Chapter 12

Multilevel IRT Model Assessment

Jean-Paul Fox
University of Twente

12.1 Introduction

Modelling complex cognitive and psychological outcomes in, for example, educational assessment led to the development of generalized item response theory (IRT) models. A class of models was developed to solve practical and challenging educational problems by generalizing the basic IRT models. An IRT model can be used to define a relation between observed categorical responses and an underlying latent trait, such as, ability or attitude. Subsequently, the latent trait variable can be seen as the outcome in a regression analysis. That is, a regression model defines the relation between the latent trait and the set of predictors. The combination of both models, a regression model imposed on the ability parameter in an IRT model, can be viewed as an extension to the class of IRT models.

Verhelst and Eggen (1989), and Zwinderman (1991, 1997) considered the combination of an IRT model with a structural linear regression model. Zwinderman showed that the correlation between the latent trait and other variables can be estimated directly without estimating the subject parameters. A straightforward extension of this model consists of a structural multilevel model imposed on the latent trait variable. Adams, Wilson and

Wu (1997) and Raudenbush and Sampson (1999) discussed a multilevel model that can be seen as a Rasch model embedded within a hierarchical structure, where the first level of the multilevel model describes the relation between the observed item scores and the ability parameters. This multilevel model can be estimated in HLM 5 (Raudenbush, Bryk, Cheong, & Congdon, 2000). A multilevel formulation of the Rasch model that can be estimated using the HLM software was developed by Kamata (2001). Maier (2001), defined a hierarchical Rasch model, that is, the person parameters in the Rasch model are modelled with a one-way ANOVA with random effects. Fox (in press) and Fox and Glas (2001) extended the two-parameter normal ogive model and the graded response model by imposing a multilevel model on the ability parameters with covariates on both levels. This multilevel IRT model describes the link between dichotomous or polychotomous response data and a latent dependent variable as the outcome in a structural multilevel model. This extension allows to model relationships between observed and latent variables on different levels using dichotomous and polytomous IRT models that relate the test performances to the latent variables. That is, relationships between abilities of students underlying the test and other observed variables of some individual or group characteristics can be analyzed taking into account the errors of measurement using dichotomous or polytomous indicators.

Verhelst and Eggen (1989) proclaimed a strict distinction between the estimation of the parameters of the measurement model and the structural model. One should first calibrate the measurement model before estimating the structural model parameters. This way it is possible to distinguish possible model violations in the measurement model and the structural model. Alternatively, a two-stage estimation procedure can cause biased parameter estimates and underestimation of some standard errors due to the fact that some parameters are held fixed at values estimated from the data, depending on the available calibration data. Furthermore, in educational testing the response data often have a hierarchical structure and the measurement model ignores this effect of the clustering of the respondents. In Fox (in press) and Fox and Glas (2001) a procedure was developed for estimating simultaneously all model parameters. This Bayesian method (Markov chain Monte Carlo, MCMC) handles all sources of uncertainty in the estimation of the model parameters.

The goal of this chapter is to develop methods to assess the plausibility of the model or some of the assumptions under the preferred Bayesian estimation method. The MCMC estimation procedure is time-consuming and it is, therefore, preferable to compute certain fit statistics during the estimation of the parameters or based on the MCMC output. In this chapter, methods are proposed for checking the fit of multilevel IRT models, which are simply byproducts of the MCMC procedure for estimating the model parameters. Samples of the posterior distributions are obtained from the MCMC estimation method. These samples can be directly used to estimate the model parameters and the posterior standard deviations, but they can also be used to test certain model assumptions or, in general, the fit of the model.

Before using a model it is necessary to investigate the adequacy and plausibility of the model. Such investigations include a residual analysis. The classical or Bayesian residuals are based on the difference between observed and predictive data under the model, but they are difficult to define and interpret due to the discrete nature of the response variable. Therefore, another approach to a residual analysis is proposed. The dichotomous or polytomous outcomes on the item-level are supposed to have an underlying normal regression structure on latent continuous data. This assumption results in an analysis of Bayesian latent residuals, based on the difference between the latent continuous and predictive data under the model. We show that the Bayesian latent residuals have continuous-valued posterior distributions and are easily estimated with the Gibbs sampler (Albert, 1992; Albert & Chib, 1995). Furthermore, Bayesian residuals have different posterior variances but the Bayesian latent residuals are identically distributed.

Different statistics to check the model fit or certain assumptions are proposed, all based on posterior distributions. First, the posterior distributions of the random errors are used to detect outliers in the multilevel IRT model. An outlier is defined as an observation with a large random error, generated by the model under consideration (Chaloner & Brant, 1988). The posterior distributions can be used to calculate the posterior probability that an observation is an outlier. These posterior probabilities of an observation being an outlier are calculated with the Gibbs sampler. Other Bayesian approaches to outlier detection can be found in, for example, Box and Tiao (1973) and Zellner (1971).

Second, hypotheses can be tested using interval estimation. The smallest interval containing 95% of the probability under the posterior is called the 95% highest posterior density (HPD) interval. According to the usual form of a hypothesis that a parameter value or a function of parameter values is zero, the HPD interval can be used to test if the value differs significantly from zero (Box & Tiao, 1973). Here, this concept is used to check heteroscedasticity at the individual level (Level 1), that is, to check whether grouped Level 1 residuals have the same posterior distribution. The parametric forms of the marginal posterior distributions are unknown, but samples of the distributions are available through the Gibbs sampler. These samples are used to check the homoscedasticity assumption at Level

1.

In section 12.3, after the introduction of the multilevel IRT model, a Bayesian residual analysis is described. Next, a method to detect outliers by examining the posterior distribution of the residuals using MCMC is discussed. Then, tests based on highest posterior density intervals are described to test the homoscedasticity at Level 1. Examples of the procedures are given by analyzing a real data set. Finally, the last section contains a discussion and suggestions for further research.

