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

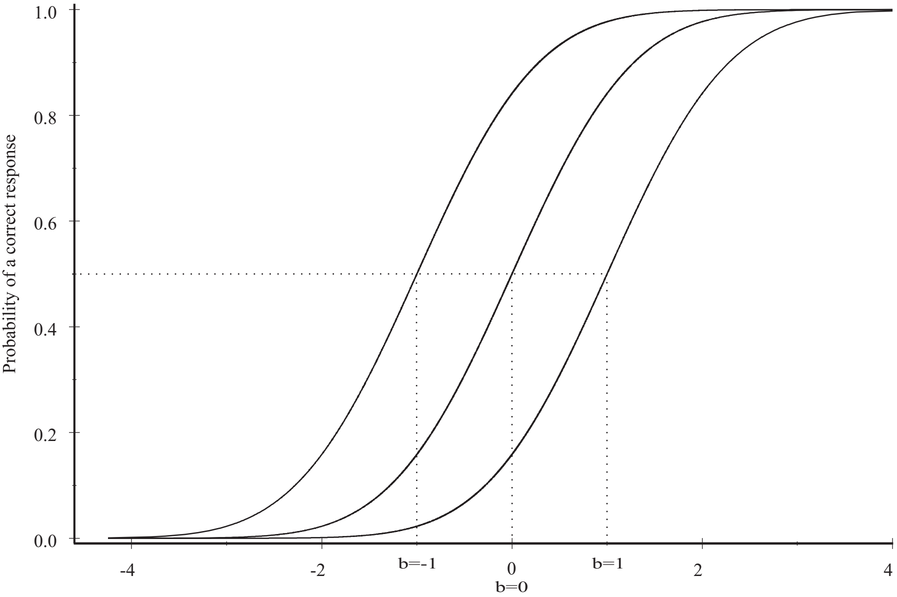

Fig. 1.1. Item characteristic curves of the one-parameter IRT model corresponding to three difficulty levels.

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

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

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

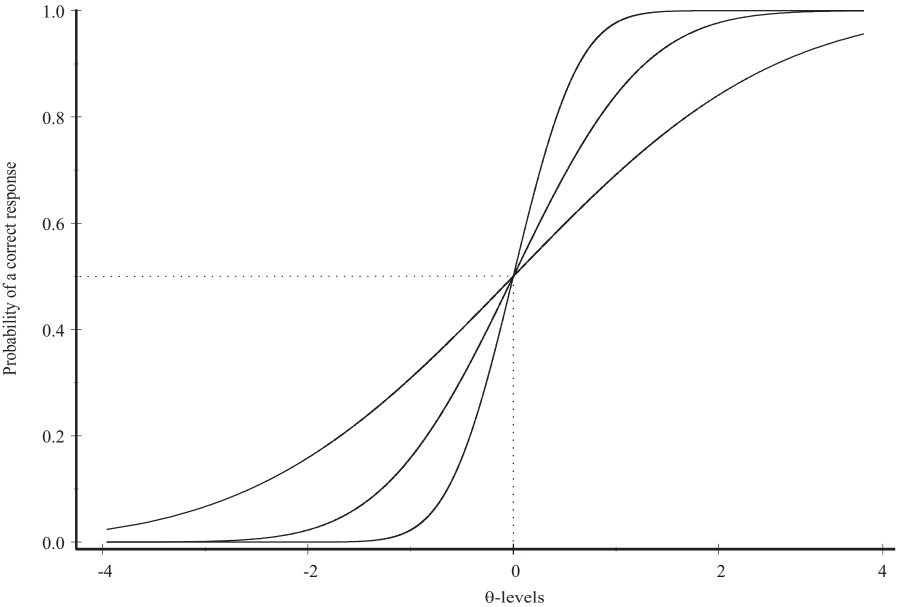

Fig. 1.2. Item characteristic curves of the two-parameter IRT model corresponding to three discrimination levels and an equal level of difficulty.

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,

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

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 $a_k\left(\theta_i - b_k^*\right)$. 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,

$$\begin{align} P(Y_{ik} = 1 \mid \theta_i, a_k, b_k, c_k) &= c_k + (1 - c_k)\Phi(a_k\theta_i - b_k) \tag{1.5} \\ &= \Phi(a_k\theta_i - b_k) + c_k\left(1 - \Phi(a_k\theta_i - b_k)\right), \tag{1.6} \end{align}$$

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

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

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.

