsymbol{\Sigma}_I$, where $\boldsymbol{\kappa}_k$ obey the order restriction in (8.27).
(b) Show that a conditional normal prior $p(\boldsymbol{\kappa}_k \mid a_k, \varphi_k, \lambda_k, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ can be derived from the multivariate normal prior.
(c) An M-H step for drawing values from the conditional posterior distribution $p(\boldsymbol{\kappa}_k \mid \mathbf{y}_k, \boldsymbol{\theta}, a_k, \varphi_k, \lambda_k, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ can be defined. Use the proposal distribution in Equation (4.31) for drawing candidate values. Define the acceptance ratio (Equation (3.1)) for evaluating the proposed values.
(d) Show that the normalizing constant of the conditional distribution of $p(a_k, \boldsymbol{\kappa}_k, \varphi_k, \lambda_k \mid \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ contains a term $P(\boldsymbol{\kappa}_k \in \mathcal{A}_k \mid \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$, where $\mathcal{A}_k = \left\{\boldsymbol{\kappa}_k \in \mathcal{R}^{C_k-1}, -\infty < \kappa_{k1} \leq \kappa_{k2} \leq \ldots \leq \kappa_{k,C_k-1} < \infty\right\}$. Why does this complicate direct sampling from the conditional posterior distribution of $\boldsymbol{\mu}_I$?
(e) The hyperparameter $\boldsymbol{\mu}_I$ given $\boldsymbol{\Sigma}_I$ has a normal distribution with mean parameter $\boldsymbol{\mu}_0$ and scale parameter $\boldsymbol{\Sigma}_I/K_0$. Define an M-H step for sampling values $\boldsymbol{\mu}_I$ from the full conditional distribution.

8.5. Consider the Amsterdam Chess Test data from players responding to computerized chess problems to measure their chess-playing proficiency (van der Maas and Wagenmakers, 2005). For each of the 40 tasks, the response accuracy and the RT were measured.
(a) Fit the RTIRT model to the data using the R-package CIRT (Fox et al., 2007).
(b) Plot the estimated ability against the speed of playing, and explain the relationship. How does the trend relate to the estimated covariance components of matrix $\boldsymbol{\Sigma}_P$?
(c) Test whether the mean speed level for correct responses differs from the mean speed level for the incorrect responses to item 7. Are players responding correctly to item 7 solving the items faster than those who respond to it incorrectly?

8.6. The likelihood model is given by Equation (8.16), and the prior structure is defined in Equations (8.6) and (8.7).
(a) Show that a (marginal) prior model for $\boldsymbol{\Omega}_{ij}$ is given by

$$\boldsymbol{\Omega}_{ij} \mid \boldsymbol{\gamma}, \boldsymbol{\Sigma}_P, \mathbf{V} \sim \mathcal{N}\left(\mathbf{x}_{ij}\mathbf{w}_j\boldsymbol{\gamma}, \mathbf{x}_{ij}\mathbf{V}\mathbf{x}_{ij}^t + \boldsymbol{\Sigma}_P\right).$$

(b) Derive the conditional distribution of $\boldsymbol{\Omega}_{ij}$ given the complete data, item parameters, fixed effects, and covariance parameters. Show that the expectation equals the term in Equation (8.22).

8.12 Appendix: DIC RTIRT Model

An expression for the number of effective parameters is constructed in which the item parameters are considered to be nuisance parameters. Instead of computing the posterior mean deviance minus the deviance of the posterior mean of $\boldsymbol{\Omega}_{ij}$, an expected penalty term is derived where the expectation is taken with respect to the posterior density of $\boldsymbol{\Omega}_{ij}$ given the complete data and covariance parameters. That is,

$$
\begin{aligned}
p_D &= \overline{D(\boldsymbol{\Omega})} - D(\bar{\boldsymbol{\Omega}}) \qquad (8.28)\\
&= E_{\boldsymbol{\Omega}}\big[D(\boldsymbol{\Omega}) \mid \mathbf{z}^*\big] - D(E(\boldsymbol{\Omega} \mid \mathbf{z}^*))\\
&= E_{\boldsymbol{\Omega}}\left[\sum_{ij}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)^t\mathbf{C}^{-1}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\right] - D(E(\boldsymbol{\Omega} \mid \mathbf{z}^*))\\
&= \operatorname{tr}\left[\sum_{ij} E_{\boldsymbol{\Omega}}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)^t\mathbf{C}^{-1}\right] -\\
&\qquad \operatorname{tr}\left[\sum_{ij}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)^t\mathbf{C}^{-1}\right]\\
&= \sum_{ij}\operatorname{tr}\Big[E_{\boldsymbol{\Omega}}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)^t\mathbf{C}^{-1} -\\
&\qquad \big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)^t\mathbf{C}^{-1}\Big]\\
&= \sum_{ij}\operatorname{tr}\big[\mathbf{C}^{-1}\operatorname{Var}(\boldsymbol{\epsilon}_{ij} \mid \mathbf{z}^*_{ij})\big] \qquad (8.29)\\
&= \sum_{ij}\operatorname{tr}\big[\mathbf{C}^{-1}\left(\operatorname{Var}(\boldsymbol{\epsilon}_{ij}) - \operatorname{Cov}(\boldsymbol{\epsilon}_{ij}, \mathbf{z}^*_{ij})\operatorname{Var}(\mathbf{z}^*_{ij})^{-1}\operatorname{Cov}(\boldsymbol{\epsilon}_{ij}, \mathbf{z}^*_{ij})\right)\big]\\
&= \sum_{ij}\operatorname{tr}\big[\operatorname{Var}(\mathbf{z}^*_{ij})^{-1}\operatorname{Var}\big(\mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\big] \qquad (8.30)\\
&= \sum_{ij}\operatorname{tr}\Big[\big(\mathbf{H}_P\mathbf{x}_{ij}\mathbf{w}_j\boldsymbol{\Sigma}_\gamma\mathbf{w}_j^t\mathbf{x}_{ij}^t\mathbf{H}_P^t + \mathbf{H}_P\mathbf{x}_{ij}\mathbf{V}\mathbf{x}_{ij}^t\mathbf{H}_P^t + \mathbf{H}_P\boldsymbol{\Sigma}_P\mathbf{H}_P^t + \mathbf{C}\big)^{-1}\\
&\qquad \big(\mathbf{H}_P\mathbf{x}_{ij}\mathbf{w}_j\boldsymbol{\Sigma}_\gamma\mathbf{w}_j^t\mathbf{x}_{ij}^t\mathbf{H}_P^t + \mathbf{H}_P\mathbf{x}_{ij}\mathbf{V}\mathbf{x}_{ij}^t\mathbf{H}_P^t + \mathbf{H}_P\boldsymbol{\Sigma}_P\mathbf{H}_P^t\big)\Big].
\end{aligned}
$$

The terms in (8.29) can be recognized as the posterior variances of the residuals, whereas those in (8.30) follow from the fact that, because of independence, the variance of $\mathbf{z}^*_{ij}$ equals the sum of the variance of $\mathbf{H}_P\boldsymbol{\Omega}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$.

9

Randomized Item Response Models

Item responses can be masked before they are observed via a randomized response mechanism. This technique is used to protect individuals and improve their willingness to answer truthfully. Various traditional randomized response sampling techniques are discussed and extended to a multivariate setting. So-called randomized item response models will be introduced for analyzing multivariate randomized response data. This class of models can also be extended to handle explanatory information at different hierarchical levels. The models discussed are particularly suitable for analyzing sensitive individual characteristics and their relationships to background variables.

9.1 Surveys about Sensitive Topics

The collection of data through surveys is subject to two main sources of error. First, there is sampling error due to the fact that a sample is studied instead of the entire corresponding population. Sampling error can be reduced by increasing the sample size, by improving the efficiency of the sampling design, or by improving the estimation procedure to improve the accuracy of the population estimates. A second type of error can be characterized as nonsampling error, which covers the systematic errors. Bias in survey data can be caused by response bias, which is the difference between the true response according to the respondent's true beliefs and the expected observed response over repeated measurements. There are two well-known sources of response bias, nonresponse (the refusal to respond) and typical response behavior, which leads to underreporting or overreporting such that the answers are systematically lower or higher, respectively.

Surveys on highly personal and sensitive issues may lead to response behavior that manifests as answering refusals and false responses, making inferences difficult. Respondents may attempt to distort answers to avoid embarrassment and mask their socially undesirable answers. Apart from social desirability, certain questions are inherently offensive because they invade the

J.-P. Fox, *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-0742-4_9,

respondent's privacy. In that case it is not the answer that is sensitive but the question. The sensitivity of the questions also depends on respondents' concerns about disclosing information to the interviewer and third parties. In general, asking sensitive questions leads to a reduction in overall and item response rates and response accuracy (percentage of truthful answers).

In this light, it is to be expected that the item nonresponse increases with the question sensitivity, but this relationship is not so evident (Tourangeau and Yan, 2007). Nevertheless, there is considerable evidence that the data collection method influences the answers and that the procedure of self-administration increases the respondents' willingness to respond truthfully to sensitive questions (e.g., Richman, Kiesler, Weisband and Drasgow, 1999; Tourangeau, Rips and Rasinski, 2000; Tourangeau and Yan, 2007). The main variety of self-administration methods are based on promises of confidentiality and can be characterized as a direct way of obtaining sensitive information. Obtaining valid and reliable information depends on the cooperation of the respondents, and the willingness of the respondents depends on the confidentiality of their responses.

9.2 The Randomized Response Technique

An indirect way of asking sensitive questions such that the respondents' answers do not reveal anything definite was developed by Warner (1965). Warner (1965) developed a data collection procedure, the randomized response (RR) technique, that allows researchers to obtain sensitive information while guaranteeing privacy to respondents. In this procedure, a respondent randomly selects one of two cards with statements of the form: I have the sensitive characteristic and (2) I do not have the sensitive characteristic. A randomization device is used to choose one of the cards (e.g., tossing a die or using a spinner). The randomization is performed by the interviewee, and the interviewer is not permitted to observe the outcome of the randomization.

The respondent is protected since the interviewer will not know which question is being answered. The interviewee responds to a question selected by the randomization device, and the interviewer only knows the response. The respondent's privacy or anonymity is well protected because only the respondent knows which question was answered. In several empirical studies it has been shown that respondents are more willing to provide honest answers with this technique because their answers do not reveal any information about themselves. Fox and Tracy (1986), Lensvelt-Mulders, Hox, van der Heijden and Maas (2005), and Tracy and Fox (1981), among others, showed via several studies that the randomized response method leads to less biased estimates of the sensitive characteristic in comparison with estimates obtained from traditional survey methods.

9.2.1 Related and Unrelated Randomized Response Designs

The model of Warner (1965) for dichotomous responses is meant for estimating a population proportion, π, that represents some sensitive characteristic. The respondent answers "True" or "False" without revealing which question was selected by the randomizing device. Let p_1 denote the probability that question (1) will be selected by the randomizing device. Then, the probability that the respondent indexed i gives a positive response equals

$$P(Y_i = 1) = p_1\pi + (1 - \pi)(1 - p_1). \tag{9.1}$$

In this model, both questions or statements deal with the sensitive topic.

Greenberg, Abul-Ela, Simmons and Horvitz (1969) proposed a different design, in which the sensitive question is paired with an innocuous question. In this unrelated question design, the second question is not related and completely innocuous, and the probability of a positive response is known. In this perspective, Warner's model is also referred to as the related randomized response design since a sensitive question is paired with its converse.

The unrelated question can also be built into the randomizing device. In that case, two probabilities are specified by the randomizing device: probability p_1 that the respondent has to answer the sensitive question and the conditional probability p_2 of a positive response in case a forced response has to be given (Edgell, Himmelfarb and Duchan, 1982). The probability of a positive response for respondent i can be stated as

$$P(Y_i = 1) = p_1\pi + (1 - p_1)p_2. \tag{9.2}$$

The extension to multiple, say $c = 1, \ldots, C$, response categories is easily made. Let $\pi(c)$ denote the proportion of respondents scoring in category c, and the randomization device determines if the item is to be answered honestly with probability p_1 or a forced response is scored with probability $1 - p_1$. If a forced response is given, it is scored in category c with conditional probability $p_2(c)$. Then, the probability of observing a score in category c equals

$$P(Y_i = c) = p_1\pi(c) + (1 - p_1)p_2(c). \tag{9.3}$$

Note that it is assumed that the response probability of scoring in category c for the nonsensitive unrelated question is known a priori. This is more efficient since it reduces the sampling variability. Furthermore, it is quite easy to define unrelated neutral questions whose response probabilities are known in advance (e.g., Greenberg et al., 1969).

In a setting where the response probabilities of the unrelated question are unknown, two independent distinct samples, say A_1 and A_2, are needed from the population where two independent randomization devices are employed. Let the randomization devices be such that p_1^1 is the probability that the respondent has to answer the sensitive question in A_1 and p_1^2 the probability

in A_2. Subsequently, given group membership, the probability of scoring in category c equals

$$P(Y_i = c) = \begin{cases} \pi(c)p_1^1 + \left(1 - p_1^1\right) p_2(c) & \text{if} \quad i \in A_1 \\ \pi(c)p_1^2 + \left(1 - p_1^2\right) p_2(c) & \text{if} \quad i \in A_2. \end{cases} \tag{9.4}$$

When selecting p_1^1 close to p_1^2, the point estimates of $\pi(c)$ and $p_2(c)$ may be unstable and greater than unity (Greenberg et al., 1969).

9.3 Extending Randomized Response Models

The traditional randomized response techniques are meant for univariate data, and the main goal is to obtain an estimate of the proportion of respondents having the sensitive characteristic. Theoretical details about the estimation of π can be found for example, in Warner (1965) and Greenberg et al. (1969). In some cases, additional information at the individual level is available that can be related to the probability of a positive response. Maddala (1983) and Scheers and Dayton (1988) incorporated explanatory variables in the randomized response model. The additional information can be used to reduce the standard errors and establish a relationship between the covariate information and the population proportion with the sensitive characteristic.

RR data can be hierarchically structured, and there is interest in group differences regarding some sensitive characteristic. One could think of an application where it is of interest to know if cheating behavior differs across faculties or if social security fraud is more likely to appear in groups with certain characteristics. The classical (e.g., related and unrelated) randomized response models do not allow a hierarchical data analysis. Statistical methods for hierarchically structured data, for example analysis of variance (ANOVA) or multilevel analysis, cannot be applied since the individual responses are randomized.

A major drawback of classical randomized response models is that only aggregate-level inferences can be obtained; that is, estimates of population proportions and related confidence intervals. It is not possible to make individual-level inferences, and this prevents insights into the possible determinants and consequences of the sensitive construct under study. This corresponds to the fact that related and unrelated randomized response methods were originally developed for analyzing univariate data.

