Communications in Statistics—Simulation and Computation®, 45: 1499–1510, 2016
Published with license by Taylor & Francis Group, LLC
ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2014.939518

Taylor & Francis
Taylor & Francis Group

Complex Latent Variable Modeling in Educational Assessment

JEAN-PAUL FOX,[1] MAARTEN MARSMAN,[1]
JORIS MULDER,[2] AND JOSINE VERHAGEN[3]

[1]Department of Research Methodology, Measurement and Data Analysis, University of Twente, The Netherlands
[2]Department of Methodology and Statistics, Tilburg University, The Netherlands
[3]Department of Psychological Methods, University of Amsterdam, The Netherlands

Bayesian item response theory models have been widely used in different research fields. They support measuring constructs and modeling relationships between constructs, while accounting for complex test situations (e.g., complex sampling designs, missing data, heterogenous population). Advantages of this flexible modeling framework together with powerful simulation-based estimation techniques are discussed. Furthermore, it is shown how the Bayes factor can be used to test relevant hypotheses in assessment using the College Basic Academic Subjects Examination (CBASE) data.

Keywords Bayes factor; Bayesian modeling; Latent variable models; MCMC.

Mathematics Subject Classification 6207; 62F15; 62F99; 62H25.

1. Introduction

In educational studies, psychometric methods are focused on measuring constructs (e.g., ability, attitudes), and developing and validating measurement instruments. The constructs, also referred to as latent variables, cannot be observed directly. Assessment data, which typically consist of multiple observed variables, are required to measure a latent variable.

Item response theory (IRT) models (Van der Linden and Hambleton, 1997) have been widely used to study the relationship between observed and latent variables. This latent variable model is particularly popular in educational and psychological measurement. The basic IRT models can be viewed as a nonlinear mixed effects model. At the level of observations, a nonlinear or generalized linear relationship is defined between the item observations and the latent variable. The correlated item observations are nested in a latent variable, while accounting for measurement error. The item observations are regressed on a

Received January 30, 2014; Accepted June 23, 2014
Address correspondence to Jean-Paul Fox, Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands; E-mail: j.p.fox@utwente.nl

latent variable through a link function that is usually the probit or logit function. A second component is needed to model the distribution of the latent variable(s). The latent variable distribution describes the population of subjects. In a common test situation, a normal population distribution is assumed. In a frequentist modeling approach, item characteristics are treated as fixed effects, although it is also possible to model them as random effects.

The common IRT models are most often not sufficient when dealing with data stemming from complex assessments. The complex nature of an assessment might lead to model violations and incorrect statistical inferences. Model adjustments are necessary to respond to non-standard test situations. In standard IRT modeling, it is assumed that the correlations between observed variables can be explained by a latent variable and (fixed) item effects. However, for example, when subjects are not independently sampled from a population but a stratified or clustered sample was taken, correlations between observations are not fully explained. The population distribution of the subjects needs to be altered appropriately to address the additional correlations between subjects.

Model adjustments are also necessary when researcher's interest is beyond the measurement of a latent variable. An appropriate model is needed to investigate relationships between explanatory observed variables and the latent variable and their relationship with the observed outcomes. The inclusion of explanatory variables in IRT modeling has received much interest but seriously complicates the modeling framework. Various model extensions have been proposed to account for multidimensional constructs and constructs measured at different levels of analysis. These extensions address the measurement of multiple constructs defined at different hierarchical levels (e.g., De Boeck and Wilson, 2004; Fox, 2010).

Model extensions have also been proposed just to improve the fit of the model, without relying on content-driven information. In the field of statistical modeling, more advanced models have been proposed to capture unobserved heterogeneity through mixture distributions, impose skewed population distributions, and to identify patterns of residual correlations over time, among other things. By accounting for the unobserved heterogeneity using statistical techniques, the model fits the data better without having to identify the practical nature of the extension.