12.2 Multilevel IRT Model

Suppose that the categorical outcome Y_{ijk} represents the item response of person i $(i = 1, \ldots, n_j)$, in group j $(j = 1, \ldots, J)$, on item k $(k = 1, \ldots, K)$. Let θ_{ij} denote the latent abilities of the persons responding to the K items. The latent ability parameters are collected in the vector $\boldsymbol{\theta}$. In the present chapter, the multilevel IRT model consists of two components, an IRT model for $p\left(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}\right)$, where $\mathbf{a}$ and $\mathbf{b}$ are the item parameters, and a model $p\left(\boldsymbol{\theta}|\boldsymbol{\beta}, \mathbf{X}, \mathbf{W}\right)$ for the relation between the latent abilities and the background variables. Explanatory variables at Level 1 containing information regarding the persons are stored in the matrix $\mathbf{X}$. In the same way, matrix $\mathbf{W}$ contains information regarding the groups at Level 2. Parameters $\boldsymbol{\beta}$ are the regression coefficients from the regression of $\boldsymbol{\theta}$ on $\mathbf{X}$. The regression coefficients may vary across groups using the explanatory variables stored in $\mathbf{W}$. The first part, $p\left(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}\right)$, is specified by a normal ogive model in case of binary response data. That is, the probability of a student, with latent ability θ, dropping for convenience reasons the subscript ij, corresponding correct to an item k is given by

$$P\left(Y_k = 1 \mid \theta, a_k, b_k\right) = \Phi\left(a_k\theta - b_k\right),$$

where $\Phi\left(.\right)$ denotes the standard normal cumulative distribution function, and a_k and b_k are the discrimination and difficulty parameter of item k. The relation between the underlying latent ability and the dichotomous outcomes can also be explained as follows. Assume a latent independent random normally distributed variable Z_k with mean $a_k\theta - b_k$ and variance 1. In addition, the response Y_k is the indicator of Z_k being positive. Thus, a correct response on item k is obtained if a positive value is drawn from this normal distribution with mean $a_k\theta - b_k$ and variance 1.

In the case of polytomous scored items, the polytomous response, Y_k, can be viewed as an indicator of Z_k falling into one of the response categories. Or, the reverse, classifying the latent variable Z_k into more than two categories is done by the cutoff or threshold parameters κ. In this case, the latent variable Z_k is defined as

$$Z_k = a_k\theta + \varepsilon_k \tag{12.1}$$

where ε_k is assumed to have a standard normal distribution. When the value of the latent variable Z_k falls between the thresholds κ_{kc-1} and κ_{kc}, the observed response on item k is classified into category c. The threshold parameters are unknown and they are estimated using the observed data. The ordering of the response categories is displayed as follows

$$-\infty < \kappa_{k1} \leq \kappa_{k2} \leq \ldots \leq \kappa_{kC_k},$$

where there are C_k categories. The number of categories may differ per item. Here, for notational convenience, $\kappa_0 = -\infty$ and the upper cutoff parameter $\kappa_{kC_k} = \infty$ for every item k $(k = 1, \ldots, K)$. The probability that an individual, given some underlying latent ability, θ, obtains a grade c, or gives a response falling into category c on item k is defined by

$$P\left(Y_k = c \mid \theta, a_k, \kappa_k\right) = \Phi\left(\kappa_{kc} - a_k\theta\right) - \Phi\left(\kappa_{kc-1} - a_k\theta\right), \tag{12.2}$$

where $\Phi\left(.\right)$ denotes the standard normal cumulative distribution function. This IRT model for polytomously scored items, called the graded response model or the ordinal probit model, has been used by several researchers, among others, Albert and Chib (1993), Johnson and Albert (1999), Muraki and Carlson (1995), and Samejima (1969).

The second component of the model, $p\left(\boldsymbol{\theta}|\boldsymbol{\beta}, \mathbf{X}, \mathbf{W}\right)$, specifies the relation between the background information and the latent variables via a multilevel model, in specific

$$\begin{aligned} \theta_{ij} &= \boldsymbol{\beta}_j' \mathbf{X}_{ij} + e_{ij} \\ \boldsymbol{\beta}_j &= \mathbf{W}_j \boldsymbol{\gamma} + \mathbf{u}_j \end{aligned} \tag{12.3}$$

where $e_{ij} \sim N\left(0, \sigma^2\right)$ and $\mathbf{u}_j \sim N\left(0, \mathbf{T}\right)$. The apostrophe defines the transpose of the vector. Parameters $\boldsymbol{\gamma}$, the so-called fixed effects, are the regression coefficients from the regression of $\boldsymbol{\beta}$ on $\mathbf{W}$. The location and scale indeterminacies can be solved by forcing the intercept in (12.3) to 0 and the variance of the latent dependent variable to 1. It is also possible to put identification restrictions on the item parameters. A Markov chain Monte Carlo method can be used to estimate the parameters of interest (Fox, in press, Fox & Glas, 2001). Computing the posterior distributions of the parameters involves high dimensional integrals but these can be carried out by Gibbs sampling (Gelfand, Hills, Racine-Poon, & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 1995). Within this Bayesian approach, all parameters are estimated simultaneously.

12.3 Bayesian Residual Analysis

The regression residuals can be used to check assumptions such as normality, conditional independence of observations and homoscedasticity of variance. There is often an interest in the magnitudes of the errors that actually occurred. The realized errors are not observed. They need to be estimated from the data together with the uncertainties associated with these estimates. In this chapter, realized residuals are viewed as random parameters with unknown values. Posterior distributions for realized errors need to be calculated and can be used to make posterior probability statements about the values of the realized errors (see, e.g., Box & Tiao, 1973; Zellner, 1971). 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. Goodness-of-fit statistics are used to summarize the model adequacy. Besides checking several model assumptions, attention is focused on examining the adequacy of the model in predicting individual data points.

In the binary case, the residuals are defined as $r_{ijk} = y_{ijk} - \Phi\left(a_k\theta_{ij} - b_k\right)$. In the classical residual analysis, residuals are usually transformed such that they approximately follow a normal distribution. The three most common normalizing transformations lead to Pearson, deviance, and adjusted deviance residuals. But in case of Bernoulli observations such transformations result in poor approximations of the distributions of the Pearson, deviance and adjusted deviance residuals by the Gaussian distribution. A fully Bayesian residual analysis does not suffer from this problem. In the Bayesian residual analysis attention is focused on the posterior distribution of each residual. Bayesian residuals have continuous-valued posterior distributions which can also be used to detect outliers.