In this chapter, the randomized response technique will be extended to a multivariate setting with the purpose of measuring an underlying sensitive construct. The motivation is that the quality of the measurements is improved by the randomized response technique such that a more reliable estimate of the sensitive individual characteristic can be obtained. Covariates concerning individual or group characteristics can be taken into account in the estimation of the individual sensitive attitudes. It is also possible, the other way

around, to explore characteristics that can be held responsible for individual differences in (sensitive) attitudes. The differential influence of individual and group characteristics on sensitive individual characteristics can be more accurately investigated.

9.4 A Mixed Effects Randomized Item Response Model

Assume that there are J groups and n_j individuals nested within each group. Let Y_{ijk} denote the randomized response for an individual, indexed ij, to an item indexed k. Subsequently, $\widetilde{Y}_{ijk}$ denotes the latent (true) item response in the randomized response design. Note that this latent response is never observed directly since each individual response is randomized and only the randomized response is observed.

9.4.1 Individual Response Probabilities

The general randomized response model

$$\begin{aligned} P\left(Y_{ijk} = c \mid \pi_{ijk}\right) &= \tilde{p}_1 \pi_{ijk}(c) + \tilde{p}_2 \\ &= f\left(\pi_{ijk}(c); \tilde{p}_1, \tilde{p}_2\right) \end{aligned} \tag{9.5}$$

includes the related, the unrelated, and the forced randomized response models, where for the related (Warner's) model $\tilde{p}_1 = (2p_1 - 1)$ and $\tilde{p}_2 = (1 - p_1)$, for the unrelated question model $\tilde{p}_1 = p_1$ and $\tilde{p}_2 = (1 - p_2)\pi^u_{ijk}(c)$, and for the forced response model $\tilde{p}_1 = p_1$ and $\tilde{p}_2 = (1 - p_1)p_2(c)$ according to Equations (9.1) and (9.3). The $\pi_{ijk}(c)$ is the response probability of person i in group j scoring a latent response $\widetilde{Y}_{ijk} = c$ and $\pi^u_{ijk}(c)$ the probability of the latent response in category c to the unrelated nonsensitive question. Although the unrelated question method fits within this framework, this randomized response design will not be pursued further. Note that $\tilde{p}_1$ and $\tilde{p}_2$ are known specific characteristics of the randomizing device that is used in the randomized response survey.

Univariate Response Data

At the second stage, the latent (true) responses are modeled. Note that, strictly speaking, from the hierarchical point of view, the latent response data are also modeled at stage one. Suppose that $\widetilde{Y}_{ijk}$ takes on value zero or one with probability $1 - \pi_{ijk}$ and π_{ijk}, respectively. Then, let Z_{ijk} be a continuous latent variable such that $\widetilde{Y}_{ijk} = 1$ if Z_{ijk} is positive and $\widetilde{Y}_{ijk} = 0$ if Z_{ijk} is negative. A probit model is defined for the latent response as

$$\pi_{ijk} = P\left(\widetilde{Y}_{ijk} = 1\right) = P(Z_{ijk} > 0), \tag{9.6}$$

where Z_{ijk} is standard normally distributed. A logistic response function can be assumed, and then Z_{ijk} is standard logistically distributed.

For polytomous ordinal data, other polytomous responses can be handled in a similar way, where the latent response $\widetilde{Y}_{ijk}$ denotes a categorical outcome and Z_{ijk} the underlying latent score such that the probability of individual ij scoring in category c equals

$$\pi_{ijk}(c) = P\big(\widetilde{Y}_{ijk} = c \mid \boldsymbol{\kappa}\big) = \Phi\left(z_{ijk} - \kappa_{k,c-1}\right) - \Phi\left(z_{ijk} - \kappa_{k,c}\right), \quad (9.7)$$

or replace $\Phi(.)$ with

$$\Psi\left(z_{ijk} - \kappa_{k,c}\right) = \frac{1}{1 + \exp[-(z_{ijk} - \kappa_{k,c})]},$$

where $\boldsymbol{\kappa}$ are the threshold parameters such that $\kappa_{k,r} > \kappa_{k,s}$ whenever $r > s$, with $\kappa_{k,0} = -\infty$ and $\kappa_{k,C} = \infty$.

Multivariate Response Data

It will be assumed that the items are composed to measure some underlying sensitive construct. For multivariate observed RR data, the second modeling stage consists of an item response model for defining the relationship between the observed randomized item responses and the underlying latent sensitive characteristic. The latent categorical outcome, $\widetilde{Y}_{ijk}$, represents the item response of person ij on item k with latent ability or attitude parameter θ_{ij}.

For dichotomous item responses, a two-parameter item response model is used for specifying the relation between the level of a latent variable and the probability of a particular item response. That is,

$$\begin{aligned}\pi_{ijk} &= P\left(\widetilde{Y}_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right) \\ &= \Phi\left(a_k\theta_{ij} - b_k\right). \end{aligned} \quad (9.8)$$

For polytomous item responses, the probability that an individual obtains a grade c on item k is defined by a graded response model

$$\begin{aligned}\pi_{ijk} &= P\left(\widetilde{Y}_{ijk} = c \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k\right) \\ &= \Phi\left(a_k\theta_{ij} - \kappa_{k,c-1}\right) - \Phi\left(a_k\theta_{ij} - \kappa_{k,c}\right) \end{aligned} \quad (9.9)$$

where the boundaries between the response categories are represented by an ordered vector of thresholds $\boldsymbol{\kappa}$. For the logistic item response model, replace $\Phi(.)$ with $\Psi(.)$. Note that these item response models correspond to the (likelihood) models in Chapter 4. The difference is that the item response models in this chapter are meant for relating, at the second stage, the underlying sensitive construct to the latent item responses.

9.4.2 A Structural Mixed Effects Model

At the third stage, the latent continuous response in Equation (9.6) can be modeled as a function of incidence matrices $\mathbf{x}_{ij}$ and $\mathbf{w}_{ij}$,

$$Z_{ijk} = \mathbf{w}_{ij}^t \boldsymbol{\gamma} + \mathbf{x}_{ij}^t \boldsymbol{\beta}_j + e_{ijk}, \tag{9.10}$$

where $\mathbf{x}_{ij}$ is the design vector for the random effects. The Q-dimensional vector $\boldsymbol{\beta}_j$ contains the random effects, and their distribution is assumed to be multivariate normal with mean zero and covariance matrix $\mathbf{T}$. Furthermore, $\boldsymbol{\gamma}$ is an S-dimensional vector of fixed effects, and the residuals e_{ijk} have mutually independent normal distributions with mean zero and variance one. There is independence between random effects of different groups, and the random effects are independent of the residuals e_{ijk}.

For the multivariate case, at the third stage, the latent sensitive construct θ_{ij} can be modeled using characteristics at the individual or group level. At this third stage, effects of group-level variables on the individual's binary or ordinal true response may vary across groups. The mixed effects model for the vector $\boldsymbol{\theta}_j$ of length n_j can be written as

$$\boldsymbol{\theta}_j = \mathbf{w}_j \boldsymbol{\gamma} + \mathbf{x}_j \boldsymbol{\beta}_j + \mathbf{e}_j, \tag{9.11}$$

where $\boldsymbol{\gamma}$ is the vector of fixed effects and $\boldsymbol{\beta}_j$ the vector of random regression effects. The prior distributional assumptions of the random effects and residuals are given by

$$\begin{aligned} \boldsymbol{\beta}_j &\sim \mathcal{N}(0, \mathbf{T}), \\ \mathbf{e}_j &\sim \mathcal{N}\left(0, \sigma_\theta^2 \mathbf{I}_{n_j}\right). \end{aligned}$$

This mixed effects model can be presented as a multilevel model. A partitioning is made in a level-1 model,

$$\theta_{ij} = \mathbf{w}_{ij}^{t(1)} \boldsymbol{\gamma}^{(1)} + \mathbf{x}_{ij}^{t(1)} \boldsymbol{\beta}_j + e_{ij}, \tag{9.12}$$

and a level-2 model,

$$\boldsymbol{\beta}_j = \mathbf{w}_j^{t(2)} \boldsymbol{\gamma}^{(2)} + \mathbf{u}_j,$$

where $\mathbf{w}_{ij}^{(1)}$ and $\mathbf{w}_j^{(2)}$ are the fixed level-1 and fixed level-2 covariates, respectively. An overview of structural multilevel models is given in Chapter 6.

The combination of a randomized response model for the observed RR data (stage 1), an individual response model for latent item response data (stage 2), and a structural mixed effects model for the latent (sensitive) construct (stage 3) constitutes a mixed effects randomized item response theory (RIRT) model.

Fox (2005c) proposed an RIRT model for analyzing multivariate binary RR data. In a Bayesian framework, the model was applied in a study on cheating behavior of students nested in studies at a Dutch university. Böckenholt

and van der Heijden (2007) introduced an RIRT model for binary RR data in a frequentist framework with a specific interest in measuring noncompliance. Fox and Wyrick (2008) developed an RIRT model that simultaneously handles binary and polytomous RR data motivated by a study for measuring alcohol dependence among students. De Jong, Pieters and Fox (2010) used an RIRT model in an empirical application to consumers' desires for adult entertainment.

In Figure 9.1, a path diagram for the RIRT model is given. This path diagram is closely related to the MLIRT model discussed in Chapter 6. It can be seen that in comparison with Figure 6.1 an ellipse is added. This ellipse depicts the relationship between the observed randomized item response and the latent item response given parameters $\tilde{\mathbf{p}}$. This relationship is specified by the randomized response sampling design, where parameters $\tilde{\mathbf{p}}$ define the characteristics of the randomizing device. The structural part of the model is used to explain between-school and within-school variability in the latent variable. The item response measurement part, also depicted as an ellipse, relates the latent variable to the latent discrete responses. Besides the three levels of uncertainty that are modeled in the item response measurement part and the structural mixed effects part, an additional level of uncertainty is added due to the randomized response sampling design.

The mixed effects randomized item response model can be identified in the same way as in the MLIRT model (see Chapter 6). The scale of the latent variable can be identified via restrictions on the item parameters or by restricting the mean and variance of the latent variable. Both ways of identifying the model lead to the same results but on a different scale, depending on the types of restrictions.

9.5 Inferences from Randomized Item Response Data

A general probabilistic relationship is defined between the randomized item response data and the latent item response data. Another variable defined is $H_{ijk} = 1$ when for respondent ij the randomizing device determines that item k is to be answered truthfully and $H_{ijk} = 0$ otherwise. The conditional probability that individual ij scores a latent response in category c' given an observed randomized response in category c can be stated as

$$\begin{aligned} P\left(\widetilde{Y}_{ijk} = c' \mid Y_{ijk} = c\right) &= \frac{P\left(\widetilde{Y}_{ijk} = c', Y_{ijk} = c\right)}{P\left(Y_{ijk} = c\right)} \\ &= \frac{\sum_{l\in(0,1)} P\left(\widetilde{Y}_{ijk} = c', Y_{ijk} = c \mid H_{ijk} = l\right) P\left(H_{ijk} = l\right)}{\sum_{l\in(0,1)} P\left(Y_{ijk} = c \mid H_{ijk} = l\right) P\left(H_{ijk} = l\right)}, \end{aligned} \tag{9.13}$$

where $c, c' = \{0, 1\}$ for binary response data or $c, c' = \{1, 2, \ldots, C_k\}$ for polytomous response data.

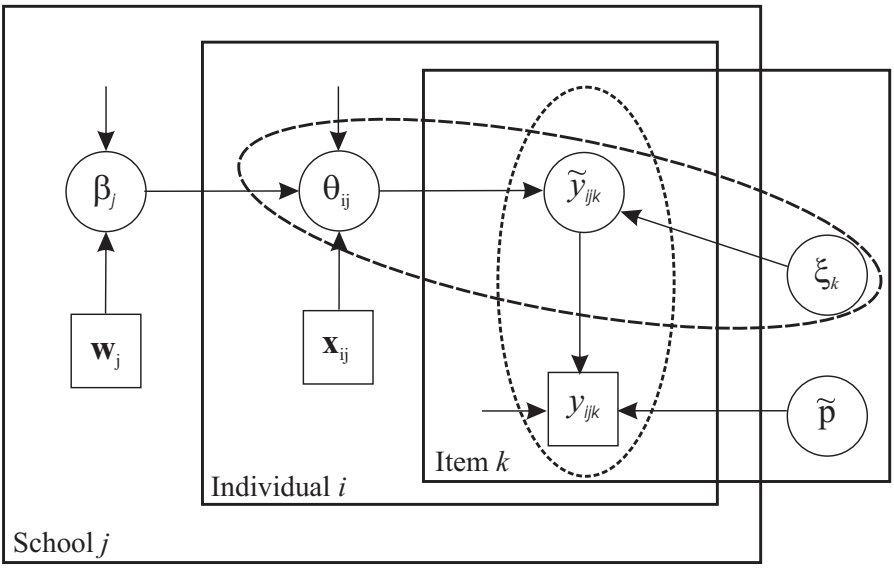

Fig. 9.1. Path diagram for the mixed effects randomized item response model.

The conditional probability of a latent response given an observed randomized response can be derived from Equation (9.13) for different randomized response sampling designs and for binary as well as polytomous item response data. To illustrate, consider the forced randomized response design for polytomous data. The denominator in Equation (9.13) equals the marginal probability of observing a response in category c. This observed response can be a latent response or a forced response. That is, the respondent, indexed ij, is instructed by the randomizing device to give a true response or a forced response. The first and second events take place with probabilities $p_1\pi_{ijk}(c)$ and $(1-p_1)p_2(c)$, respectively. Let the numerator in Equation (9.13) be equal to the simultaneous probability of the event that a response is observed in category c and that individual ij gives a latent response in category c. With probability $\pi_{ijk}(c)$, a latent response is given that is observed with probability p_1 or a forced response is observed with probability $(1-p_1)p_2(c)$. Both events take place with probability $\pi_{ijk}(p_1 + (1-p_1)p_2(c))$. As a result, the fraction gives the conditional probability of a latent response in category c for item k of individual ij given an observed randomized response in category c. The other cases can be derived in the same way.

As a result, the conditional probability of a latent polytomous response given an observed polytomous response using the forced randomized response sampling design can be defined:

$$P\left(\widetilde{Y}_{ijk} = c' \mid Y_{ijk} = c\right) = \begin{cases} \frac{\pi_{ijk}(c')(p_1+(1-p_1)p_2(c))}{\pi_{ijk}(c)p_1+(1-p_1)p_2(c)} & \text{if} \quad c = c' \\ \frac{\pi_{ijk}(c')(1-p_1)p_2(c)}{\pi_{ijk}(c)p_1+(1-p_1)p_2(c)} & \text{if} \quad c \neq c'. \end{cases} \tag{9.14}$$

Assume the structural model defined in Equation (9.10). Then the likelihood part of the RIRT model can be stated as

$$p\left(\mathbf{y} \mid \boldsymbol{\xi}, \sigma_\theta^2, \boldsymbol{\gamma}, \mathbf{T}\right) = \prod_{j=1}^{J} \left[\int \left[\prod_{i=1|j}^{n_j} \int \sum_{\tilde{\mathbf{y}}_{ij} \in \tilde{\mathcal{Y}}_{ij}} p\left(\mathbf{y}_{ij} \mid \tilde{\mathbf{y}}_{ij}\right) p\left(\tilde{\mathbf{y}}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}_k\right) \right.\right.$$
$$\left.\left. p\left(\theta_{ij} \mid \mathbf{x}_{ij}, \mathbf{w}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}, \sigma_\theta^2\right) \mathrm{d}\theta_{ij} \right] p\left(\boldsymbol{\beta}_j \mid \mathbf{T}\right) \right] \mathrm{d}\boldsymbol{\beta}_j, \quad (9.15)$$

where $\tilde{\mathcal{Y}}_{ij}$ is the set of all possible latent response vectors. The various levels of the RIRT model (Figure 9.1) can also be recognized from the likelihood in Equation (9.15).