Recently, Bayesian latent variable modeling has received considerable attention in different research fields. Through a Bayesian approach, advanced latent variable models have been applied to analyze complex item response data leading to powerful statistical inferences. The Bayesian approach has several advantages. It allows to incorporate prior information in the analysis, besides the data information. The prior information can come from expert views or a previous analysis. It is also possible to include parameter restrictions, a functional or probabilistic relationship, as prior information. Data and prior information are used, reflecting the accumulation of evidence, to make statistical inferences. Another advantage is that uncertainty can be quantified (i.e., expressing certain beliefs), and inferences can be made through probabilistic reasoning. The Bayesian modeling approach stimulates the use of uncertainty in parameter estimation but also in subsequent analysis as model fit. Furthermore, the whole procedure is conceptually the same for simple and advanced problems. Unknown quantities (e.g., missing data, latent variables, parameters) are treated in a similar way. After specifying a prior for an unknown, posterior inferences can be made by conditioning on all information using the posterior distribution.

The attractive features of Bayesian modeling are accompanied with powerful simulation methods to estimate model parameters and to test hypothesis. Recent developments in simulation-based estimation techniques show potential to give support to the demand for more advanced Bayesian latent variable models. Bayesian estimation methods have several

advantages over maximum likelihood (ML) estimation methods. First, Bayes estimates do not rely on large sample theory and the distribution of the estimates is not necessarily normal, where the distribution of ML estimates is assumed to be normal (large sample theory) and the standard errors are based on a symmetric distribution. A Bayesian credible interval is based on the posterior distribution, which is allowed to be strongly skewed. The Bayesian exploration of model fit can also be directly based on the posterior distributions. Second, when dealing with categorical outcomes and many latent variables, ML estimates can be difficult or even impossible to compute, since it requires high-dimensional numerical integration. In this situation, the Bayesian computations are less cumbersome, and with diffuse priors the ML and Bayes estimates are comparable for large sample sizes. Third, the Bayesian estimation methods give support to investigate complex models with a huge number of parameters, where ML methods do not provide a natural way.

This has led to many Bayesian applications of item response data analyses in different research fields. Under the name Bayesian IRT models, referred to as BIRTs, different applications can be given, for example, in health (He et al., 2010; Van den Berg et al., 2007), marketing (De Jong et al., 2007), and education (Wang et al., 2013) sciences.

The main objective is to give an overview of new developments in Bayesian modeling of item response data. Therefore, a short introduction to Bayesian item response modeling is given. Then, a more general Bayesian modeling framework will be given, with connections to different popular model extensions. Then, it is shown and illustrated that new developments in powerful simulation techniques support the complex Bayesian modeling approaches. Further developments in Markov chain Monte Carlo (MCMC) methods stimulate joint parameter estimation, and testing complex models using Bayes factors. In this light, the use of Bayes factor testing is illustrated and discussed using the CBASE data. Then, a discussion is given and suggestions for further research.

2. Basic Concepts of BIRTs

Consider multivariate item response data (level 1), which are nested within subjects (level 2). An IRT model is defined at level 1. Characteristic of IRT is that subject's observations are assumed to be conditionally independent given the latent variable level. The level 1 component handles the nesting of item responses within subjects. At level 2, the heterogeneity among subjects is modeled using a population distribution.

Let y_{ik} denote the response to item k ($k = 1, \ldots, K$) of subject i. In a unidimensional setting, the observed variables $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})^t$ are related to a single latent variable, denoted as θ_i. For binary response items, where a correct or incorrect response is observed, a two-parameter IRT model is considered to link the observed item responses to the latent variable. The two-parameter model will not account for guessing but will account for differences in item difficulties (represented by difficulty parameter b_k). Furthermore, it accounts for slope differences of the item-specific curves (represented by discrimination parameter a_k) such that a positive (additional) contribution to the latent variable will lead to an item-specific increase in the success probability.