The Gibbs sampler can be used to estimate the posterior distribution of the residuals. Denote an MCMC sample from the posterior distribution of the parameters (θ_{ij}, a_k, b_k) by $\left(\theta_{ij}^{(m)}, a_k^{(m)}, b_k^{(m)}\right)$, $m = 1, \ldots, M$. It follows that sampled values from the residual posterior distribution corresponding to observation ijk are defined by

$$r_{ijk}^{(m)} = y_{ijk} - \Phi\left(a_k^{(m)}\theta_{ij}^{(m)} - b_k^{(m)}\right),\ m = 1, \ldots, M. \tag{12.4}$$

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

12.3.1 Computation of Bayesian Latent Residuals

The Bayesian latent residuals are based on the introduction of the latent variable $\mathbf{Z}$. For binary response data, this latent continuous score is defined as Z_{ijk}, where $Z_{ijk} > 0$ if $Y_{ijk} = 1$ and $Z_{ijk} \leq 0$ if $Y_{ijk} = 0$. Then, the Bayesian latent residuals corresponding to observations Y_{ijk} are defined as

$$\varepsilon_{ijk} = Z_{ijk} - a_k\theta_{ij} + b_k. \tag{12.5}$$

From the definition of the augmented data it follows that given a_k, b_k and θ_{ij}, the Bayesian latent residuals ε_{ijk} are standard normally distributed. For more detailed information regarding Bayesian latent residuals, in case of binary data, see Albert and Chib (1995) and Johnson and Albert (1999). According to Equation 12.1, the Bayesian latent residuals, in case of polytomous response data, are defined as

$$\varepsilon_{ijk} = Z_{ijk} - a_k\theta_{ij}. \tag{12.6}$$

Both Bayesian latent residuals (Equations 12.5 and 12.6) are easily estimated as a byproduct of the Gibbs sampler. That is, MCMC samples from Z_{ijk}, a_k, b_k and θ_{ij} produce samples ε_{ijk} from its posterior distribution. Accordingly, posterior means and standard deviations of the Bayesian latent residuals can be computed from the sampled values. A more efficient estimator of the Bayesian latent residuals is the conditional expectation given a sufficient statistic, called a Rao-Blackwellised estimator (Gelfand & Smith, 1990). That is, the sampling error attributable to the Gibbs sampler is reduced to obtain a more efficient estimate of the posterior means. The unbiased character of the Monte Carlo estimator remains while reducing its variance.

For the binary response data, it follows that, the conditional expectation of the Bayesian latent residuals can be computed given the model parameters. Suppose that $Y_{ijk} = 1$, it follows that

$$\begin{aligned} E\left(\varepsilon_{ijk} \mid Y_{ijk} = 1, \theta_{ij}, a_k, b_k\right) &= \int_0^{\infty} E\left(\varepsilon_{ijk} \mid z_{ijk}, y_{ijk}, \theta_{ij}, a_k, b_k\right) \\ &\quad \cdot \frac{f\left(z_{ijk}, y_{ijk} \mid \theta_{ij}, a_k, b_k\right)}{f\left(y_{ijk} \mid \theta_{ij}, a_k, b_k\right)} dz_{ijk} \\ &= \frac{\phi\left(b_k - a_k\theta_{ij}\right)}{\Phi\left(a_k\theta_{ij} - b_k\right)}, \end{aligned} \tag{12.7}$$

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

$$E\left(\varepsilon_{ijk} \mid Y_{ijk} = 0, \theta_{ij}, a_k, b_k\right) = \frac{-\phi\left(b_k - a_k\theta_{ij}\right)}{\Phi\left(b_k - a_k\theta_{ij}\right)}. \tag{12.8}$$

It follows, in the same way, for polytomous data using equation 12.2 and 12.6, that

$$E\left(\varepsilon_{ijk} \mid Y_{ijk} = c, \theta_{ij}, a_k\right) = \frac{\phi\left(\kappa_{kc} - a_k\theta_{ij}\right) - \phi\left(\kappa_{kc-1} - a_k\theta_{ij}\right)}{\Phi\left(\kappa_{kc} - a_k\theta_{ij}\right) - \Phi\left(\kappa_{kc-1} - a_k\theta_{ij}\right)}, \tag{12.9}$$

Some elementary calculations have to be done to find expressions for the posterior variances of the residuals, but they can be derived in the same way. As a result, sampled values of the model parameters can be used to compute the estimates for the residuals and their variance. The estimates of the Bayesian latent residuals are easily computed within the Gibbs sampling procedure. Then, it can be checked if the Bayesian latent residuals are normally distributed given the observations by a quantile-quantile plot.

12.3.2 Detection of Outliers

The outlier detection problem is addressed from a Bayesian perspective. As just discussed, realized regression error terms are treated as unknown parameters, see Zellner (1971). The posterior distribution of these residuals can be used to calculate the posterior probability that an observation is an outlier. Outliers can be detected by examining the posterior distribution of the error terms. An observation can be considered to be outlying if the posterior distribution of the corresponding residual is located far from its mean (Albert & Chib, 1995). Here, the posterior distributions of the Bayesian latent residuals are examined to detect outliers among the observations. The Bayesian latent residuals are a function of unknown parameters (Equations 12.5 and 12.6) and the posterior distributions are therefore straightforward to calculate.

Following Chaloner and Brant (1988), Johnson and Albert (1999) and Zellner (1971), Y_{ijk} 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_{ijk} is marked as an outlier if $P\left(|\varepsilon_{ijk}| > q \mid y_{ijk}\right)$. In fact the augmented continuous scores Z_{ijk} are marked as outliers but Z_{ijk} has a one-to-one correspondence with Y_{ijk}, given the ability and item parameters. The probability that an observation exceeds a prespecified value is called the outlying probability. The outlying probabilities can be estimated with the Gibbs sampler.

Consider the residuals at the IRT level. It follows that, analogous to Equation 12.7, if $Y_{ijk} = 1$

$$\begin{aligned} P\left(|\varepsilon_{ijk}| > q \mid Y_{ijk} = 1, \theta_{ij}, a_k, b_k\right) &= \int_q^\infty \frac{f\left(z_{ijk}, y_{ijk} \mid \theta_{ij}, a_k, b_k\right)}{f\left(y_{ijk} \mid \theta_{ij}, a_k, b_k\right)} dz_{ijk} \\ &= \frac{\Phi\left(-q\right)}{\Phi\left(a_k \theta_{ij} - b_k\right)} \end{aligned} \tag{12.10}$$

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

$$P\left(|\varepsilon_{ijk}| > q \mid Y_{ijk} = 0, \theta_{ij}, a_k, b_k\right) = \frac{\Phi\left(-q\right)}{1 - \Phi\left(a_k \theta_{ij} - b_k\right)}. \tag{12.11}$$