The distribution of the randomized item response data given latent item response data is defined by the randomized response sampling design, $p\left(\mathbf{y} \mid \tilde{\mathbf{y}}\right)$. The distribution of the latent item responses, $p(\tilde{\mathbf{y}}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}_k)$, represents the item response model at the lowest level of the RIRT model. The second level is represented by the conditional distribution of the latent variable, $p\left(\theta_{ij} \mid \mathbf{x}_{ij}, \mathbf{w}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}, \sigma_\theta^2\right)$, and at the third level, the distribution of the random group effects is given, $p\left(\boldsymbol{\beta}_j \mid \mathbf{T}\right)$. Note that the randomized response model is positioned at the first level of the RIRT model since the latent item observations are masked before they are observed.

Before an MCMC implementation is discussed, a short overview of the prior distributions is given. The joint prior density of $\left(\boldsymbol{\xi}, \boldsymbol{\gamma}, \sigma_\theta^2, \mathbf{T}\right)$ can be factorized as

$$p\left(\boldsymbol{\xi}, \boldsymbol{\gamma}, \sigma_\theta^2, \mathbf{T}\right) = p(\boldsymbol{\xi}) p(\boldsymbol{\gamma}) p\left(\sigma_\theta^2\right) p(\mathbf{T}).$$

For the item parameters, noninformative proper priors for the discrimination and difficulty parameters are used, although other (conjugated) priors can be used as well; see Chapter 4. The prior for item parameters in the graded response model is specified as

$$p(\boldsymbol{\xi}) \propto \prod_k \phi\left(a_k; \mu_a, \sigma_a^2\right) I(a_k > 0) I(\kappa_{k,1}, \ldots, \kappa_{k,C_k} \in \mathcal{A}), \quad (9.16)$$

subject to the condition $\kappa_{k,0} < \kappa_{k,1} < \ldots < \kappa_{k,C_k}$ with $\kappa_{k,0} = -\infty$ and $\kappa_{k,C_k} = \infty$. Furthermore, $\mathcal{A}$ is a sufficiently large bounded interval.

According to the specification of the linear mixed effects model, the latent variable is assumed to have a normal distribution with mean $\mathbf{w}\boldsymbol{\gamma} + \mathbf{x}\boldsymbol{\beta}$ and variance σ_θ^2. The fixed effects, $\boldsymbol{\gamma}$, are assumed to have independent normal priors, with mean zero and variance σ_γ, and the hyperparameter σ_γ equals a large number that reflects a noninformative prior.

The prior for the covariance matrix $\mathbf{T}$ is taken to be an inverse Wishart density

$$p\left(\mathbf{T} \mid n_q, \mathbf{S}\right) \propto |\mathbf{T}|^{-(n_q+Q+1)/2} \exp\Big(-\frac{1}{2} \operatorname{tr}\left(\mathbf{S}\mathbf{T}^{-1}\right)\Big)$$

with unity matrix $\mathbf{S}$ and hyperparameter $n_q \geq Q$ equal to a small number to specify a diffuse proper prior. The conventional prior for σ_θ^2 is the inverse gamma with prior parameters g_1 and g_2.

9.5.1 MCMC Estimation

The joint posterior distribution, combining the likelihood in Equation (9.15) with the specified prior distributions, is intractable analytically, but MCMC methods such as the Gibbs sampler and the M-H algorithm can be used to draw samples. An MCMC scheme will be presented for binary as well as polytomous response data observed via the related or forced randomized response sampling design. The MCMC scheme consists of three parts. In part A, the latent (discrete) response data are sampled given the observed RR data. In part B, latent continuous responses are sampled given the latent discrete responses. Finally, in part C, the parameters are sampled given the latent continuous responses. Let $\pi_{ijk}^{(m)}(c)$ denote the probability of a latent response in category c according to a normal ogive or logistic item response model given latent variable and item parameter values sampled at iteration m.

MCMC Scheme 7 (RIRT)

A1) Binary response data

1. *The related-question design. Via Equation* (9.13)*,*

$$\widetilde{Y}_{ijk} \mid Y_{ijk} = 1, \pi_{ijk}^{(m)} \sim \mathcal{B}\left(\lambda = \frac{p_1 \pi_{ijk}}{p_1 \pi_{ijk} + (1-p_1)(1-\pi_{ijk})}\right),$$
$$\widetilde{Y}_{ijk} \mid Y_{ijk} = 0, \pi_{ijk}^{(m)} \sim \mathcal{B}\left(\lambda = \frac{(1-p_1)\pi_{ijk}}{p_1(1-\pi_{ijk}) + (1-p_1)\pi_{ijk}}\right),$$

   where λ *defines the success probability of the Bernoulli distribution.*

2. *The unrelated-question design. The latent item responses are Bernoulli distributed,*

$$\widetilde{Y}_{ijk} \mid Y_{ijk} = 1, \pi_{ijk}^{(m)} \sim \mathcal{B}\left(\lambda = \frac{\pi_{ijk}\left(p_1 + p_2(1-p_1)\right)}{p_1 \pi_{ijk} + p_2(1-p_1)}\right),$$
$$\widetilde{Y}_{ijk} \mid Y_{ijk} = 0, \pi_{ijk}^{(m)} \sim \mathcal{B}\left(\lambda = \frac{\pi_{ijk}(1-p_1)(1-p_2)}{1 - \left(p_1 \pi_{ijk} + p_2(1-p_1)\right)}\right).$$

A2) Polytomous response data

For the unrelated-question design, latent variable $\widetilde{Y}_{ijk}$ *given* $Y_{ijk} = c$ *and* $\pi_{ijk}^{(m)}$ *is multinomially distributed with cell probabilities*

$$\Delta(c) = \frac{\pi_{ijk}(c')p_1 I(c = c') + \pi_{ijk}(c')(1-p_1)p_2(c)}{\pi_{ijk}(c)p_1 + (1-p_1)p_2(c)}.$$

B1) Binary latent response data
Sample augmented data $\mathbf{z}^{(m+1)}$ *and item parameters* $\boldsymbol{\xi}^{(m+1)}$ *given latent response data* $\widetilde{\mathbf{y}}^{(m+1)}$ *and* $\boldsymbol{\theta}^{(m)}$ *according to step A1 in MCMC scheme 4 with* $\mathbf{y}^{(m+1)} = \widetilde{\mathbf{y}}^{(m+1)}$.

B2) Polytomous latent response data
Sample augmented data $\mathbf{z}^{(m+1)}$ *and item parameters* $\boldsymbol{\xi}^{(m+1)}$ *given latent response data* $\widetilde{\mathbf{y}}^{(m+1)}$ *and* $\boldsymbol{\theta}^{(m)}$ *according to step A2 in MCMC scheme 4 with* $\mathbf{y}^{(m+1)} = \widetilde{\mathbf{y}}^{(m+1)}$.

C) Sample structural mixed effects model parameters

1. *Given normally distributed latent continuous responses, the full conditional density of* $\boldsymbol{\theta}$ *equals*

$$\theta_{ij} \mid \mathbf{z}_{ij}^{(m+1)}, \boldsymbol{\xi}^{(m+1)}, \boldsymbol{\beta}_j^{(m)}, \boldsymbol{\gamma}^{(m)}, \sigma_\theta^{2(m)} \sim \mathcal{N}\left(\mu_\theta, \Omega_\theta\right), \quad (9.17)$$

where

$$\mu_\theta = \Omega_\theta \left(\left(\mathbf{a}^t\mathbf{a}\right) \widehat{\theta}_{ij} + (\mathbf{w}_{ij}^t\boldsymbol{\gamma} + \mathbf{x}_{ij}^t\boldsymbol{\beta}_j)/\sigma_\theta^2 \right), \quad (9.18)$$

$$\Omega_\theta = \left(\mathbf{a}^t\mathbf{a} + \sigma_\theta^{-2}\right)^{-1}, \quad (9.19)$$

and $\widehat{\theta}_{ij}$ *is the least squares estimate following from the regression of* $\mathbf{z}_{ij} + \mathbf{b}$ *on* $\mathbf{a}$ *for binary data and* $\mathbf{z}_{ij}$ *on* $\mathbf{a}$ *for polytomous data. This full conditional is used as a proposal density in an M-H algorithm when the continuous latent responses are logistically distributed.*

2. *The random regression effects* $\boldsymbol{\beta}_j$ *are conditionally multivariate normally distributed,*

$$\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j^{(m+1)}, \boldsymbol{\gamma}^{(m)}, \sigma_\theta^{2(m)}, \mathbf{T}^{(m)} \sim \mathcal{N}\left(\mathbf{D}\mathbf{x}_j^t(\boldsymbol{\theta}_j - \mathbf{w}_j\boldsymbol{\gamma})/\sigma_\theta^2, \mathbf{D}\right),$$

where

$$\mathbf{D}^{-1} = \left(\mathbf{x}_j^t\mathbf{x}_j/\sigma_\theta^2 + \mathbf{T}^{-1}\right).$$

3. *The fixed effects are conditionally multivariate normally distributed,*

$$\boldsymbol{\gamma} \mid \boldsymbol{\theta}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \sigma_\theta^{2(m)}, \mathbf{T}^{(m)} \sim \mathcal{N}\left(\mu_\gamma, \Omega_\gamma\right),$$

where

$$\mu_\gamma = \Omega_\gamma \mathbf{w}^t(\boldsymbol{\theta} - \mathbf{x}\boldsymbol{\beta}),$$
$$\Omega_\gamma = \left(\mathbf{w}^t\mathbf{w} + \sigma_\gamma^{-1}\sigma_\theta^2\mathbf{I}_S\right)^{-1}.$$

4. *Variance parameter* σ_θ^2 *is conditionally distributed as inverse gamma with parameter* $g_1 + N/2$ *and scale parameter* $g_2 + \sum_{i|j}(\theta_{ij} - (\mathbf{w}_{ij}\boldsymbol{\gamma} + \mathbf{x}_{ij}\boldsymbol{\beta}_j))^2/2$.
5. *Covariance matrix* $\mathbf{T}$ *is conditionally distributed as inverse Wishart with degrees of freedom* $n_q + J$ *and scale parameter* $\mathbf{S} + \sum_j \boldsymbol{\beta}_j\boldsymbol{\beta}_j^t$.

9.5.2 Detecting Noncompliance Behavior

Although the randomized response design protects the privacy of respondents, some of them may not be convinced and/or do not trust the privacy protection. Respondents may not follow the randomization scheme and always answer the least stigmatizing category regardless of the question asked. This behavior is called cheating (Clark and Desharnais, 1998) or noncompliance (e.g, Böckenholt and van der Heijden, 2007; Cruyff, van den Hout, van der Heijden and Böckenholt, 2007; van den Hout and Klugkist, 2009). This noncompliance behavior is characterized as consistently selecting the least self-incriminating response categories.

Clark and Desharnais (1998) developed a method for estimating the magnitude of the noncompliance behavior using two sample groups that are confronted with different randomized response designs. Böckenholt and van der Heijden (2007) developed a latent two-class model where one group consists of respondents that follow the randomized response design instructions and a second group of respondents do not follow the instructions and show noncompliance behavior. The latent class approach opens the possibility of handling different types of noncompliance behavior induced by different randomized response designs.

The RIRT model can be extended to handle noncompliance response behavior via a latent class structure. In the most straightforward way, assume that noncompliance response behavior is characterized by consistently responding to the least stigmatizing option (say zero in the case of binary data or one in the case of polytomous data). Respondents exhibiting noncompliance behavior answer with probability one the least self-incriminating response, and their response patterns do not provide any information about the items' characteristics.

A binary latent class variable G_{ijk} is defined such that $G_{ijk} = 1$ when respondent ij answers item k according to self-protective response behavior (noncompliance) and $G_{ijk} = 0$ when respondent ij answers item k according to the randomized response mechanism. Let $\nu_{ij} = P(G_{ijk} = 1)$ denote the conditional probability that respondent ij answers question k in the least self-incriminating way. For binary response data, it follows that

$$\begin{aligned} P(Y_{ijk} = 0) &= (1 - \nu_{ij}) P(Y_{ijk} = 0 \mid \boldsymbol{\xi}_k, \theta_{ij}) + \nu_{ij} I(Y_{ijk} = 0) \\ &= (1 - \nu_{ij}) f(1 - \pi_{ijk}; \tilde{p}_1, \tilde{p}_2) + \nu_{ij} I(Y_{ijk} = 0), \end{aligned} \tag{9.20}$$

where the first category is considered to be the least incriminating response and π_{ijk} is the success probability of respondent ij in the compliance class (Equation (9.8)) given a randomized response design (Equation (9.5)).

The full model is a mixture model consisting of an RIRT model for the compliance class and a model for the noncompliance class. The mixing weight parameter ν_{ij} combines the two models. This mixture model can be recognized from Equation (9.20), where the observed response is distributed with

probability $(1 - \nu_{ij})$ according to the RIRT model and with probability ν_{ij} according to the noncompliance response model. When there is no noncompliance response behavior, the mixture model resembles the RIRT model.

Inferences concerning the parameters of interest in the RIRT model are to be based on the response data of the respondents in the compliance class. This requires knowledge of the status of the latent class variable G_{ijk}. The conditional distribution of the latent class variable G_{ijk} is Bernoulli; that is,

$$G_{ijk} \mid \nu_{ij}, \mathbf{y}_{ij}, \boldsymbol{\pi}_{ij} \sim \mathcal{B}\left(\frac{\nu_{ij} I\left(Y_{ijk} = 0\right)}{\nu_{ij} I\left(Y_{ijk} = 0\right) + (1 - \nu_{ij}) f\left(1 - \pi_{ijk}\right)}\right), \quad (9.21)$$

where the success probability specifies the conditional probability that respondent ij answers in the least self-incriminating way to item k. Assume a beta prior with parameters α and β for the mixing weight ν_{ij}. The full conditional posterior of ν_{ij} is beta; that is,

$$\nu_{ij} \mid \mathbf{g}_{ij} \sim \mathcal{B}e\left(\alpha + n_g, \beta + K - n_g\right), \quad (9.22)$$

where $n_g = \sum_k I(G_{ijk} = 1)$. MCMC scheme 7 is easily extended to handle noncompliance behavior by adding the sampling steps in Equations (9.21) and (9.22). Subsequently, in part A, latent discrete response data are sampled for observed responses that stem from the RIRT model. In that case, the corresponding latent class variables are zero. The other sampling steps remain the same.