For the two-parameter model, the probability of a correct response of subject i to item k is given by

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

where $F(.)$ denotes a cumulative distribution function. The item parameters (a_k, b_k) are often referred to as the discrimination and difficulty parameter. Although, they can also

be recognized as a factor loading and intercept value, according to the terminology of structural equation modeling. This follows from representing the measurement model in terms of a nonlinear mixed effect model. Therefore, assume a (mixed effect) probit model in Eq. (1) and let F^{-1} denote the inverse of the cumulative normal distribution function, it follows that

$$F^{-1}\left[P\left(Y_{ik}=1 \mid \theta_i, a_k, b_k\right)\right] = a_k(\theta_i - b_k). \tag{2}$$

The right-hand side is linearly related to the transformed expected response, since $E(Y_{ik} \mid a_k, b_k, \theta_i) = P\left(Y_{ik}=1 \mid \theta_i, a_k, b_k\right)$. When $\theta_i - b_k = 0$, the probability of success equals .50. So, the intercept or difficulty parameter can be interpreted as the level that needs to be matched to have a success of .50. The discrimination parameter a_k is a slope effect and quantifies the growth in success when increasing the construct level. In the literature, the logistic response formulation is often used, which would lead to

$$\text{logit}\left[P\left(Y_{ik}=1 \mid \theta_i, a_k, b_k\right)\right] = a_k(\theta_i - b_k), \tag{3}$$

where the logit is the inverse function of the logistic distribution function. The interpretation of the model parameters remains the same.

2.1. *Population Models*

At a higher level, the (population) distribution of the latent variable is described. Most often it is assumed that subjects are sampled independently from a population using a normal distribution. In that case,

$$\theta_i = \mu_\theta + e_{ik}, \tag{4}$$

where $e_{ik} \sim N\left(0, \sigma_\theta^2\right)$. The error term defines the between-subject heterogeneity and μ_θ the average performance level.

Although the normal population model is often used representing simple random sampling of subjects from a population, the subjects can also be sampled in a different way. When, for example, subjects are sampled in a stratified way, addressing a clustering of subjects in groups, the use of a normal distribution will give an incorrect representation. Scores from the same group contain less information than scores from different groups, and the scores should be differently weighted in the analysis. Otherwise stated, scores of subjects within a group are more correlated than those from different groups. As a result, the population distribution of subjects needs to address the characteristics of the sampling design.

Several modeling alternatives have been proposed to describe more complex sampling designs, where subjects are not independently sampled. The IRT model, Eq. (1), is extended with a population distribution, which reflects the structure of the sampling design. In case of a multistage sampling design, Fox (2010) and Fox and Glas (2001), among others, introduced a multilevel population model to describe the within-group dependencies and defined it as the level 2 component. This extension, defined as a multilevel IRT model, takes the survey design explicitly into account. The heterogeneity between subjects is explained due to differences within groups (between subjects) and differences between groups.

Related to this, consider the multiple-group IRT model, where the population consists of a fixed number of groups and specific interest is focused on differences between those groups. This multiple group IRT model has been generalized in several ways. Azevedo

et al. (2012) proposed other item response functions, such as the skew probit, logit, and the log-log, to improve the link between the response observations and the latent mean structure. The population distribution was also generalized in several ways by allowing, for example, finite mixture of normals to describe the heterogeneity of subjects in the population. Another extension is based on describing the nesting of students in unobserved groups (latent classes) to capture associations between latent groups of subjects. Vermunt (2003) and Cho and Cohen (2010) defined complex mixture distributions to describe the unobserved clustering of students and possibly groups of students.

2.2. *Models for Random Item Parameters*

Much attention has been given to describe the characteristics of the items within the test. The item characteristics can be assumed to be independent, but it is more realistic to assume a within-item correlation. In more complex sampling designs, it is also possible to model item characteristic differences between groups to describe the variability in item functioning over groups or time.

Therefore, consider multiple groups, denoted by $j = 1, \ldots, J$, and let the probability of a correct response of person i in group j be represented by an IRT model with a cumulative normal link function:

$$P(Y_{ijk} = 1 \mid \theta_{ij}, \tilde{b}_{kj}) = F(\theta_{ij} - \tilde{b}_{kj}), \tag{5}$$

with group-specific item parameter $\tilde{b}_{kj}$. To make sure the parameters in this model are identified, the group-specific item parameters are restricted to sum to zero within each group (for each group j, $\sum_k \tilde{b}_{kj} = 0$).