In the same way, in case of polytomous data, it follows from Equation 12.2 and Equation 12.9 that

$$P\left(|\varepsilon_{ijk}| > q \mid Y_{ijk} = c, \theta_{ij}, a_k\right) = \frac{\Phi\left(\kappa_{kc}\right) - \Phi\left(q\right)}{\Phi\left(\kappa_{kc} - a_k \theta_{ij}\right) - \Phi\left(\kappa_{kc-1} - a_k \theta_{ij}\right)}.$$

Again, these expressions can be used to estimate the outlying probabilities of the estimated Bayesian latent residuals given sampled values of the model parameters. It is possible to find q such that the probability $P\left(|\varepsilon_{ijk}| > q \mid y_{ijk}\right)$ assumes a given percentage, say ν. Therefore, in every Gibbs iteration q must be solved in the equation $P\left(|\varepsilon_{ijk}| > q \mid y_{ijk}\right) = \frac{\nu}{100}$. The mean of these values is an estimate of the unique root, that is, the q-percent value, or the probability that Z_{ijk} 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 high posterior probability of being greater than two standard deviations.

Notice that other complex posterior probabilities can be computed with the Gibbs sampler by keeping track of all the possible outcomes of the relevant probability statement. However, this method has the drawback that a lot of iterations are necessary to get a reliable estimate. It could be possible, for example, that in case of multiple outliers, a test for a single outlier does not detect one outlier in the presence of another outlier. This so-called masking occurs when two posterior probabilities, for example, $P\left(|\varepsilon_{ijk}| > q \mid y_{ijk}\right)$ and $P\left(|\varepsilon_{sjk}| > q \mid y_{sjk}\right)$, do not indicate any outliers but the posterior probability $P\left(|\varepsilon_{ijk}| > q \text{ and } |\varepsilon_{sjk}| > q \mid \mathbf{y}\right)$ shows that ε_{ijk} and ε_{sjk} are both outliers. This simultaneous probability can be estimated by counting the events that both absolute values of the residuals are greater than q times the standard deviation divided by the total number of iterations.

12.4 Heteroscedasticity

In a standard linear multilevel model (Equation 12.3) the residuals at Level 1 and 2 are assumed to be homoscedastic. It is possible that the variances of the residuals are heteroscedastic when they depend on some explanatory variables. Homoscedastic variances can be obtained when modelling the variation as a function of the explanatory variables. Neglecting the heteroscedasticity may lead to incorrect inferences concerning the hypotheses tests for variables which are responsible for the heteroscedasticity (Snijders & Bosker, 1999, pp. 126-128). In a Bayesian framework, complex variance structures can be defined as prior information. Here, Level 1 variation is considered but the same principles apply to higher levels. General functions of more than one explanatory variable can be considered to model the variance at Level 1. Examples of complex variation modelling are given in, for example, Goldstein (1995, p. 50) and Snijders & Bosker (1999, p. 110-119).

Two tests for heteroscedasticity at Level 1 in case of two or more groups are considered that are easy to compute using the MCMC output, sampled under the assumption of homoscedasticity. Notice that the groups considered here, denoted as $l = 1, \ldots, L$, may differ from the groups, $j = 1, \ldots, J$, defined at Level 2 of the multilevel model. Testing the equality of variances of two or more grouped residuals at Level 1 coincides with testing the hypothesis, $\sigma_1^2 = \ldots = \sigma_L^2$ against the alternative $\sigma_l^2 \neq \sigma_{l'}^2$ for at least one $l \neq l'$. Highest posterior density intervals (HPD) can be used to test the equality of the group specific variances. In the first case, $L = 2$, the posterior distribution of the group specific variances can be derived, and in the general case, the posterior distribution of a function of the group specific variances can be approximated to obtain the HPD regions. The second approach is based on a normal approximation to the posterior distribution of the group specific variances. Testing heteroscedasticity at Level 1 can be transformed to testing the equality of the means of normal distributed variables, which is a much easier problem.

12.4.1 Highest Posterior Density Intervals

In a Bayesian posterior inference the marginal posterior distributions are summarized. Often, $100(1-\alpha)\%$ posterior credible intervals are given which are easy to obtain, but a highest posterior density interval (HPD) may be more desirable when the marginal posterior distributions are not symmetric (Box & Tiao, 1973). HPD regions are very appealing because they group the most likely parameter values and do not rely on normality or asymptotic normality assumptions. Chen and Shao (1999) showed how to obtain these HPD intervals given the MCMC samples generated from the marginal posterior distributions. From this it may seem that HPD intervals can only be used when MCMC samples from the parameters of interest are available. In some cases, HPD intervals can be computed without sampled values from the marginal posterior distributions and without evaluating the marginal posterior distributions analytically or numerically. This can be useful in hypothesis testing when some of the parameters of interest are not estimated in the MCMC procedure. Here, an example is provided for testing a particular hypothesis, homoscedasticity, without having to estimate the complete model, including all parameters.

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

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

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

$$\frac{\sigma_2^2}{\sigma_1^2} \sim \frac{s_2^2}{s_1^2} F\left(n_1-1, n_2-1\right), \tag{12.12}$$

see Box and Tiao (1973, pp. 110-112). The mode of the distribution of F is 1, thus the mode of the posterior distribution of σ_2^2/σ_1^2 is s_2^2/s_1^2. The limits of the HPD interval are specified by the F distribution in combination with an estimate of s_2^2/s_1^2, using the sampled values of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ from their marginal posterior distribution. In general, the group specific variance s_l^2, of group l can be estimated with M samples of $(\boldsymbol{\theta}, \boldsymbol{\beta})$, that is

$$\hat{s}_l^2 = \frac{1}{M} \sum_{m=1}^{M} \left[\sum_{i \in l} \left(\theta_{ij}^{(m)} - \mathbf{X}_{ij} \boldsymbol{\beta}_j^{(m)} \right)^2 \right]. \tag{12.13}$$

In a more general case, assume L group specific Level 1 variances. To assure that comparisons of L scale parameters $\left(\sigma_1^2, \ldots, \sigma_L^2\right)$ are unaffected by any linear recoding of the data, consider $(L-1)$ linearly independent contrasts in $\log \sigma_l^2$. So, let $\Delta_l = \log \sigma_l^2 - \log \sigma_L^2$. The vector $\boldsymbol{\Delta}_0 = \mathbf{0}$ is included in the highest posterior density region of content $(1-\alpha)$ if and only if

$$P\left[p\left(\boldsymbol{\Delta} \mid \mathbf{y}\right) > p\left(\boldsymbol{\Delta}_0 \vert\, \mathbf{y}\right) \mid \mathbf{y}\right] < 1 - \alpha.$$