The prior distribution of the mixing weights can be made more informative by adding related background information on the respondents. This might also decrease the uncertainty in the estimation of the mixing weights and other model parameters. When interest is focused on the mixing weights, a structural model can be defined on the parameters $\boldsymbol{\nu}$ to explain individual and group differences and to investigate relationships to background variables.

9.5.3 Testing for Fixed-Group Differences

In the structural mixed effects model, Equation (9.12), a grouping structure is modeled via random effects parameters. The groups are assumed to be sampled from a population, and interest is focused on the characteristics of this population distribution. A different analysis is required when inferences are to be made about those groups that are in the sample.

Assume that a grouping structure is defined by indicator variables that are stored in a design matrix $\mathbf{w}$ and that matrix $\mathbf{w}^t\mathbf{w}$ is nonsingular. Attention is focused on the grouping structure in the latent variable $\boldsymbol{\theta}$. A linear hypothesis is considered where the expectation of a $\boldsymbol{\theta}$ is a known linear function of unknown parameters $\boldsymbol{\gamma}$, $E(\boldsymbol{\theta}) = \mathbf{w}\boldsymbol{\gamma}$.

Interest is in a point null hypothesis of the form $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ and an alternative hypothesis $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0$. Therefore, an HPD region is constructed from the conditional posterior distribution, and the null hypothesis is rejected if and only if

$\boldsymbol{\gamma}_0$ is outside the HPD region. A p-value p_0 can be defined as one minus the content of the HPD region that just covers $\boldsymbol{\gamma}_0$, which equals

$$p_0 = 1 - P\left[p\left(\boldsymbol{\gamma} \mid \boldsymbol{\theta}\right) > p\left(\boldsymbol{\gamma}_0 \mid \boldsymbol{\theta}\right) \mid \boldsymbol{\theta}\right]. \tag{9.23}$$

The null hypothesis is rejected when the p-value is less than or equal to a significance level α.

For the moment, the factor effects $\boldsymbol{\gamma}$ and the log of the within-group variance, $\log \sigma_\theta^2$, are assumed to have independent prior distributions that are uniform over the real line. The joint posterior distribution of $\boldsymbol{\gamma}$ and σ_θ^2 given $\boldsymbol{\theta}$ is

$$\begin{aligned} p\left(\boldsymbol{\gamma}, \sigma_\theta^2 \mid \boldsymbol{\theta}\right) &\propto \sigma_\theta^{-n-2} \exp\left(\frac{-1}{2\sigma_\theta^2}\left(\boldsymbol{\theta} - \mathbf{w}\boldsymbol{\gamma}\right)^t\left(\boldsymbol{\theta} - \mathbf{w}\boldsymbol{\gamma}\right)\right) \\ &\propto \sigma_\theta^{-n-2} \exp\left(\frac{-1}{2\sigma_\theta^2}\left(\left(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\right)^t \mathbf{w}^t\mathbf{w}\left(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\right) + S^2\right)\right), \end{aligned}$$

where $\hat{\boldsymbol{\gamma}} = \left(\mathbf{w}^t\mathbf{w}\right)^{-1}\mathbf{w}^t\boldsymbol{\theta}$ and S^2 is the residual sum of squares. Integrating with respect to σ_θ^2 leads to (see Box and Tiao, 1973, Chapter 2)

$$\begin{aligned} p\left(\boldsymbol{\gamma} \mid \boldsymbol{\theta}\right) &\propto \left[1 + \frac{\left(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\right)^t \mathbf{w}^t\mathbf{w}\left(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\right)}{S^2}\right]^{-\frac{n}{2}} \\ &\propto \left[1 + \frac{S_s^2(\boldsymbol{\gamma})/(n-s)}{S^2/(n-s)}\right]^{-\frac{n}{2}}. \end{aligned} \tag{9.24}$$

Lindley (1965, Chapter 8) and Box and Tiao (1973) showed that the marginal posterior in Equation (9.24) is a monotonically decreasing function of the $S_s^2(\boldsymbol{\gamma})$ and $S_s^2(\boldsymbol{\gamma}) = c$ defines a contour of the distribution in the s-dimensional space of $\boldsymbol{\gamma}$. Furthermore, they show that the term $(S_r^2(\boldsymbol{\gamma})/r)/(S^2/(n-s))$ is F-distributed with $(r, n-s)$ degrees of freedom for any $r \leq s$. That is, the contours of the marginal distribution of any subset r of the fixed factor effects $\boldsymbol{\gamma}$ are defined by an F-distribution and will define a confidence region.

The conditional posterior probability that $\boldsymbol{\gamma}_0$ is not included in the HPD region (Equation (9.23)) depends on the usually unknown parameter $\boldsymbol{\theta}$. The corresponding marginal posterior probability that a certain point is included can be computed via MCMC. The marginal posterior probability that an r-dimensional subset $\boldsymbol{\gamma}_0$ is not included in a $(1-\alpha)$ HPD region is specified by

$$\begin{aligned} \alpha &\geq 1 - P\left[p\left(\boldsymbol{\gamma} \mid \mathbf{y}\right) > p\left(\boldsymbol{\gamma}_0 \mid \mathbf{y}\right) \mid \mathbf{y}\right] \\ &\geq \int 1 - P\left(F(r, n-s) \leq \frac{S_r^2(\boldsymbol{\gamma}_0)/r}{S^2/(n-s)} \,\middle|\, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \mathrm{d}\boldsymbol{\theta} \\ &\geq \int p_0\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \mathrm{d}\boldsymbol{\theta} \approx \sum_m p_0\left(\boldsymbol{\theta}^{(m)}\right) \Big/ M, \end{aligned} \tag{9.25}$$

where $\boldsymbol{\theta}^{(m)}$ is an MCMC sample from the marginal posterior distribution. The conditional posterior probability is easily computed using the fact that the quantity is F-distributed and the marginal posterior probability is obtained using the MCMC sample. That is, the marginal posterior probability of interest is obtained by integrating over the parameter $\boldsymbol{\theta}$ using an MCMC sample from the marginal posterior.

9.5.4 Model Choice and Fit

The assumptions of the RIRT model can be evaluated via a Bayesian residual analysis (see Chapter 5). The Bayesian residual $R_{ijk} = Y_{ijk} - \sum_c c \cdot f(\pi_{ijk}(c))$ is the difference between the observed randomized response and the expected response. These residuals provide information about the fit at the level of the randomized response model. The residuals can be estimated in MCMC scheme 7. At another level of the model, latent residuals can be defined as $\widetilde{R}_{ijk} = \widetilde{Y}_{ijk} - \sum_c c \cdot \pi_{ijk}(c)$, which represents the difference between the latent response and the expected latent response. The latent residuals can also be estimated from MCMC scheme 7.

The fit of the items and the persons can be tested via two functions of sums of squared residuals,

$$Q_k(\mathbf{y}) = \sum_{i=1}^{N} R_{ik}^2,$$
$$Q_i(\mathbf{y}) = \sum_{k=1}^{K} R_{ik}^2.$$

Both functions can be defined in the same way for the latent residuals. Subsequently, the functions can be used as discrepancy measures in a posterior predictive check to evaluate the fit; that is,

$$P(Q_i(\mathbf{y}^{rep}) \geq Q_i(\mathbf{y}) \mid \mathbf{y}) = \sum_{\mathbf{y}^{rep} \in \mathcal{Y}^{rep}} I(Q_i(\mathbf{y}^{rep}) \geq Q_i(\mathbf{y}))\, p(\mathbf{y}^{rep} \mid \mathbf{y}),$$

where $\mathcal{Y}^{rep}$ is the set of all possible randomized response vectors of $\mathbf{y}^{rep}$. The extremeness of the realized discrepancy is evaluated via replicated data under the model. Relatively high function values correspond to a poor fit of the item or person using the marginal distribution of the replicated data to quantify the extremeness of the realized discrepancy.

A test of local independence can be performed by evaluating the conditional covariance between residuals concerning two items given the person parameters. Let $R_k(\tilde{\mathbf{y}}, \boldsymbol{\theta})$ denote the vector of residuals of item k given the latent response data and the latent person parameters. The covariance between two vectors of item residuals k and k' is denoted as $\sigma_{k,k'} = \mathrm{Cov}(R_k(\tilde{\mathbf{y}}, \boldsymbol{\theta}), R_{k'}(\tilde{\mathbf{y}}, \boldsymbol{\theta}))$. The assumption of local independence is violated

with significance level α when the point $\sigma_{k,k'} = 0$ is not included in the $(1-\alpha)$ HPD region; that is, when the posterior probability

$$\begin{aligned} p_0 &= P\left(p\left(\sigma_{k,k'} \mid \mathbf{y}\right) > p\left(\sigma_{k,k'} = 0 \mid \mathbf{y}\right) \mid \mathbf{y}\right) \\ &= \sum_{\tilde{\mathbf{y}} \in \tilde{\mathcal{Y}}} P\left(p\left(\sigma_{k,k'} \mid \tilde{\mathbf{y}}\right) > p\left(\sigma_{k,k'} = 0 \mid \tilde{\mathbf{y}}\right) \mid \tilde{\mathbf{y}}\right) p\left(\tilde{\mathbf{y}} \mid \mathbf{y}\right) \\ &= \sum_{\tilde{\mathbf{y}} \in \tilde{\mathcal{Y}}} \int P\left(p\left(\sigma_{k,k'} \mid \boldsymbol{\theta}, \tilde{\mathbf{y}}\right) > p\left(\sigma_{k,k'} = 0 \mid \boldsymbol{\theta}, \tilde{\mathbf{y}}\right) \mid \tilde{\mathbf{y}}\right) p\left(\boldsymbol{\theta} \mid \tilde{\mathbf{y}}\right) \mathrm{d}\boldsymbol{\theta} p\left(\tilde{\mathbf{y}} \mid \mathbf{y}\right) \end{aligned} \tag{9.26}$$

is greater than $1-\alpha$. This posterior probability can be computed via MCMC scheme 7.

A DIC can be computed to compare models with each other but requires an analytical expression for the likelihood. In a similar way as in deriving an expression for the likelihood of the MLIRT model in Appendix 6.10, an augmented likelihood expression is derived for the augmented data $\tilde{\mathbf{y}}$ such that the likelihood is expressed as an integrated augmented likelihood. Let the parameters of interest be defined as $\boldsymbol{\Lambda} = (\boldsymbol{\gamma}, \boldsymbol{\xi}, \sigma_\theta^2, \mathbf{T})$. Then

$$\begin{aligned} p\left(\mathbf{y} \mid \boldsymbol{\Lambda}\right) &= \sum_{\tilde{\mathbf{y}} \in \tilde{\mathcal{Y}}} p\left(\mathbf{y} \mid \tilde{\mathbf{y}}\right) p\left(\tilde{\mathbf{y}} \mid \boldsymbol{\Lambda}\right) \\ &= \int \sum_{\tilde{\mathbf{y}} \in \tilde{\mathcal{Y}}} p\left(\mathbf{y} \mid \tilde{\mathbf{y}}\right) p\left(\tilde{\mathbf{y}} \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \boldsymbol{\Lambda}\right) \mathrm{d}\mathbf{z} \\ &= \int \int \sum_{\tilde{\mathbf{y}} \in \tilde{\mathcal{Y}}} p\left(\mathbf{y} \mid \tilde{\mathbf{y}}\right) p\left(\tilde{\mathbf{y}} \mid \mathbf{z}\right) \frac{p\left(\mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\Lambda}\right)}{p\left(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\Lambda}\right)} \mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta} \\ &= \int \int \int p\left(\mathbf{y} \mid \mathbf{z}\right) \frac{p\left(\mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)}{p\left(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)} \frac{p\left(\boldsymbol{\beta} \mid \mathbf{T}\right)}{p\left(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Lambda}\right)} \mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\beta}. \end{aligned} \tag{9.27}$$

As defined in MCMC scheme 7, there are two types of augmented data. The observed randomized responses are augmented with latent (discrete) responses. The discrete latent responses are augmented with continuous latent responses denoted as $\mathbf{z}$. The likelihood part $p\left(\mathbf{y} \mid \tilde{\mathbf{y}}\right)$ presents the fit of the randomized response model, and the part $p\left(\tilde{\mathbf{y}} \mid \mathbf{z}\right)$ is equal to one since the values of $\mathbf{z}$ completely specify the values of $\tilde{\mathbf{y}}$. These likelihood parts do not contain the parameters of interest and are not interesting for comparing different RIRT models since they are similar with respect to the randomized response model part. A DIC for comparing RIRT models that differ in the structural part and share the same randomized response part is constructed from the other likelihood terms in Equation (9.27). The corresponding log-likelihood term, $\log p\left(\mathbf{z} \mid \boldsymbol{\Lambda}\right)$, can be expressed as

$$\begin{aligned} \log p\left(\mathbf{z} \mid \boldsymbol{\Lambda}\right) = \frac{1}{2} \sum_j \Big[&-K n_j \log(2\pi) + n_j \log\left(\Omega_\theta\right) - n_j \log\left(\sigma_\theta^2\right) - \log\left|\mathbf{T}\right| \\ &+ \log\left|\boldsymbol{\Sigma}_{\tilde{\beta}_j}\right| - S\left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\beta}}_j\right) \Big], \end{aligned} \tag{9.28}$$

where

$$S\left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\beta}}_j\right) = \sum_{i,k}\left(z_{ijk} - \left(a_k\tilde{\theta}_{ij} - b_k\right)\right)^2 + \sigma_\theta^{-2}\left(\tilde{\boldsymbol{\theta}}_j - \mathbf{w}_j\boldsymbol{\gamma} - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right)^t \cdot \left(\tilde{\boldsymbol{\theta}}_j - \mathbf{w}_j\boldsymbol{\gamma} - \mathbf{x}_j\tilde{\boldsymbol{\beta}}_j\right) + \tilde{\boldsymbol{\beta}}_j^t\mathbf{T}^{-1}\tilde{\boldsymbol{\beta}}_j. \tag{9.29}$$

Furthermore, each $\tilde{\theta}_{ij}$ is independently normally distributed with mean μ_θ (Equation (9.18)) and variance Ω_θ (Equation (9.19)) with $\boldsymbol{\beta}_j = \tilde{\boldsymbol{\beta}}_j$. The $\tilde{\boldsymbol{\beta}}_j = E\left(\boldsymbol{\beta}_j \mid \mathbf{z}_j, \boldsymbol{\Lambda}\right)$ and $\boldsymbol{\Sigma}_{\tilde{\beta}_j}$ are defined in Equations (6.42) (with mean $\mathbf{w}_j\boldsymbol{\gamma}$) and (6.43), respectively.