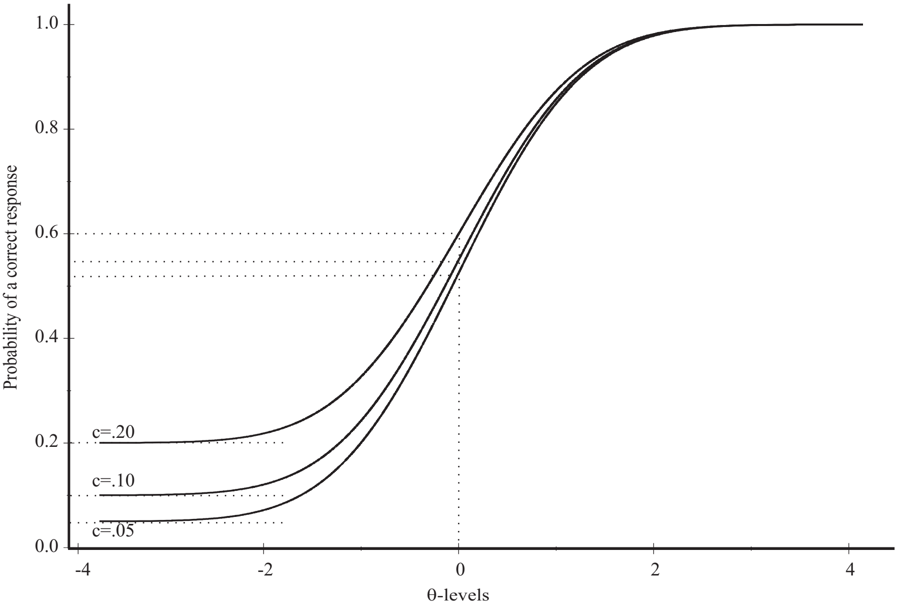

Fig. 1.3. Item characteristic curves of the three-parameter IRT model corresponding to three different levels of guessing and an equal level of discrimination and difficulty.

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

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

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

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

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

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}) &= p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{y}) && (1.9)\\ &\propto p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}), && (1.10)\end{aligned}$$

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.

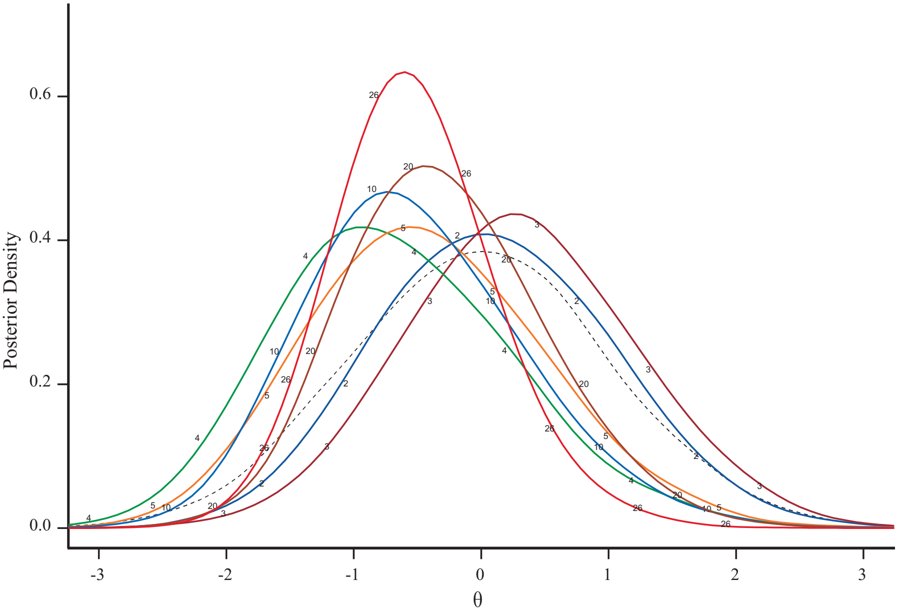

Fig. 1.4. Updated posterior densities of a person's cognitive impairment for 2–26 item observations.

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

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

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.

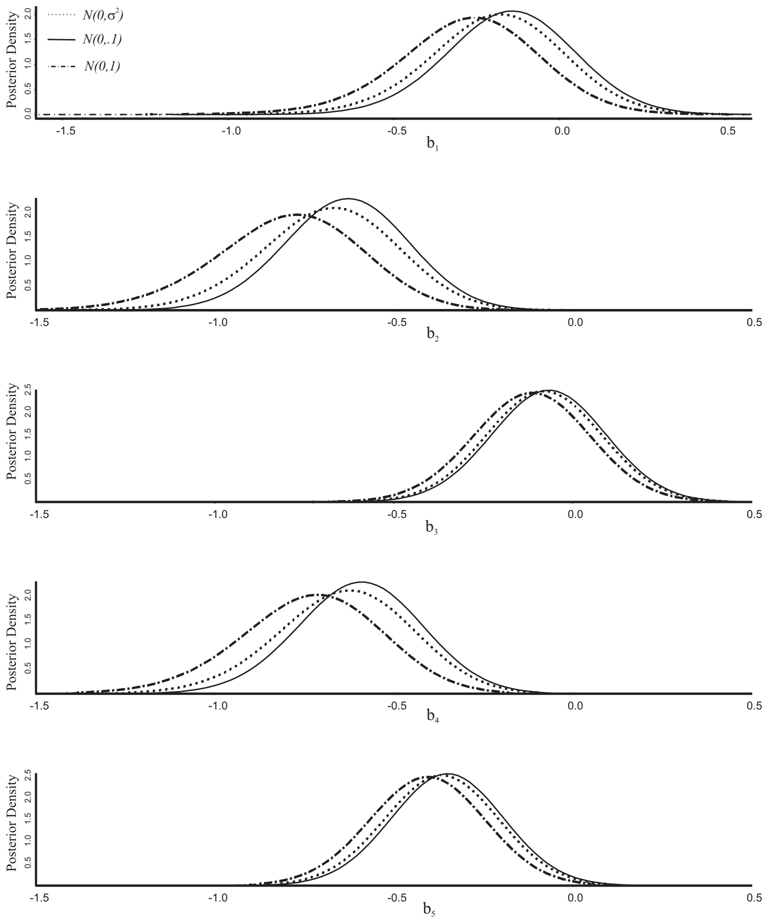

Fig. 1.5. Estimated posterior densities of the difficulty parameters for different prior choices.

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.