The density function $p\left(\boldsymbol{\Delta} \mid \mathbf{y}\right)$ is a monotonic decreasing function of a function with parameters σ_l^2 and s_l^2 which is asymptotically distributed as χ_{L-1}^2, as $n_l \to \infty$, $l = 1, \ldots, L$, where s_l^2 is the mean sum of squares in group l (Box & Tiao, 1973, pp. 133-135). In case of the hypothesis $\boldsymbol{\Delta}_0 = \mathbf{0}$, which corresponds to the situation $\sigma_1^2 = \ldots = \sigma_L^2$, this function becomes

$$M_0 = -\sum_{l=1}^{L} n_l \left(\log s_l^2 - \log \overline{s}^2 \right) \tag{12.14}$$

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

$$\lim_{n_l \to \infty} P\left[p\left(\boldsymbol{\Delta} \mid \mathbf{y}\right) > p\left(\boldsymbol{\Delta}_0 \vert\, \mathbf{y}\right) \mid \mathbf{y}\right] = P\left(\chi_{L-1}^2 < M_0\right).$$

Hence, for large samples, the point $\boldsymbol{\Delta}_0 = \mathbf{0}$ is included in the $(1-\alpha)$ highest posterior density region if

$$M_0 < \chi_{L-1,\alpha}^2.$$

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

$$P\left[p\left(\boldsymbol{\Delta} \mid \mathbf{y}\right) > p\left(\boldsymbol{\Delta}_0 \vert\, \mathbf{y}\right) \mid \mathbf{y}\right] \approx P\left(\chi_{L-1}^2 < \frac{M_0}{1+A}\right), \tag{12.15}$$

where $A = \frac{1}{3(L-1)} \left(\sum_{l=1}^{L} n_l^{-1} - N^{-1} \right)$. The difficulty in practice with this test for equal variances is the sensitivity to the assumption of normality (Lehmann, 1986, p. 378).

The sampled values of the parameters, $(\boldsymbol{\theta}, \boldsymbol{\beta})$, can be used to compute the righthand side of Equation 12.14 using Equation 12.13. Notice that it is not necessary to estimate the model with the assumption of heteroscedasticity at level 1. It is possible to compute the highest posterior density of $\left(\sigma_1^2,\ldots,\sigma_L^2\right)$ given the observed data by integrating over the random effects $(\boldsymbol{\theta},\boldsymbol{\beta})$ and computing the probability density, in every iteration of the Gibbs sampler. The highest posterior density region should be constructed in such a way that the probability of every set of interior points is at least as large that of any set of exterior points. Furthermore, the region should be such that for a given probability, it occupies the smallest possible volume in the parameter space. The obtained vectors of parameter values can be used to construct such a region. Accordingly, the equality of variances can be tested by checking if the vector $\left(\sigma_1^2,\ldots,\sigma_L^2\right)=\mathbf{0}$ lies within the highest posterior density region.

12.4.2 Normal Approximation to the Posterior Distribution

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

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

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

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

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

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

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

$$\int_{C}^{\infty} \chi_{L-1}^{2}\left(y\right) dy = \alpha, \tag{12.16}$$

where α denotes the significance level. If the number of observations per group differ then the transformation s_l'/λ_l, with $\lambda_l = 2/\left(n_l - 1\right)$, results in a test which rejects the hypothesis of equal variances when

$$\sum_{l=1}^{L}\left(\frac{s_l'}{\lambda_l}\right)^2 - \frac{\left(\sum_{l=1}^{L} s_l'/\lambda_l^2\right)^2}{\sum_{l=1}^{L}\left(1/\lambda_l^2\right)} > C, \tag{12.17}$$

where C is determined by (12.16), see Lehmann (1986, p. 377). The Gibbs sampler is used to estimate the s_l' for every group l using the sampled values for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and Equation 12.13. That is, after a sufficient number of iterations, the test statistic is computed to test the homogeneity of variances.

12.5 An Analysis of a Dutch Primary School Mathematics Test

This section is concerned with the study of a primary school advancement test. In Fox and Glas (2001), this data set was analyzed to compare parameter estimates of a multilevel IRT model and an hierarchical linear model using observed scores. Here, the goodness of fit of the multilevel IRT model is analyzed. Residuals at different levels are analyzed, outliers are identified and different models are compared. Also, heteroscedasticity at Level 1 is tested.

The data set consisted of responses from 2156 grade 8 students, unequally divided over 97 schools, to 18 dichotomously scored mathematics items taken from the school advancement examination developed by the National Institute for Educational Measurement (Cito). The 97 schools were fairly representative of all Dutch primary schools (Doolaard, 1999). Of the 97 schools sampled, 72 schools regularly participated in the school advancement examination, denoted as Cito schools and the remaining 25 schools are denoted as the non-Cito schools. Socioeconomic status (SES),

Table 12.1: Parameter Estimates of a Multilevel IRT Model With Explanatory Variable End at Level 2.

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

scores on a nonverbal intelligence test (ISI; Van Boxtel, Snijders, & Welten, 1982), and gender were used as predictors for the students' mathematical ability. SES was based on four indicators: The education and occupation level of both parents (if present). Predictors SES and ISI were standardized. The dichotomous predictor Gender was an indicator variable equal to 0 for males and equal to 1 for females. Finally, a predictor variable labelled End equaled 1 if the school participated in the school advancement test, and 0 if this was not the case.

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

$$\begin{aligned}\theta_{ij} &= \beta_{0j} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\mathrm{End}_j + u_{0j}\end{aligned} \tag{12.18}$$

where $e_{ij} \sim N\left(0, \sigma^2\right)$, $u_{0j} \sim N\left(0, \tau_0^2\right)$. The model contains random group effects and random variation within groups. The dependent latent variable equals the sum of a general mean γ_{00}, a random effect at the school level, u_{0j}, and a random effect at the individual level, e_{ij}, corrected for the predictor End. The two-parameter normal ogive model was used as the measurement model. In Table 12.1, the estimates of the parameters issued from the Gibbs sampler are given. The reported standard deviations and HPD regions are the posterior standard deviations and the 90% highest posterior density intervals, respectively.