The deviance for model comparison is defined as $D\left(\boldsymbol{\Lambda}\right) = -2\log p\left(\mathbf{z} \mid \boldsymbol{\Lambda}\right)$, and the DIC is defined as

$$\begin{aligned}\text{DIC} &= \sum_{\tilde{\mathbf{y}}\in\tilde{\mathcal{Y}}}\int \left[\text{DIC} \mid \mathbf{z}\right] p\left(\mathbf{z} \mid \tilde{\mathbf{y}}\right) \mathrm{d}\mathbf{z}\, p\left(\tilde{\mathbf{y}} \mid \mathbf{y}\right) \\ &= \sum_{\tilde{\mathbf{y}}\in\tilde{\mathcal{Y}}}\int \left[D\left(\widehat{\boldsymbol{\Lambda}}\right) + 2p_D \mid \mathbf{z}\right] p\left(\mathbf{z} \mid \tilde{\mathbf{y}}\right) \mathrm{d}\mathbf{z}\, p\left(\tilde{\mathbf{y}} \mid \mathbf{y}\right) \\ &= E_{\mathbf{z},\tilde{\mathbf{y}}}\left[D\left(\widehat{\boldsymbol{\Lambda}}\right) + 2p_D \mid \mathbf{y}\right],\end{aligned} \tag{9.30}$$

where $\widehat{\boldsymbol{\Lambda}}$ is the posterior mean of $\boldsymbol{\Lambda}$. MCMC scheme 7 can be used to estimate the DIC using the likelihood expression in Equation (9.28).

9.6 Simulation Study

In the first study, univariate randomized item response data were simulated under different randomized response designs. In the second study, hierarchically structured multivariate randomized item response data were simulated under different randomizing device characteristics. Interest was focused on the differential effects of different randomized response sampling designs or different design properties on the structural model parameter estimates.

9.6.1 Different Randomized Response Sampling Designs

A total of N binary latent responses, $\tilde{\mathbf{y}}$, divided at random across ten groups, were generated according to the (nonlinear) mixed effects model

$$P\left(\tilde{Y}_{ij} = 1\right) = \pi_{ij} = \Phi\left(\gamma_0 + u_{0j} + x_i\gamma_1\right), \tag{9.31}$$

where $u_{0j} \sim \mathcal{N}(0, \tau^2)$. The values of vector $\mathbf{x}$ were simulated from a normal distribution with mean zero and standard deviation $1/2$. Randomized response data, $\mathbf{y}$, were generated via Equation (9.5) with $p_1 = 4/5$ according

Table 9.1. Re-estimating mixed effects model parameters given latent or randomized responses.

			Latent Response		Warner		Forced	
N	Par.	True	Mean	SD	Mean	SD	Mean	SD
1000	γ_0	.00	−.06	.16	−.06	.16	−.07	.16
	γ_1	2.00	1.94	.17	1.81	.26	1.80	.20
	τ	.25	.29	.18	.20	.16	.24	.16
5000	γ_0	.00	−.12	.15	−.11	.16	−.12	.16
	γ_1	2.00	2.01	.07	2.01	.12	1.97	.09
	τ	.25	.23	.14	.22	.15	.24	.15

to Warner's model (related response design) and with $p_1 = 4/5$ and $p_2 = 2/3$ according to the forced response model (unrelated response design).

MCMC algorithm 7 was run for 20,000 iterations, convergence was obtained after 5,000 iterations, and the cumulative averages of sampled parameter values resembled the true parameter values. Table 9.1 presents the true values, posterior mean estimates, and standard deviations given the latent response vector $\tilde{\mathbf{y}}$, labeled Latent Response, for the mixed effects Warner model, labeled Warner, and for the mixed effects forced response model, labeled Forced. It is apparent that the point estimates resemble the true values for a sample size of N=5,000 and the estimates are close to the true values for a sample size of N=1,000. The Warner model has the largest estimated standard deviations with respect to parameter γ_1, which was also found by Scheers and Dayton (1988).

The proportion, π, of positive responses was estimated using M MCMC sampled values of latent response data $\tilde{\mathbf{y}}^{(m)}$,

$$\hat{\pi} = \frac{1}{MN} \sum_m \sum_{i,j} \tilde{y}_{ij}^{(m)}.$$

For N=5,000, the simulated proportion equals .464. The estimated proportion under the mixed effects Warner model equals .464 with standard deviation .007, and equals .465 with standard deviation .004 under the mixed effects forced response model. These point estimates resemble the estimated proportion using Warner's model, $\hat{\pi} = .462$ with standard deviation .010, and using the forced response model, $\hat{\pi} = .464$ with standard deviation .008.

For N=1,000, the simulated proportion equals .485 and the estimated proportions under the mixed effects Warner model and the mixed effects forced response model equal .488 (.014) and .480 (.009), respectively, where the standard deviations are given in parentheses. The point estimates under Warner's model and the forced response model equal .490 (.023) and .481 (.019), respectively. For both sample sizes, it can be concluded that the estimated proportions are comparable but that there is a reduction in sampling error

since the mixed effects model accounts for the grouping of individuals and utilizes the covariate information.

9.6.2 Varying Randomized Response Design Properties

Estimates of mixed effects model parameters are compared given directly observed and RR data for different randomizing proportions. The probability distribution of a latent behavior parameter, θ_{ij}, has the same form for each individual, but the parameters of that distribution vary over 20 groups. A level-2 model describes a linear relationship between the behaviors of $N = 1{,}000$ respondents and explanatory variables $\mathbf{x}$ and $\mathbf{w}$, and a level-3 model represents the distribution of the random effect parameters

$$\begin{aligned}\theta_{ij} &= \mathbf{x}_{ij}^t \boldsymbol{\beta}_j + w_{ij}\gamma^{(1)} + e_{ij},\\ \beta_{0j} &= \gamma_0^{(2)} + u_{0j},\\ \beta_{1j} &= \gamma_1^{(2)} + u_{1j},\end{aligned}$$

where $e_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$, $\mathbf{u}_j \sim \mathcal{N}\left(0, \mathbf{T}\right)$, and the random effects are independent of the residuals.

The first column of $\mathbf{x}$ consists of ones, and the second column and the vector $\mathbf{w}$ contain values generated from a normal distribution with mean zero and standard deviation .30. Latent responses to 10 items per subject, each with three ordinal response categories, were simulated using a graded response model. The latent responses were randomized via the unrelated-question design with $p_2 = 1/3$. The probability that the latent response matched the observed response (e.g., a truthful response was demanded), p_1, was considered to be one (which resembles a direct response), .80, and .60. A total of 100 datasets were analyzed for each value of p_1. Discrimination parameter values were sampled from a lognormal distribution, $a_k \sim \log N(\exp(1), 1/4)$. Threshold parameters, κ_{k1} and κ_{k2}, were sampled from a normal distribution with means $-1/2$ and $1/2$ (taking order restrictions into account), respectively, and variance $1/4$.

For identification of the model, the mean and variance of the latent behavior variable were scaled to the true simulated mean and variance, respectively. For each dataset, the graded response model parameters and the mixed effects model parameters were estimated simultaneously using 50,000 posterior draws. The burn-in period consisted of 5,000 iterations. Table 9.2 presents, for each model parameter, the true setup value, the average of the means, and standard deviations over the MCMC samples corresponding to the 100 datasets.

It can be seen that there is close agreement between the true and the average estimated means. For each model parameter, the average of posterior standard deviations resembled the standard deviation in the 100 estimated posterior means. Note that even for $p_1 = .60$, which means that 40% of the

Table 9.2. Generating values, means, and standard errors of recovered values.

		Direct Response		Forced			
		$p_1 = 1$		$p_1 = .80$		$p_1 = .60$	
Parameter	True	Mean	SD	Mean	SD	Mean	SD
Fixed Effects							
$\gamma^{(1)}$	.5	.50	.03	.50	.04	.49	.04
$\gamma_0^{(2)}$	0	.03	.11	.02	.11	−.08	.10
$\gamma_1^{(2)}$	0	.04	.11	.04	.12	.08	.10
Random Effects							
σ_e^2	1	1.00	.05	1.00	.05	.98	.05
τ_{00}	.3	.36	.08	.36	.09	.31	.08
τ_{11}	.2	.20	.05	.20	.06	.20	.06
τ_{01}	0	.13	.04	.14	.05	.00	.04

responses were forced responses, the estimated values resemble the true simulated values. The standard deviations of the mixed effect parameter estimates were not increasing, due to the incorporation of a randomized response sampling design. Furthermore, additional variance in the item parameter estimates, due to the randomized response sampling design, did not result in biased estimates.

9.7 Cheating Behavior and Alcohol Dependence

Data from two randomized response research studies were analyzed. In the first study, attention was focused on measuring cheating behavior of students at a Dutch university. The questionnaire consists of binary items. In the second study, attention was focused on measuring alcohol dependence of students of different colleges and universities in North Carolina. This questionnaire consists of polytomous items and was described in Section 5.3.1.

9.7.1 Cheating Behavior at a Dutch University

Detecting fraud is hard, and educational organizations often are not willing to expend the effort required to get to the bottom of cheating cases. On the other hand, student cheating and plagiarism is becoming an important problem with the rising number of ways to cheat on exams. The introduction of mobile phones and handheld computers has led to high-tech cheating with Web-equipped cell phones or handheld organizers. In 2002, a study was done to assess cheating behavior of students at a university in the Netherlands. The main targets were to investigate the number of students committing fraud, the reasons, and the different ways that students are cheating on exams.

Students received an email in which they were asked to participate in the study. The forced randomized response method was explained in the email to gain the respondents' confidence, so that they were willing to participate and to answer the questions truthfully. A website was developed containing 36 questions concerning cheating on exams and assignments. Note that types of cheating or academic dishonesty are measured by different items to capture the whole range of cheating. The 36 questionnaire items can be found in Fox and Meijer (2008). Each item was a statement, and respondents were asked whether they agreed or disagreed with it. When a student visited the website, an on-Web dice server rolled two dice before a question could be answered. The student answered "yes" when the sum of the outcomes equaled 2, 3, or 4, answered "no" when the sum equaled 11 or 12, or otherwise answered the sensitive question. In practice, a forced response was automatically given since it is known that some respondents find it difficult to lie (Fox and Tracy, 1986). As a result, the forced response technique was implemented with $p_1 = 3/4$ and $p_2 = 2/3$.

Data were available from 349 students (229 male students and 120 female students) from one of seven main disciplines: computer science (CS), educational science and technology (EST), philosophy of science (PS), mechanical engineering (ME), public administration and technology (PAT), science and technology (ST), and applied communication sciences (ACS). Within these seven disciplines, a stratified sample of students was drawn such that different majors were represented in proportion to their total number of students.

In the cheating literature, several variables, such as personal attributes, background characteristics, or situational factors, have been shown to be related to cheating behavior. For example, earlier cheating research focused on estimating population proportions of cheating across locations (small/large university), school types (private/public), or individual differences (male/female) (see, for example, Davis, Grover, Becker and McGregor, 1992). In the present study, background information was collected with respect to age, gender, year of major, number of hours per week spent on the major (less than 10 hours, 10–20 hours, 20–30 hours, 30–40 hours, and more than 40 hours), residence (on campus, in the city, with their parents), and lifestyle (active or passive). Respondents were guaranteed confidentiality, and the questionnaires were filled in anonymously.

Differences in Attitudes

In this study, the three fixed factors, gender, major, and year of major, were considered as student characteristics that are likely to explain differences in cheating behavior. A design matrix $\mathbf{w}$ was constructed that represent different grouping structures via indicator variables. The latent attitude parameter was assumed to be normally distributed with mean $\mathbf{w}\boldsymbol{\gamma}$ and standard deviation σ_θ. This structural part is the third stage of the RIRT model. The other stages of

the model are defined in Equations (9.5) and (9.8), where the discrimination parameters were fixed to one.

The RIRT model with the fixed factors was estimated given the randomized responses to the 36 items. The model was identified by fixing the mean of the scale of the attitude parameter to zero. The MCMC procedure contained 50,000 iterations and a burn-in period of 5,000 iterations. Convergence of the MCMC chain was checked using the standard convergence diagnostics from the BOA program. Plots of the runs and the diagnostic tests suggested a convergence of the MCMC chain.

In Table 9.3, the number of observations per group and the parameter estimates, including the estimated factor effects, are given. It can be seen that the estimated mean attitude for the males is slightly below zero. This means that female students cheat slightly more than male students. There is variation in attitudes across the seven majors, where computer science students hardly cheat and applied communication science students often cheat. It is apparent that freshmen cheated less than others. Third-year students and those in the sixth-year and above were more inclined to cheat.

Let γ denote the effect of being male. Then, using Equation (9.25), the tail-area probability of one minus the marginal posterior probability that $\gamma = 0$ (no gender differences) is just included in the HPD region was computed. This tail-area probability corresponds to the null hypothesis that there are no gender differences in attitudes. The effect of gender is not significant with a significance level of $\alpha = .01$. In the same way, it is concluded that there are no significant differences in levels of cheating across years in college.

Interpretation of the relationship between cheating and year of study is difficult because many characteristics (such as motivation, age, and experience) change as students progress through the grade levels. For example, it is known that cheating is negatively correlated with age. Finally, it is mentioned that attitudes towards cheating differ significantly across majors. The largest difference was found between CS and ACS students, with ACS students more inclined to cheat than CS students.

Item Level Analysis

To illustrate the advantages of the item response modeling approach, Figure 9.2 presents the ICCs of six items about different ways of cheating. The ICCs reveal that the items have different threshold values.

The ICCs of item 10 ("use of crib notes during an exam") and item 31 ("lying to postpone a deadline") have the lowest thresholds and thus represent the most popular ways of cheating of the six selected items (they are also the most popular methods among all 36 items). In contrast, the ICCs of item 2 ("received information using signals or sign language during an exam") and item 7 ("obtaining information outside the classroom or examination area") have higher thresholds and are thus less popular. Note that here the term popular refers to a preference of those respondents who are likely to cheat and

Table 9.3. Cheating study: Parameter estimates of the RIRT model with fixed factors gender, major, and year of major.

		RIRT Model			
	n_j	Mean	SD	HPD	p_0
Fixed Effects					
$\gamma^{(2)}$ (Intercept)		−.020	.123	[−.266, .217]	
Student variables					
Gender					.015
$\gamma_1^{(1)}$ (Male)	229	−.020	.163	[−.342, .294]	
Major					.001
$\gamma_2^{(1)}$ (CS)	50	−.438	.184	[−.821, −.090]	
$\gamma_3^{(1)}$ (PAT)	53	.180	.157	[−.144, .469]	
$\gamma_4^{(1)}$ (ACS)	53	.448	.161	[.129, .762]	
$\gamma_5^{(1)}$ (ST)	46	−.045	.176	[−.401, .288]	
$\gamma_6^{(1)}$ (EST)	66	.157	.159	[−.159, .466]	
$\gamma_7^{(1)}$ (ME)	49	−.157	.173	[−.498, .168]	
$\gamma_8^{(1)}$ (PS)	32	−.141	.211	[−.565, .259]	
Year of Major					.030
$\gamma_9^{(1)}$ (First)	52	−.661	.192	[−1.063, −.298]	
$\gamma_{10}^{(1)}$ (Second)	73	.026	.140	[−.250, .299]	
$\gamma_{11}^{(1)}$ (Third)	66	.235	.137	[−.041, .495]	
$\gamma_{12}^{(1)}$ (Fourth)	61	.085	.145	[−.210, .360]	
$\gamma_{13}^{(1)}$ (Fifth)	45	.078	.167	[−.268, .390]	
$\gamma_{14}^{(1)}$ (>Fifth)	52	.237	.160	[−.083, .546]	
Random Effects					
σ_θ^2 Residual		.857	.077	[.710, 1.007]	

does not refer to a general preference among the population of respondents. Item 34 ("submitted coursework from others without their knowledge") and item 12 ("looking at others' work with their knowledge during an exam") have the same item characteristics. Most interesting, however, is that from inspecting the ICCs (Figure 9.2) it can be concluded that using signals is a popular method for students with high cheating attitudes. The analysis showed that students who use signals are likely to be frequent cheaters.

In conclusion, not "everyone's doing it," but about 25% of the students admitted they have cheated on an exam. The analysis with the RIRT model revealed that cheating behavior varied across majors and that computer sci-

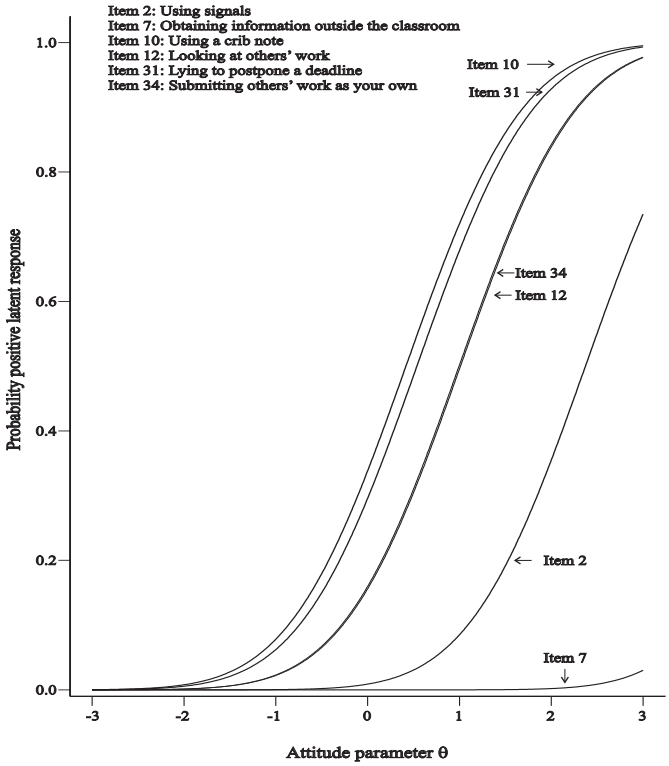

Fig. 9.2. Cheating study: Item characteristic functions for ways of cheating.

ence students are less likely to cheat. More details about the RIRT model cheating analysis can be found in Fox (2005c) and Fox and Meijer (2008).

9.7.2 College Alcohol Problem Scale

In Section 5.3.1, the response data of 351 students answering 13 items from the CAPS questionnaire (Appendix 5.9) were analyzed. The responses of the 351 students were obtained via direct questioning. The CAPS questions can be marked as sensitive since they will elicit answers that are socially unacceptable. Although confidentiality was promised, misreporting was expected due to socially desirable response behavior.

Whether the randomized response technique improved the quality of the self-reports was investigated. Therefore, each class of respondents (5–10 participants) was randomly assigned to either the direct questioning (DQ) or the randomized response technique condition. Random assignment at the individual level was not logistically feasible. The 351 students assigned to the DQ condition, denoted as the DQ group, were instructed to answer the questionnaire as they normally would. They served as the study's control group.

Table 9.4. CAPS: Gender and ethnicity demographics.

	Total Count	DQ Group (%)	RR Group (%)
Gender			
Female	502	65	62
Male	291	35	38
Ethnicity			
Asian	17	3	1
White	655	81	84
Black	93	12	11
Other	28	4	4

Note: DQ Group = direct questioning condition, RR Group = randomized response condition.

The 442 students in the RR condition, denoted as the RR group, received a spinner to assist them in completing the questionnaire. For each CAPS item, the participant spun the spinner, and wherever the arrow landed determined whether the item was to be answered honestly or dictated the answer choice to be recorded by the participant. The spinner was developed such that 60% of the area was comprised of answer honestly space and 40% of the area was divided into equal sections to represent the five possible answer choices. Each answer choice was given 8% of the area of the circle, 4% in two different places on the circle. The respondents from the DQ group and the RR group were assumed to be selected from the same population.

In total, 793 student participants from four local colleges and universities, Elon University (N=495), Guilford Technical Community College (N=66), University of North Carolina (N=166), and Wake Forest University (N=66), voluntarily responded to a questionnaire with 16 items in 2002. Three items of this questionnaire asked participants about their age, gender, and ethnicity (demographic information). In Table 9.4, the demographic (gender and ethnicity) percentages are given, and it follows that they are similar for the DQ group and RR group.

A unidimensional latent variable representing alcohol dependence, denoted as θ, was measured by the items, where a higher level indicated that a participant was more likely to have a drinking problem. A forced randomized response sampling design was implemented with $p_1 = .60$ and $p_2(c) = .20$ $(c = 1, \ldots, 5)$. All response data, observed via direct questioning and via the randomized response technique, were used to measure the latent behaviors (alcohol dependence) of the respondents on a common scale using the graded response model. This results in the level-1 part of the RIRT model,

$$P(Y_{ijk} = c \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k) = \begin{cases} p_1 \pi_{ijk}(c) + (1 - p_1) p_2(c) & i \in \text{RR} \\ \pi_{ijk}(c) & i \in \text{DQ}, \end{cases} \quad (9.32)$$

where

$$\pi_{ijk}(c) = \Phi\left(a_k\theta_{ij} - \kappa_{k,c-1}\right) - \Phi\left(a_k\theta_{ij} - \kappa_{k,c}\right).$$

An indicator variable w_{1ij} was defined that equaled one if respondent i at university j was assigned to the RR group and zero otherwise. Other observed explanatory information (gender and ethnicity) was also stored in the matrix $\mathbf{w}$ via effect coding such that the intercept equaled the (unweighted) grand mean. This led to the structural mixed effects model for the latent variable,

$$\theta_{ij} = \gamma^{(2)} + \beta_{0j} + \mathbf{w}_{ij}^t\boldsymbol{\gamma}^{(1)} + e_{ij}, \tag{9.33}$$

where $e_{ij} \sim \mathcal{N}\left(0, \sigma_\theta^2\right)$ and $\beta_{0j} \sim \mathcal{N}\left(0, \tau^2\right)$.

It was assumed that the item response functions were the same across groups, that is, the response probabilities given the alcohol dependence level did not depend on group membership (e.g., DQ group and RR group). The general intercept, $\gamma^{(2)}$, in Equation (9.33) presents the general mean level of the males of different ethnic origin in the DQ group and parameter $\gamma_1^{(1)}$ denotes the contribution of the treatment effect, that is, being questioned via the randomized response technique. A significant positive treatment effect was expected, which means that the randomized response technique induced respondents to answer the items more truthfully.

The model was identified by fixing the mean and variance of the scale of the latent variable to zero and one, respectively. The MCMC algorithm was used to estimate all parameters simultaneously, using 50,000 iterations with a burn-in period of 5,000 iterations. The estimated value of the parameter corresponding to the RR indicator variable is .232 and significantly different from zero while controlling for other population differences. This estimate indicates that the RR group scored significantly higher in comparison with the DQ group on the standardized alcohol dependence scale. It is concluded that the randomized response technique led to an improved willingness of students to answer truthfully.

Fixed Versus Random Effects

The structural relationships between students' alcohol dependence and observed background variables were estimated using the observations from the RR group since those students were more likely to give honest answers. Besides a mixed effects model, an alternative fixed effects model was estimated where interest was focused on students of the four selected colleges/universities that took part in the experiment and not on the underlying population. In a similar way, the clustering of students in colleges/universities was represented using dummy variables. The corresponding structural model contained only fixed effects.

In Table 9.5, the parameter estimates are given for both models. The estimates of the mean and posterior standard deviation of the random effects

(college/university) are given under the label Mixed Effects Model. The estimated variance of the random effects, τ^2, indicates that alcohol dependence of students varies across colleges/universities. However, the corresponding estimated posterior standard deviation is too large to make substantial inferences. The fixed effects estimates show a slightly stronger variation in alcohol dependence across colleges/universities. From the DICs it follows that the fixed effects RIRT model fits the data better than the mixed effects RIRT model.

Furthermore, it follows that male students scored significantly higher in comparison with female students. That is, male students are more likely to experience alcohol-related problems. There are inequalities in reporting alcohol-related problems across ethnic groups, and it turns out that the mean score of black students is much lower than that of other ethnic groups. The mean score of students from Guilford Technical Community College is higher than the other college/university means. The results indicate that gender, ethnicity, and college/university are associated with alcohol-related problems. From Equation (9.25), it followed that both null hypotheses, $\gamma_3^{(1)} = \gamma_4^{(1)} = \gamma_5^{(1)} = \gamma_6^{(1)}$ and $\gamma_7^{(1)} = \gamma_8^{(1)} = \gamma_9^{(1)} = \gamma_{10}^{(1)}$ are rejected with $\alpha = .05$. It can be concluded that there is a main effect for ethnicity and college/university.

Explaining Item Response Variation

The latent category response probabilities for item 10 ("drove under the influence") were investigated for respondents in the RR group. Each respondent i had a probability $\pi_{ik}(c)$ of giving a latent response in category c. Interest was focused on comparing the latent category response probabilities of item 10 across gender and colleges/universities.

The latent response model of the fixed effects RIRT model was expanded for item 10 and the factors gender and college/university,

$$Z_{ik} = a_k\theta_i + \mathbf{w}_i^t\boldsymbol{\gamma} + \epsilon_{ik} \tag{9.34}$$

for $k = 10$, were included, where ϵ_{ik} is standard normally distributed and $\mathbf{w}$ is a design matrix that records the factor information via effect coding. In this case, differences in latent category response probabilities are explained by a random person effect and fixed factors gender and college/university. Note that in the structural model of Equation (9.33) differences in the individual alcohol dependence levels are explained by the fixed factors. In that case, the fixed factor effects are assumed to be the same across items.

In Figure 9.3, the estimated latent category response probabilities per factor level are given from the fixed effects RIRT analysis that includes the specific latent response model for item 10 (Equation (9.34)). Visual inspection shows that males have lower response probabilities for the category "never" and higher response probabilities for the other categories. The estimated category response probabilities also differ visually across levels of factor college/university. It follows that students from Guilford Technical Community College

Table 9.5. CAPS: Parameter estimates of a mixed and fixed effects RIRT model using the RR data.

	Mixed Effects Model			Fixed Effects Model		
	Mean	SD	HPD	Mean	SD	HPD
Fixed Effects						
$\gamma^{(2)}$ (Intercept)	.118	.445	[−.780, .969]	.140	.157	[−.179, .442]
Student variables						
$\gamma_1^{(1)}$ (Female)	−.261	.102	[−.465, −.065]	−.264	.109	[−.483, −.052]
Ethnicity						
$\gamma_2^{(1)}$ (Asian)	−.180	.293	[−.758, .375]	−.198	.324	[−.854, .412]
$\gamma_3^{(1)}$ (White)	.089	.118	[−.141, .321]	.085	.141	[−.195, .357]
$\gamma_4^{(1)}$ (Black)	−.474	.178	[−.837, −.137]	−.465	.186	[−.824, −.095]
$\gamma_5^{(1)}$ (Other)	.543	.247	[.079, 1.027]	.587	.240	[.092, 1.050]
School variables						
University						
$\gamma_6^{(1)}$ (Elon)	.188	.127	[−.073, .417]	.041	.092	[−.144, .217]
$\gamma_7^{(1)}$ (UNCG)	−.148	.137	[−.439, .105]	−.288	.122	[−.529, −.054]
$\gamma_8^{(1)}$ (Wake Forest)	−.014	.159	[−.370, .262]	−.150	.154	[−.452, .149]
$\gamma_9^{(1)}$ (Guilford)	.474	.149	[.143, .722]	.406	.161	[.080, .712]
Random Effects						
Within schools						
σ_θ^2 Residual	.913	.066	[.787, 1.049]	.912	.065	[.789, 1.038]
Between schools						
$\tau^2 Intercept$	.721	.773	[.106, 1.861]			
Information Criteria						
−2 log-likelihood			11768.90			11666.75
DIC			17979.75			17907.82

are more likely to drive under the influence. The fixed factor effects $\boldsymbol{\gamma}$ in Equation (9.34) were tested in a similar way as in Equation (9.25). It was concluded that the factor levels were not significantly different from zero. This means that the factors explained significant structural differences in the individual levels of alcohol dependence (see Table 9.5) but no significant additional item-specific differences in the latent category response probabilities of item 10.

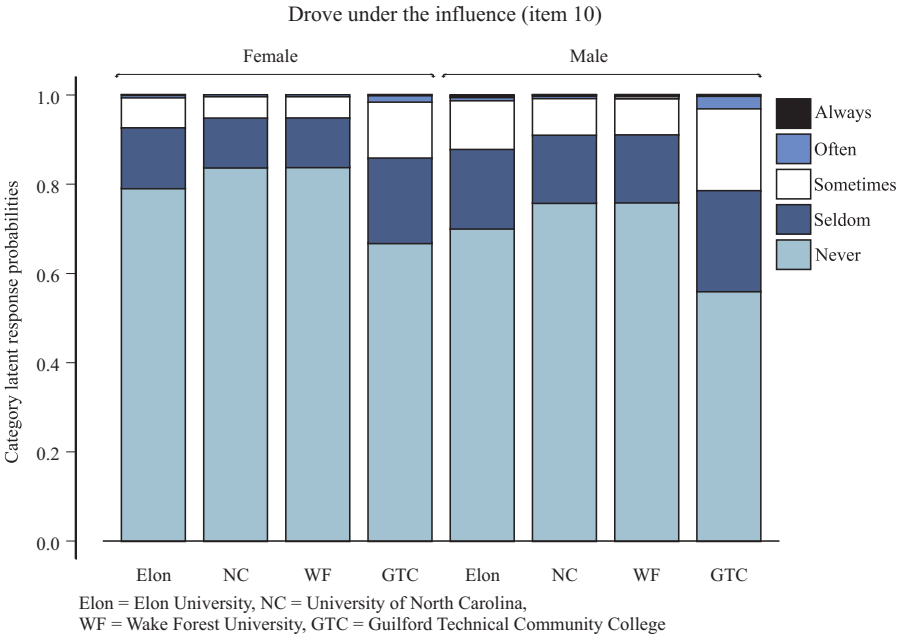

Fig. 9.3. The estimated latent category response probabilities of item 10 across gender and colleges/universities.