The general mean ability, γ_{00}, of the students attending non-Cito schools was not significantly different from zero. The positive significant value of

γ_{01} indicates a positive effect of participating in the school advancement exam on the students' abilities. The intraclass-correlation coefficient was approximately .26, which is the proportion of variance accounted for by group membership given the explanatory variable End.

The behavior of the Bayesian latent residuals for this data set were considered. The Bayesian latent residuals, the probabilities of a correct response, and the outlying probabilities, that is, the probabilities that the residuals were larger than 2, were estimated using Equations 12.7, 12.8, 12.10, and 12.11. In Figure 12.1, all the Bayesian residuals (Equation 12.4), and all Bayesian latent residuals (Equation 12.5) are plotted against the corresponding fitted probability of a correct response. The observed item responses determine the domain of the Bayesian residuals, that is, if $Y_{ijk} = 1$ then $r_{ijk} \in (0, 1)$ and otherwise $r_{ijk} \in (-1, 0)$. The Bayesian latent residuals are also grouped by the value of Y. If the answer is correct, $Y_{ijk} = 1$, the Bayesian latent residual, ε_{ijk}, is positive, otherwise, it is negative, but there is no ceiling-effect for the Bayesian latent residuals. Figure 12.1 shows that extreme valued Bayesian latent residuals are discovered more easily because they are not restricted in size as the Bayesian residuals. Next, we show how to identify outliers. The extreme Bayesian residuals and Bayesian latent residuals correspond to the same observed data.

Figure 12.2 displays marginal posterior distributions of Bayesian latent residuals and Bayesian residuals corresponding to, the same, 25 randomly selected answers to item 17. The order of the posterior means of the Bayesian residuals and the Bayesian latent residuals is the same. Negative posterior means correspond to incorrect answers and positive posterior means correspond to correct answers. As a result, the marginal posterior distributions of the Bayesian residuals are defined on $(-1, 0)$ if the corresponding observation equals zero, and on $(0, 1)$ otherwise. The marginal posterior distributions of the Bayesian residuals differ. This makes it is difficult to assess how extreme the marginal posterior distributions are. Subsequently, it is difficult to identify outliers given the posterior means that are estimates of the Bayesian residuals and their marginal posterior distributions. The marginal posterior distributions of the Bayesian latent residuals are standard normal, according to Equation 12.5. This provides a convenient basis to test the presence of outliers by examining whether the posterior means of the marginal posterior distributions are significantly different from zero. As a result, the Bayesian latent residuals are easy to interpret and interesting for identifying outliers. In Figure 12.2, the four smallest posterior means of the Bayesian latent residuals are significantly smaller than zero using a 5% significance level, and the corresponding observations can be regarded as outliers. These outliers cannot be identified directly by visual inspection of the marginal posterior distribution of the

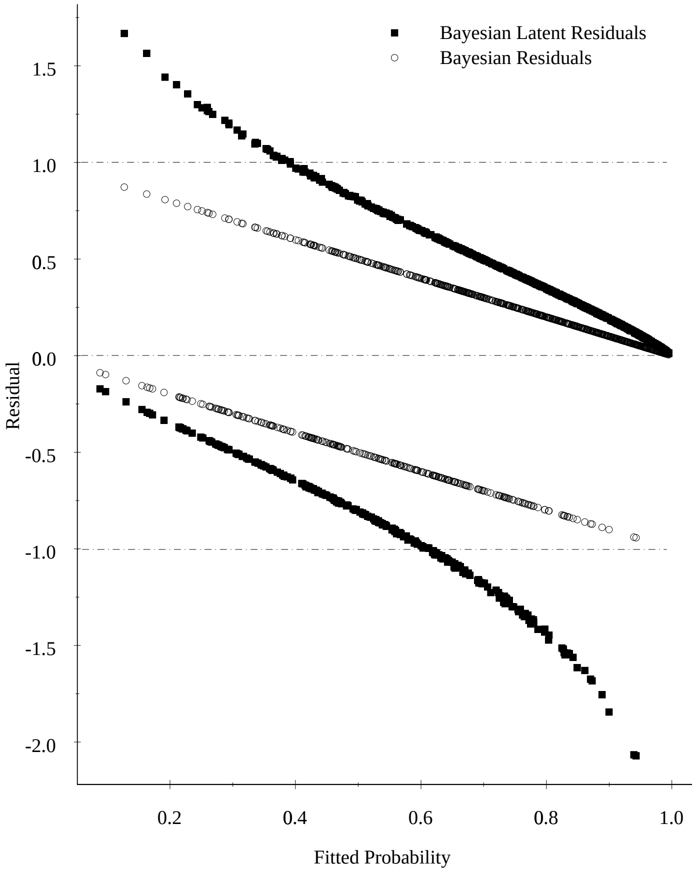

Figure 12.1: Bayesian Latent Residuals and Bayesian Residuals Plotted Against the Probabilities of a Correct Response.

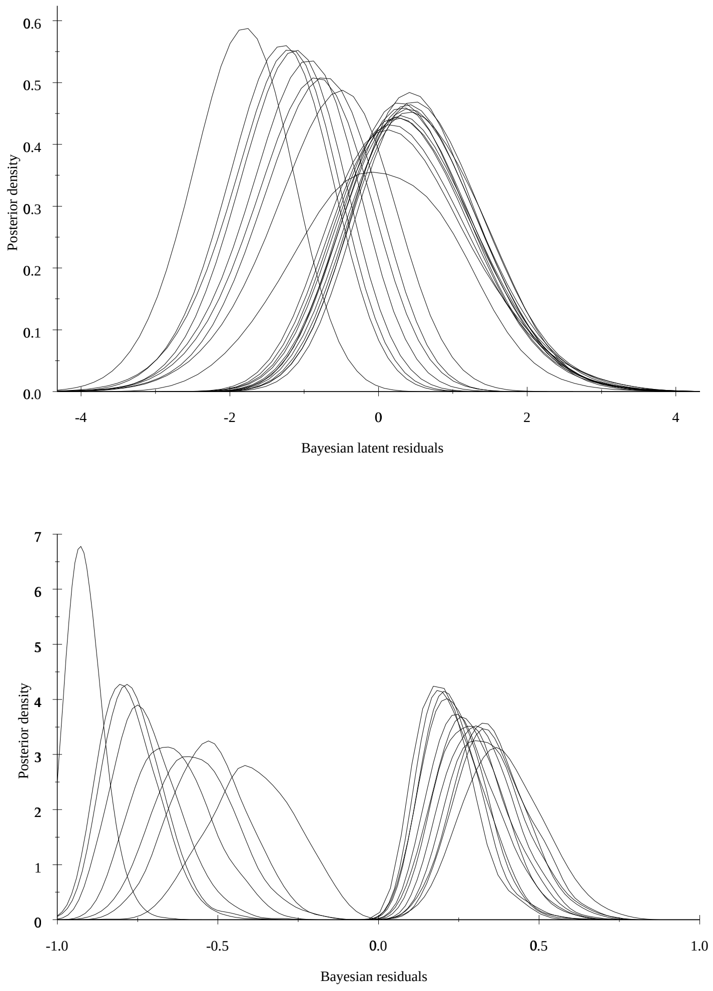

Figure 12.2: Posterior Distributions of Bayesian Latent Residuals and Bayesian Residuals Corresponding to Item 17.