9.8 Discussion

An RIRT model is developed for analyzing binary or polytomous hierarchically structured randomized response data. The proposed RIRT model is capable of treating a variety of special problems in a unified framework. The statistical inference of RR data is improved and/or expanded by assuming an individual response model that specifies the relation between the randomized item response data and an (individual) underlying behavior (multivariate response data) or an individual true response probability (univariate response data). Although the (latent) responses are masked and only randomized responses are observed, it is possible to compute (1) individual estimates of a sensitive characteristic, (2) relationships between background variables and the sensitive characteristic, and (3) item characteristics. It is also possible to estimate simultaneously respondent and item effects when observing a mixture of binary/polytomous randomized response and direct-questioning data.

When a researcher believes that distorted results will be obtained because of the sensitive nature of the research, the randomized response methodology reduces the distortions caused by nonresponses or inaccurate responses. On the other hand, the randomized response technique has limitations. The simulation studies showed that the model parameters can be accurately estimated given direct-questioning and/or randomized response observations.

However, the randomized response technique requires larger sample sizes to obtain parameter estimates with the same precision as those obtained via direct questioning. In the unrelated-question design, there is an efficiency loss due to observing responses to the unrelated question. This loss of efficiency can be improved using relevant prior information. There is also additional time needed to administer and explain the procedures to respondents. Besides, tabulating and calculating statistics is more difficult, which increases the cost of using the procedure. In this area, Moshagen (2008) proposed a new administrative procedure for multiple items by introducing a particular distribution scheme for the outcomes of a randomization device. In this design, multiple item outcomes are masked without needing multiple randomization processes.

Respondents in a randomized response study are to be informed about the levels of information that can and cannot be obtained from their answers. Most often, via instructions that are given before starting the survey, it is said that it is impossible to know the true (latent) answers to the questions because the outcome of the randomization device is only visible to the respondent. Furthermore, it is said that it is possible that individual scores on some latent scale can be inferred. It is important to inform respondents that the randomized response design masks the observed item scores. This will improve the willingness of the respondents to cooperate and provide truthful answers. However, from an ethical point of view, it is important to inform respondents about the level of inference that can be made. It would be interesting to investigate the inferences that respondents make based on the instructions. In this light, the method is powerful for respondents who are fully informed about the potential risks of participation in the survey and are still encouraged to provide truthful answers.

9.9 Exercises

9.1. The object is to explore gender differences in responses to item 10 of the cheating study in Section 9.7.1.
(a) Define priors for the parameters of a univariate randomized response model with fixed factor gender, and complete the implementation given in Listing 9.1.

Listing 9.1. WinBUGS code: Univariate randomized item response model.

```
model {
    for (i in 1:N) {
        Y10[i] ~ dbern(pr[i])
        pr[i] <- p1*p[i] + (1-p1)*p2
        p[i] <- phi(theta[i])

        theta[i] ~ dnorm(mu[i],1)
        mu[i]    <- b0 + b1*Male[i]
    }
}
```

(b) Estimate the posterior mean effect of being male, test whether there are significant gender differences, and state your conclusions.
(c) Estimate the true average success probability for the males and females.
(c) Do exercises (b) and (c) for different (incorrect) design probabilities p_1 and p_2, and evaluate the influence of using incorrect randomized response design parameters.

9.2. Consider the randomized response data of the cheating study in Section 9.7.1. Use a one-parameter item response model to analyze the data and identify the model by restricting the mean of the scale.
(a) Implement the (one-parameter) randomized item response model in WinBUGS and estimate the parameters.
(b) Use the MLIRT program (Fox, 2007) to estimate the parameters by indicating that randomized item-responses were observed.
(c) Assume that the (randomized) observations were directly observed and fit the one-parameter item response model. Explain that the items have lower thresholds when assuming directly observed item responses.

9.3. (continuation of Exercise 9.2) Consider the outfit person statistic in Equation (5.40) for detecting noncompliance behavior and misfit of item response patterns.
(a) Implement a posterior predictive check in WinBUGS using the outfit person statistic at the level of the randomized item responses,

$$T\left(\mathbf{Y}_i, \theta_i\right) = K^{-1} \sum_k \left(\frac{Y_{ik} - P_{ik}\left(\theta_i\right)}{\sqrt{P_{ik}\left(\theta_i\right)\left(1 - P_{ik}\left(\theta_i\right)\right)}} \right)^2 ,$$

where $P_{ik}\left(\theta_i\right) = p_1 \Phi\left(\theta_i - b_k\right) + \left(1 - p_1\right) p_2$.
(b) Using the posterior predictive check defined in (a), compute the proportion of respondents with extreme response patterns under the model using a 5% significance level.
(c) Implement a posterior predictive check using the outfit person statistic at the level of the latent item responses,

$$T\left(\tilde{\mathbf{Y}}_i, \theta_i\right) = K^{-1} \sum_k \left(\frac{\tilde{Y}_{ik} - P_{ik}\left(\theta_i\right)}{\sqrt{P_{ik}\left(\theta_i\right)\left(1 - P_{ik}\left(\theta_i\right)\right)}} \right)^2 ,$$

with $P_{ik}\left(\theta_i\right) = \Phi\left(\theta_i - b_k\right)$.
(d) Using the posterior predictive check defined in (c), compute the proportion of respondents with extreme response patterns under the model using a 5% significance level.
(e) Compare the estimated proportions of (b) and (d), and argue that person misfits are likely to be caused by noncompliance behavior.

9.4. Consider observed randomized sum scores, denoted as Y, each consisting of K randomized item responses. Assume a constant individual response

probability or response rate, denoted as θ, for the K binary items. The object is to make posterior inferences about θ.
(a) Show that an observed randomized sum score is binomially distributed with success probability

$$L(\theta) = P(Y_k = 1) = p_1\theta + (1 - p_1)\, p_2$$

when the forced randomized response technique is used.
(b) Assume a beta prior with parameters α and β for the success probability $L(\theta)$. Show that the posterior density of $L(\theta)$ equals

$$p(L(\theta) \mid y) = \frac{\Gamma(\alpha + \beta + K)}{\Gamma(\alpha + y)\,\Gamma(\beta + K - y)} L(\theta)^{\alpha+y-1} (1 - L(\theta))^{\beta+K-y-1},$$

which is a beta density; see Exercise 3.8(b).
(c) Show that the posterior expected response rate equals

$$E(\theta \mid y) = \frac{\alpha + y}{p_1(\alpha + \beta + K)} - (1 - p_1)\, p_2/p_1$$

using $L(\theta)$ as a linear transformation function with an inverse function.
(d) Derive a 95% confidence interval for θ using that

$$\frac{\beta'}{\alpha'} \frac{L(\theta)}{1 - L(\theta)} \sim F_{2\alpha',2\beta'}$$

and that $L(\theta)$ is a posteriori beta distributed with parameters $\alpha' = \alpha + y$ and $\beta' = \beta + K - y$ according to (b).

9.5. The CAPS data (see Sections 5.3.1 and 9.7.2) that are obtained via direct questioning (DQ condition) and the randomized response technique (RR condition) can be analyzed simultaneously.
(a) Use the MLIRT program to fit an RIRT model (Equation (9.32)) with a main effect of RR treatment on the alcohol dependence and invariant CAPS items.
(b) Investigate whether respondents of the RR group score higher on the alcohol dependence scale than respondents of the DQ group.
(c) Assume characteristic differences of the CAPS items across the groups (RR group and DQ group). Assume a standard normal population distribution for alcohol dependence and only one invariant item. Use the MLIRT program to fit the RIRT model (Equation (9.32)).
(d) Compare the results of (a) and (c), and state your conclusion(s).

9.6. (continuation of Exercise 9.5) Investigate differences in alcohol dependence among students using RR treatment and gender as independent explanatory variables. Assume invariant CAPS items across groups.
(a) Estimate the RIRT model with RR treatment and gender (female coded as one) as independent explanatory categorical variables. Interpret the estimated main effects.

(b) Construct an interaction variable by multiplying the RR treatment variable with the gender variable.
(c) Estimate the additional effect of being female in the RR group conditional on the main effects of RR treatment and gender. Investigate whether there is an RR treatment effect on alcohol dependence only for females .

9.7. The observed number of randomized responses per category of the CAPS response data are modeled, where n_{ic} refers to the number of randomized observations in category c of respondent i over K items.
(a) Given a constant response rate of person i per category c, denoted as θ_{ic}, show that the observations are multinomially distributed such that

$$p\left(n_{i1},\ldots,n_{iC} \mid \theta_{i1},\ldots,\theta_{iC}\right) = \frac{K!}{\prod_c n_{ic}!}\prod_c L\left(\theta_{ic}\right)^{n_{ic}}$$

for C response categories, where $L\left(\theta_{ic}\right) = p_1\theta_{ic} + \left(1-p_1\right)p_2(c)$.
(b) Define a conjugated Dirichlet prior for $L\left(\boldsymbol{\theta}_i\right)$,

$$p\left(L\left(\boldsymbol{\theta}_i\right) \mid \boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_c \alpha_c\right)}{\prod_c \Gamma\left(\alpha_c\right)}\prod_c L\left(\boldsymbol{\theta}_i\right)^{\alpha_c - 1},$$

where $\Gamma(.)$ is the gamma function (Equation (2.7)). Show that the posterior of $L\left(\boldsymbol{\theta}_i\right)$ is also a Dirichlet distribution.
(c) Show that the posterior mean of θ_{ic} equals

$$E\left(\theta_{ic} \mid \boldsymbol{\alpha}, \mathbf{n}_i\right) = \frac{\alpha_c + n_{ic}}{p_1\sum_c\left(\alpha_c + n_{ic}\right)} - \left(1-p_1\right)p_2(c)/p_1$$

in a similar way as in Exercise 9.4.

References

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

Afshartous, D. and De Leeuw, J. (2005). Prediction in multilevel models. *Journal of Educational and Behavioral Statistics*, *30*, 109–139.

Aitkin, M. (1997). The calibration of p-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, *7*, 253–261.

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

Albers, W., Does, R. J. M. M., Imbos, T., and Janssen, M. P. E. (1989). A stochastic growth model applied to repeated tests of academic knowledge. *Psychometrika*, *54*, 451–466.

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

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

Albert, J. H. and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, *82*, 747–769.

Anderson, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Hoboken, NJ: Wiley.

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education (2000). *Standards for Educational and Psychological Testing 1999,* 2nd ed. Washington, DC: AERA.

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques,* 2nd ed. New York: Marcel Dekker.

Barnard, J., McCullogh, R. E., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281–1311.

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis,* 2nd ed. London: Arnold.

Bayarri, M. J. and Berger, J. O. (1999). Quantifying surprise in the data. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 6* (pp. 53–82). New York: Oxford University Press.

Bayarri, M. J. and Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, *95*, 1127–1142.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418.

Béguin, A. A. and Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis,* 2nd ed. New York: Springer.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–335.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for linear models. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 5* (pp. 25–44). New York: Oxford University Press.

Berger, J. O. and Selke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, *82*, 112–122.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* New York: Wiley.

Best, N. G., Cowles, M. K., and Vines, K. (2010). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, version 0.5-1 [computer software and manual]. Retrieved from `http://www.mrc-bsu.cam.ac.uk/bugs/classic/coda04/readme.shtml`.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*, 258–276.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: The Massachusetts Institute of Technology.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practices*, *16*, 21–33.

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

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

Böckenholt, U. and van der Heijden, P. G. M. (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika*, *72*, 245–262.

Boscardin, W. J. and Zhang, X. (2004). Modeling the covariance and correlation matrix of repeated measures. In A. Gelman and X.-L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (pp. 215–226). New York: Wiley.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, *143*, 383–430.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

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

Braun, H. I., Jones, D. H., Rubin, D. B., and Thayer, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. *Psychometrika*, *48*, 171–181.

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

Browne, W. J. (2006). MCMC algorithms for constrained variance matrices. *Computational Statistics and Data Analysis*, *50*, 1655–1677.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Casella, G. and Berger, R. L. (2002). *Statistical Inference,* 2nd ed. Pacific Grove, CA: Duxbury Thomson Learning.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, *75*, 651–659.

Chen, M.-H. and Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, *8*, 69–92.

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

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327–335.

Clark, S. J. and Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods*, *3*, 160–168.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

Commenges, D. and Jacqmin, H. (1994). The intraclass correlation coefficient: Distribution-free definition and test. *Biometrics*, *50*, 517–526.

Congdon, P. (2001). *Bayesian Statistical Modelling*. Chichester: Wiley.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, *6*, 101–111.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Jounal of the American Statistical Association*, *91*, 883–904.

Cruyff, M. J. L. F., van den Hout, A., van der Heijden, P. G. M., and Böckenholt, U. (2007). Log-linear randomized-response models taking self-protective response behavior into account. *Sociological Methods and Research*, *36*, 266–282.

Davis, S. F., Grover, C. A., Becker, A. H., and McGregor, L. N. (1992). Academic dishonesty: Prevalence, determinants, techniques, and punishments. *Teaching of Psychology*, *19*, 16–20.

De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

De Jong, M. G., Pieters, R., and Fox, J.-P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, *47*, 14–27.

De Jong, M. G. and Steenkamp, J. B. E. M. (2009). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika*, *(online)*.

De Jong, M. G., Steenkamp, J. B. E. M., and Fox, J.-P. (2007). Relaxing cross-national measurement invariance using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278.

De Jong, M. G., Steenkamp, J. B. E. M., Fox, J.-P., and Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*, 104–115.

De Leeuw, J. and Kreft, I. G. G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, *11*, 57–85.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: Wiley.

DeIorio, M. and Robert, C. P. (2002). Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, *64*, 629–630.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*, 341–353.

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

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, *56*, 363–375.

Donders, F. C. (1868). Over de snelheid van psychische processen [On the speed of mental processes]. Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868–1869, Tweede reeks, II, 92–120.

Doolaard, S. (1999). *Schools in Change or Schools in Chains?* PhD dissertation, University of Twente.

Edgell, S. E., Himmelfarb, S., and Duchan, K. L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods and Research*, *11*, 89–100.

Edwards, A. W. F. (1963). The measure of association in a 2x2 table. *Journal of the Royal Statistical Society, Series A*, *126*, 109–114.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalization. *Journal of the American Statistical Association*, *70*, 311–319.

Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Emons, W. H. M., Sijtsma, K., and Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, *10*, 101–119.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models,* 2nd ed. New York: Springer.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental/state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

Fox, J. A. and Tracy, P. E. (1986). *Randomized Response: A Method for Sensitive Surveys.* Beverly Hills, CA: Sage.

Fox, J.-P. (2001). *Multilevel IRT: A Bayesian Perspective on Estimating Parameters and Testing Statistical Hypotheses.* PhD dissertation, University of Twente, Faculty of Behavioural Sciences.

Fox, J.-P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology*, *56*, 65–81.

Fox, J.-P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, *15*, 261–280.

Fox, J.-P. (2005a). Multilevel IRT model assessment. In A. van der Ark, M. A. Croon, and K. Sijtsma (Eds.), *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences* (pp. 227–252). Mahwah, NJ: Lawrence Erlbaum.