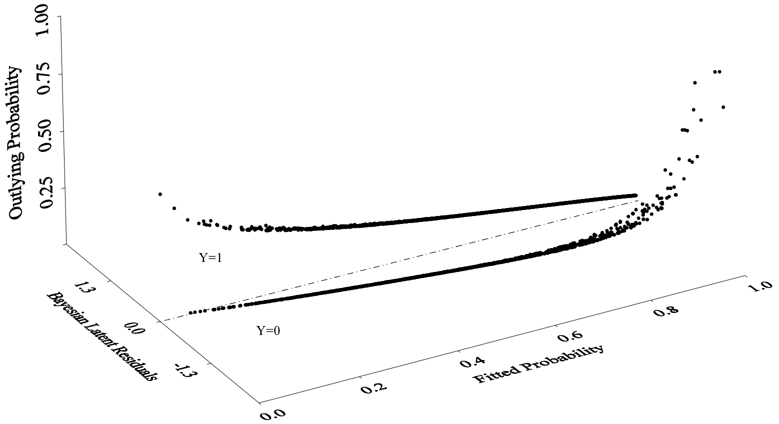

Figure 12.3: Bayesian Latent Residuals Plotted Against the Probabilities of a Correct Response and the Outlying Probabilities.

Bayesian residuals.

In Figure 12.3, all Bayesian latent residuals, ε_{ijk}, are plotted against the probabilities of a correct response of person i in group j to item k, and the outlying probabilities, where the outlying probabilities were computed for $q = 2$, using Equation 12.10 and Equation 12.11. Successes, $Y_{ijk} = 1$, with fitted probabilities close to one and failures, $Y_{ijk} = 0$, with fitted probabilities close to zero correspond to small absolute values of the residuals. The outlying probability increases if the value of the residual increases. The points with low fitted probabilities corresponding to correct answers and high fitted probabilities corresponding to incorrect answers can be marked as outliers. More specific, when the corresponding outlying probability is higher than a 5% significance level the corresponding observation can be marked as an outlier. Obviously, Figure 12.3 shows that there are a lot of outliers, approximately 6% of the observations, so the model doesn't fit the data very well.

Fitted probabilities close to one corresponding to successes and fitted probabilities close to zero corresponding to failures have residual distributions that resemble standard normal curves. That is, the distributions of the residuals are not influenced by the observations. However, the observations have a large influence on the posterior distributions of the residuals when the fitted probabilities are in conflict with the observations. In Figure 12.4, posterior distributions of the Bayesian latent residuals corresponding to Item 17 of the math test of several students are plotted. Some of the residuals can be marked as outliers because their posterior distributions

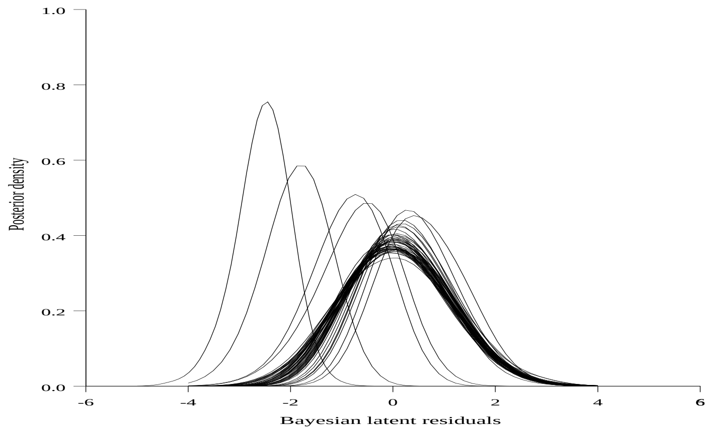

Figure 12.4: Posterior Densities of the Bayesian Latent Residuals Corresponding to Item 17.

differ from the standard normal distribution. The nonzero location and the smaller standard deviation of the posterior distributions of these residuals express the conflict between the observations and the fitted probabilities. For example, the outlying probability of the largest residual in Figure 12.4 is .982. The corresponding response pattern showed that all items were scored correct except Item 17, although it was answered correctly by 88% of the students.

It was assumed that the nonverbal intelligence test and the socio-economic status provide information about the math abilities. Therefore, Model M_1 (Equation 12.18) was extended with these Level 1 predictors, that is

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

where $e_{ij} \sim N\left(0, \sigma^2\right)$ and $u_{0j} \sim N\left(0, \tau_0^2\right)$. In the sequel, this model is labelled M_2. Here, it was assumed that the effects of the scores of the intelligence test and the SES of the students did not differ per school, that is, the random regression coefficients were fixed over schools. The parameter estimates resulting from the Gibbs sampler are shown in Table 12.2.

The residual variance at Level 1 and Level 2 decreased due to the pre-

Table 12.2: Parameter Estimates of a Multilevel IRT Model With Explanatory Variables ISI and SES at Level 1 and End at Level 2.

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

dictors at Level 1. The coefficients of both predictors are significant. As expected, SES and intelligence (ISI) have a positive effect on the achievements. The likelihood of model M_1 is higher than the likelihood of model M_2 indicating that model M_2 fits the data better. On the other hand, there are no significant differences between the Bayesian latent residuals of model M_1 and M_2. Many outliers under model M_1 are also outliers under model M_2. Also, the estimated posterior means of the residuals are similar. Changing the structural multilevel model did not result in major differences in the measurement model. It turned out that the explanatory variables, ISI and SES, explained variance in the latent dependent variable but did not result in different parameter estimates of the measurement model. So, the structural multilevel model M_2 is preferred, but the introduction of the explanatory variables did not reduce the number of outliers.