Fox, J.-P. (2005b). Multilevel IRT using dichotomous and polytomous items. *British Journal of Mathematical and Statistical Psychology*, *58*, 145–172.

Fox, J.-P. (2005c). Randomized item response theory models. *Journal of Educational and Behavioral Statistics*, *30*, 189–212.

Fox, J.-P. (2007). Multilevel IRT modeling in practice. *Journal of Statistical Software*, *20*, Issue 5.

Fox, J.-P. and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

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

Fox, J.-P., Klein Entink, R. E., and van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*, Issue 7.

Fox, J.-P. and Meijer, R. R. (2008). Using item response theory to obtain individual information from randomized response data: An application using cheating data. *Journal of Applied Psychological Measurement*, *32*, 595–610.

Fox, J.-P. and Wyrick, C. (2008). A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics*, *33*, 389–415.

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

Fuller, W. A. (1991). Regression estimation in the presence of measurement error. In P. P. Biemer, R. M. Groves, L. E. Lyberg, and N. A. Mathiowetz (Eds.), *Measurement Errors in Surveys* (pp. 617–635). New York: Wiley.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*, 320–328.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, *56*, 501–514.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling based methods (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 147–167). Oxford: Oxford University Press.

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

Gelman, A. (1995). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 131–143). London: Chapman and Hall.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*, 445–450.

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

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, A. and King, D. G. (1990). Estimating the electoral consequences of legislative redirecting. *Journal of the American Statistical Association*, *85*, 274–282.

Gelman, A., Meng, X.-L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 169–193). Oxford: Oxford University Press.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Hoboken, NJ: Wiley.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, *4*, 473–483.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics and Probability Letters*, *23*, 165–170.

Ghosh, M., Ghosh, A., Chen, M.-H., and Agresti, A. (2000). Noninformative priors for one-parameter item response models. *Journal of Statistical Planning and Inference*, *88*, 99–115.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Glas, C. A. W. and Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*, 217–233.

Glas, C. A. W. and van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247–261.

Goldstein, H. (2003). *Multilevel Statistical Models,* 3rd ed. London: Hodder Arnold.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education*, *11*, 319–330.

Goldstein, H., Bonnet, G., and Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*, 252–286.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., and Healy, M. (1998). *A User's Guide to MLwiN*. London: Multilevel Models Project, Institute of Education.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.

Greenberg, B. G., Abul-Ela, A., Simmons, W. R., and Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *The American Statistician*, *64*, 520–539.

Gulliksen, H. O. (1950). *Theory of Mental Tests*. New York: Wiley.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B*, *29*, 83–100.

Hall, D. B. and Clutter, M. (2004). Multivariate multilevel nonlinear mixed effects models for timber yield predictions. *Biometrics*, *60*, 16–24.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. London: Sage.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and related problems. *Journal of the American Statistical Association*, *72*, 320–340.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Hedeker, D. R. (1999). MIXNO: A computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software*, *4*, 1–92.

Hedeker, D. R. and Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 157–176.

Hedeker, D. R. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Hoboken, NJ: Wiley.

Heidelberg, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144.

Higdon, D. M. (1998). Auxiliary variable methods for Markov Chain Monte Carlo with applications. *Journal of the American Statistical Society*, *93*, 585–595.

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Society*, *60*, 806–825.

Hoijtink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A. J. van Duijn, and T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 109–129). New York: Springer.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.

Holland, P. W. and Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Janssen, R., Schepers, J., and Peres, D. (2004). Models with item and item group predictors. In P. de Boeck and M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 189–212). New York: Springer.

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

Jeffreys, H. J. (1961). *Theory of Probability*. New York: Oxford University Press.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. New York: Springer.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics*, *48*, 631–644.

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

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, *25*, 163–176.

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. J., and Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika*, *59*, 405–421.

Klein Entink, R. H. (2009). *Statistical Models for Responses and Response Times*. PhD dissertation, University of Twente, Faculty of Behavioural Sciences.

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., and Fox, J.-P. (2009b). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*, 54–75.

Kuha, J. (1997). Estimation by data augmentation in regression models with continuous and discrete covariates measured with error. *Statistics in Medicine*, *16*, 189–201.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. New York: Houghton Mifflin.

Lee, P. M. (2004). *Bayesian Statistics: An Introduction,* 3rd ed. New York: Wiley.

Lee, S.-Y. and Zhu, H.-T. (2000). Statistical analysis of non-linear equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *53*, 209–232.

Lehmann, E. L. and Casella, G. (2003). *Theory of Point Estimation,* 2nd ed. New York: Springer.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., and Maas, C. J. M. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods and Research*, *33*, 319–348.

Leonard, T. and Hsu, J. S. J. (1999). *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge: Cambridge University Press.

Levy, R. (2006). *Posterior Predictive Model Checking for Multidimensionality in Item Response Theory and Bayesian Networks*. PhD dissertation, University of Maryland.

Lindley, D. V. (1965). *An Introduction to Probability and Statistics from a Bayesian Viewpoint (Parts 1 and 2).* Cambridge: Cambridge University Press.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, *34*, 1–41.

Little, R. J. A. and Rubin, D. A. (2002). *Statistical Analysis with Missing Data,* 2nd ed. Hoboken, NJ: Wiley.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, *81*, 27–40.

Liu, L. C. and Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, *62*, 261–268.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*, 817–827.

Longford, N. T. (1993). *Random Coefficient Models.* New York: Oxford University Press.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157–162.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization.* New York: Oxford University Press.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility [computer software]. *Statistics and Computing*, *10*, 325–337.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, *48*, 188–190.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307–330.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445–469.

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

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

Masters, G. N. and Wright, B. D. (1997). The partial credit model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 101–121). New York: Springer.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,* 2nd ed. New York: Chapman and Hall.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, *99*, 173–193.

McDonald, R. P. (1967). *Nonlinear Factor Analysis* (Psychometric Society Monograph No. 15). Richmond, VA: William Byrd Press.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*, 72–87.

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

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Mellenbergh, G. J. (1994a). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

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

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*, 1142–1160.

Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, *86*, 301–320.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Meredith, W. and Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*, 289–311.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087–1092.

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

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

Mislevy, R. J. and Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*, 661–679.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fisher and I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications* (pp. 3–14). New York: Springer.

Molenaar, I. W. and Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 75–106.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*, 47–55.

Moshagen, M. (2008). *Multinomial Randomized Response Models*. Phd dissertation, Heinrich-Heine-Universität Dusseldörf, Mathematisch Naturwissenschaftlichen Fakultät.

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

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*, 351–363.

Muraki, E. and Bock, R. D. (1997). *PARSCALE: IRT Based Test Scoring and Item Analysis for Graded Items and Rating Scales* [computer software]. Chicago, IL: Scientific Software International.

Muthén, B. O. (1992). Latent variable modeling in epidemiology. *Alcohol Health and Research World*, *16*, 286–292.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides and R. E. Schumacker (Eds.), *New Developments and Techniques in Structural Equation Modeling* (pp. 1–33). New York: Lawrence Erlbaum Associates.

Muthén, B. O. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*, 463–469.

Muthén, L. K. and Muthén, B. O. (1998). *Mplus: The Comprehensive Modeling Program for Applied Researchers* [computer software]. Los Angeles, CA: Muthén and Muthén.

Nandram, B. and Chen, M.-H. (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation*, *54*, 129–144.

Neal, R. M. (1997). Markov Chain Monte Carlo methods based on 'slicing' the density function. Technical Report No. 9722, University of Toronto, Department of Statistics.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, *56*, 3–48.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16*, 1–32.

Novick, M. R., Lewis, C., and Jackson, P. H. (1973). The estimation of proportions in m groups. *Psychometrika*, *38*, 19–46.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Hoboken, NJ: Wiley.

OECD (Organisation for Economic Co-operation and Development) (2004). *Learning From Tomorrow's World. First Results from PISA 2003*. Paris: OECD.

O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika*, *63*, 329–333.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society, Series B*, *57*, 99–138.

O'Hare, T. M. (1997). Measuring problem drinkers in first time offenders: Development and validation of the college alcohol problem scale (CAPS). *Journal of Substance Abuse Treatment*, *14*, 383–387.

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

Orlando, M. and Thissen, D. (2003). Further investigation of the performance of $s - x^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298.

Ostini, R. and Nering, M. L. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–356.

Patz, R. J. and 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*, 342–366.

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

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*, 341–384.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics*. Hoboken, NJ: Wiley.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabe-Hesketh, S. and Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics*, *57*, 1256–1264.

Raftery, A. E. (1995). Bayesian model selection in social research. *Social Methodology*, *25*, 111–163.

Raftery, A. L. and Lewis, S. (1992). Comment: One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, *7*, 493–497.

Rasch, G. (1960). *Probabilistic Models for some Intelligence Tests and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods,* 2nd ed. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., and Congdon, R. T. (2000). *HLM 5: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W. and Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, *29*, 1–41.

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, *77*, 190–195.

Reinsel, G. (1983). Some results on multivariate autoregressive index models. *Biometrika*, *70*, 145–156.

Richman, W. L., Kiesler, S., Weisband, S., and Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*, 754–775.

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

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

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

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer.

Roberts, G. O. (1995). Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 45–57). London: Chapman and Hall.

Roberts, G. O. and Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and Their Applications*, *80*, 211–229; Correction *91*, 337–338.

Robins, J. M., van der Vaart, A. W., and Ventura, V. (2000). Asymptotic distribution of P values in composite null models: Rejoinder. *Journal of the American Statistical Association*, *95*, 1171–1172.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*, 349–359.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, *90*, 558–566; Correction *91*, 1136.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 187–208). New York: Springer.

Rossi, P. E., Allenby, G. M., and McCullogh, R. E. (2005). *Bayesian Statistics and Marketing*. Chichester: Wiley.

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

Rost, J. and von Davier, M. (1995). Mixture distribution Rasch models. In G. Fischer and I. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications* (pp. 257–268). New York: Springer.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151–1172.

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

Rupp, A. A., Dey, D. K., and Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling*, *11*, 424–451.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, *72*, 217–232.

Samejima, F. (1997). The graded response model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). New York: Springer.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437–457.

Scheerens, J. (1992). *Effective Schooling: Research, Theory and Practice*. London: Cassell.

Scheerens, J., Glas, C. A. W., and Thomas, S. M. (2003). *Educational Evaluation, Assessment, and Monitoring*. Lisse: Swets and Zeitlinger.

Scheers, N. J. and Dayton, C. (1988). Covariate randomized response model. *Journal of the American Statistical Association*, *83*, 969–974.

Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*, 18–38.

Schnipke, D. L. and Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method for measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Schnipke, D. L. and Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, and W. C. Ward (Eds.), *Computer-Based Testing: Building the Foundation for Future Assessments* (pp. 237–266). Mahwah, NJ: Lawrence Erlbaum.

Schulz-Larsen, K., Kreiner, S., and Lomholt, R. K. (2007a). Mini-mental status examination: A short form of MMSE was as accurate in predicting dementia. *Journal of Clinical Epidemiology*, *60*, 260–267.

Schulz-Larsen, K., Kreiner, S., and Lomholt, R. K. (2007b). Mini-mental status examination: Mixed Rasch model item analysis derived two different

cognitive dimensions of the MMSE. *Journal of Clinical Epidemiology*, *60*, 268–279.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.

Shalabi, F. (2002). *Effective Schooling in the West Bank*. PhD dissertation, University of Twente.

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

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375–394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429–449.

Sinharay, S., Johnson, M. S., and Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321.

Sinharay, S. and Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*, 196–201.

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

Smith, B. (2010). BOA: Bayesian Output Analysis Program, version 1.1.5 [computer software and manual]. Retrieved from `http://www.public-health.uiowa.edu/boa/`.

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

Snijders, T. A. B. and Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Soares, T. M., Gonçalves, F. B., and Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*, 348–377.

Song, X.-Y. and Lee, S.-Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical and Statistical Psychology*, *54*, 237–263.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583–639.

Steenkamp, J. B. E. M. and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Stern, H. S. (2000). Asymptotic distribution of P values in composite null models: Comment. *Journal of the American Statistical Association*, *95*, 1157–1159.

Stone, C. A. and Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement*, *60*, 974–991.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., and Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354.

Swaminathan, H. and Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, *7*, 175–192.

Swaminathan, H. and Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349–364.

Swaminathan, H. and Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*, 589–601.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.

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

Theil, H. (1963). On the use of incomplete prior information in regression analysis. *Journal of the American Statistical Association*, *58*, 401–414.

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

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing* (pp. 179–203). New York: Academic Press.

Thissen, D. (1991). *MULTILOG: Multiple Category Item Analysis and Test Scoring Using Item Response Theory* [computer software]. Chicago, IL: Scientific Software International.

Thissen, D. and Wainer, H. (2001). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum.

Tiao, G. C. and Tan, W. Y. (1965). Bayesian analysis of random-effects models in the analysis of variance. I: Posterior distribution of variance components. *Biometrika*, *52*, 37–53.

TIBCO Software (2009). *TIBCO Spotfire S+ 8.1: Programmer's Guide and Computer Program* [computer software]. TIBCO Software Inc.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, *22*, 1701–1762.

Tombaugh, T. N. (1992). The mini-mental state examination: A comprehensive review. *The Journal of the American Geriatrics Society*, *40*, 922–935.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York: Wiley.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859–883.

Tracy, P. E. and Fox, J. A. (1981). The validity of the randomized response for sensitive measurements. *American Sociological Review*, *46*, 187–200.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics*, *9*, 263–276.

Tsutakawa, R. K. and Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, *51*, 251–267.

Tsutakawa, R. K. and Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics*, *13*, 117–130.

Tucker, L. R. (1952). A level of proficiency scale for a unidimensional skill. *American Psychologist*, *7*, 408 (Abstract).

Tutz, G. and Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics and Data Analysis*, *22*, 537–557.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*, 351–370.

van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, *70*, 359–376.

van den Hout, A. and Klugkist, I. (2009). Accounting for non-compliance in the analysis of randomized response data. *Australian and New Zealand Journal of Statistics*, *51*, 353–372.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

van der Linden, W. J. and Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

van der Maas, H. L. J. and Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, *118*, 29–60.

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

Vandenberg, R. J. and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

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

Verhelst, N. D., Glas, C. A. W., and Verstralen, H. H. F. M. (1995). *OPLM: One Parameter Logistic Model* [computer software]. Arnhem: Cito.

Verhelst, N. D. and Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. van Duijn, and T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 89–106). New York: Springer.

Verhelst, N. D., Verstralen, H. H. F. M., and Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 169–185). New York: Springer.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*, 33–51.

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. Cambridge: Cambridge University Press.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69.

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement*, *14*, 219–225.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; A Gibbs sampling approach. *Journal of the American Statistical Association*, *86*, 79–86.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Zellner, A. (1997). *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*. Cheltenham: Edward Elgar.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. (1996). *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items* [computer software]. Chicago, IL: Scientific Software International.

Index