The residuals at Level 1 were assumed to have a constant variance, that is, they were assumed to be homoscedastic. It was investigated whether the residual variance at Level 1 differed between male and female students. Model M_2 was estimated again under the assumption of unequal variances, that is, each group specific residual error variance was sampled during the parameter estimation of model. Also, the other model parameters were estimated given the sampled values of the group specific variances. The marginal posterior distributions of the group specific error variances, for both the male and the female group, are given in the top-figure of Figure 12.5. Because the posterior distributions are overlapping, it can be concluded that the group specific error variances at Level 1 are not significantly different.

It was investigated whether the statistics for testing heteroscedasticity at Level 1 (Equations 12.12, 12.15, and 12.17) yielded the same conclusion using the MCMC output generated under the assumption of homoscedastic variances at Level 1. The MCMC output was used to compute the sum of squares of the group specific residuals (Equation 12.13). A HPD region for the ratio of variances was derived from Equation 12.12. The 90% HPD region of the ratio of the two group specific residual variances is $[.84, 1.04]$. Thus, the point of equal variances is included in the 90% region. In Figure 12.5, the bottom figure shows the posterior distribution of the variance ratio and illustrates the 90% HPD region. This ratio consists of the residual error variance within the male group divided by the residual variance within the female group. The posterior mean of the variance ratio is shifted toward the left of one. Therefore, the residual variance within the female group is slightly, but not significantly, higher. This corresponds with the estimated posterior means of the group specific residual variances in the top-figure of Figure 12.5. The other test statistics (Equations 12.14 and 12.17) were computed using the same MCMC output. Both means of the computed test statistics correspond with a p-value of .27. Therefore, it can be concluded that there are no indications of residual variance differences between the male and female groups at Level 1.

Two multilevel IRT models were investigated using the methods described in this chapter. It was shown that the measurement model did not fit the data very well because many outliers were detected. Model M_2 was analyzed to illustrate that changing the structural part will not improve the fit of the measurement part. Conclusions drawn from the multilevel analysis can be wrong when the measurement part does not fit the observed data. Therefore, a further analysis should at least consider other measurement models.

12.6 Discussion

Methods for evaluation of the fit of a multilevel IRT model were discussed. It was shown that Bayesian latent residuals are easily estimated and particularly useful in case of dichotomous and polytomous data. Estimates of these Bayesian latent residuals can be used to detect outliers. Moreover, outlying probabilities of the residuals are easily computed using the MCMC output. Together, these estimates provide useful information regarding the

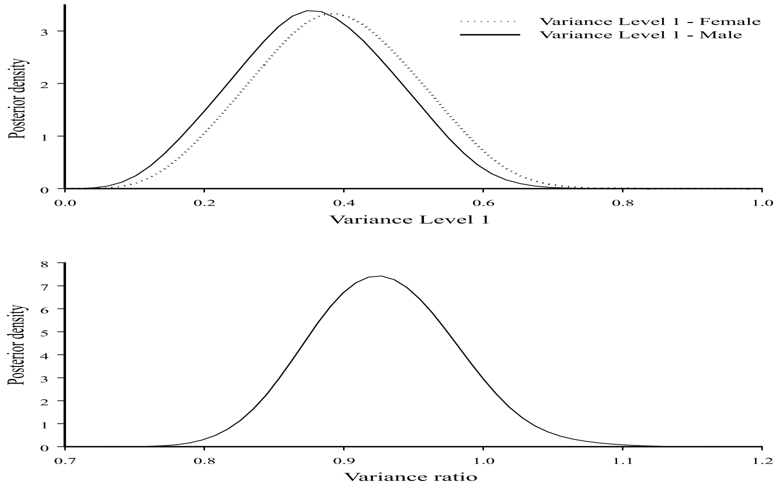

Figure 12.5: The Marginal Posterior Distribution of the Residual Variance at Level 1 in the Male and Female Group and the Posterior Distribution of the Ratio of Both Variances.

fit of the model. One particular assumption of the multilevel IRT model is homoscedasticity at Level 1. Several tests are given to check this assumption. They can be computed as a byproduct of the Gibbs sampler.

Further research will focus on summarizing the information regarding the detected outliers. Then, diagnostic tests can be developed to detect respondents with misfitting response patterns or items that induce outliers. These tests will provide more assistance in the search for a better model instead of just providing information regarding the fit of the model.

Another class of tests not discussed in this chapter, to check the discrepancy between the model and the data, are the so called *posterior predictive checks*, introduced by Rubin (1984). Posterior predictive checks consist of quantifying the extremeness of the observed value of a selected discrepancy. Several general discrepancies are developed but this can be any function of the data and the model parameters (Meng, 1994; Gelman, Meng, & Stern, 1996). Obviously, these tests can be used to judge the fit of a multilevel IRT model. More research is required into the relation between the tests described here and posterior predictive checks.

The connection of the discrete observed responses Y_{ijk} to continuous latent responses Z_{ijk} has several advantages. The problem of estimating all parameters reduces to sampling from standard distributions. The Bayesian latent residuals provide information concerning the fit of the model and possible outliers are easily detected. This simulation technique introduces extra randomness in the estimation procedure, therefore, establishing the convergence of the algorithm requires extra attention.

References

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

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

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

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

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

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

Chaloner, K., & Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika, 75*, 651-659.

Chen, M. -H., & Shao, Q. -M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics, 8*, 69-92.

Doolaard, S. (1999). *Schools in change or schools in chains.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

Fox, J. -P. (in press). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology.*

Fox, J. -P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 269-286.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*, 972-985.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gelman, A., Meng, X. -L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Arnold.

Johnson, V. E., & 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.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Springer.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*, 307-330.

Meng, X. -L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22,* 1142-1160.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2000). *HLM 5. Hierarchical linear and nonlinear modeling.* Lincolnwood, IL; Scientific Software International, Inc.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12,* 1151-1172.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17.*

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis.* London: Sage.

Van Boxtel, H. W., Snijders, J., & Welten, V. J. (1982). *ISI: Interesse, Schoolvorderingen, Intelligentie.* [ISI: Interest, school progress, intelligence.] Publicatie 7. Vorm III. Groningen, The Netherlands: Wolters-Noordhoff.

Verhelst, N. D., & Eggen, T. J. H. M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek* [Psychometric and statistical aspects of measurement research.] (PPON rapport 4). Arnhem, The Netherlands: Cito.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics.* New York: Wiley.

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

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245-256). New York: Springer.