For the group-specific parameters, the prior structure can be defined in several different ways. One way is to assume a multilevel structure for the group-specific item parameters (De Jong et al., 2007; Fox, 2010; Fox and Verhagen, 2010; Verhagen and Fox, 2013a, 2013b). Let μ_{b_0} denote the average difficulty of all items and let u_k denote the average deviation of item k from the population mean difficulty. The group-specific item difficulty can now be specified via a two-level structure:

$$\tilde{b}_{kj} = \mu_{b_0} + u_k + e_{kj}, \tag{6}$$

where e_{kj} denotes the group-specific deviation of the item difficulty of item k from the population average for item k, $\mu_{b_0} + u_k$. For each item k, the group-specific deviations e_{kj} are assumed to be normally distributed with mean zero and variance $\sigma^2_{b_k}$. This variance component defines the variability in item functioning over groups in the population. The random component u_k is assumed to be normally distributed with mean zero and variance σ^2_b, and this variance component defines the variability in difficulties of items in the item bank.

Another way is to treat the groups as fixed, and to assume that the group-specific item characteristics in the different groups are related when they refer to the same item. Items that are more difficult in one group will probably be more difficult in the other groups as well. Hence, a multivariate normal model is imposed on the group-specific item characteristics, to model the variance in the random deviations of the item difficulties within groups and the correlations between group-specific deviations of the same item (see also De Boeck, 2008; Frederickx et al., 2010). The group-specific item parameters are specified as:

$$\tilde{b}_{kj} = e_{kj}. \tag{7}$$

As the average item difficulty in each group j is restricted to zero due to the choice of identification restrictions, e_{kj} displays the deviation of item difficulties from the general mean in group j.

The covariance matrix consists of the item parameter variance within each group (σ_{bj}^2) on the diagonal, and the covariance of item parameters between each pair of groups $\sigma_{b_j b_{j'}}$, $j \neq j'$ on the off-diagonal:

$$\mathbf{e}_k \mid \boldsymbol{\Sigma}_b \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_b)$$

$$\boldsymbol{\Sigma}_b = \begin{bmatrix} \sigma_{b_1}^2 & \sigma_{b_1 b_2} & \dots & \sigma_{b_1 b_J} \\ \sigma_{b_1 b_2} & \sigma_{b_2}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_{b_{J-1} b_J} \\ \sigma_{b_1 b_J} & \ddots & \sigma_{b_{J-1} b_J} & \sigma_{b_J}^2 \end{bmatrix}, \tag{8}$$

where $\mathbf{e}_k = (e_{k1}, \dots, e_{kJ})$. The variance of the item parameters over items (σ_{bj}^2) can differ over groups, indicating that there is more variation in item difficulties in one group than in the other.

In the next section, it will be shown that parameters of both models can also be estimated using a general sampling-based estimation method. Furthermore, inferences from both models can be obtained in a similar way through the Bayesian machinery.

3. Simulation-based Methods for Estimation

Over the last two decades, MCMC approaches to Bayesian inference for IRT models have become increasingly popular. Most applications use the Gibbs sampler, which is a well-known abstract divide-and-conquer algorithm for generating a dependent sample from a complex multivariate distribution. In each iteration, a sample is drawn from so-called full conditional distributions: that is, distributions of one (set of) variable(s) conditionally on all the other variable(s).

Straightforward application of the Gibbs sampler to IRT models results in intractable full conditional distributions. To enable sampling from these distributions, one of two procedures are often used: a data augmentation procedure or a Metropolis-type procedure. We will briefly summarize the procedures and discuss potential developments to address estimation problems in large-scale assessments.

3.1. *Data Augmentation*

In the normal ogive model, random observations, where Y_{ik} equals 1 if the response of person i to item k is correct and Y_{ik} equals 0 otherwise, are represented by Bernoulli random variables with probability of success $\pi_{ik} = F(\theta_i - b_k)$ and $F()$, where represents the standard Gaussian CDF. In a Bayesian framework, both the person and item parameters receive a prior distribution denoted by $p(\theta_i)$ and $p(b_k)$, respectively. Unfortunately, the posterior distribution of the person and item parameters is largely intractable since there are no prior distributions that are conjugate to the standard Gaussian. The key idea of Albert (1992) for this problem was to introduce latent responses, with $Z_{ik} \sim \mathcal{N}(\theta_i - b_k, 1)$, where Y_{ik} equals one if $Z_{ik} > 0$ and zero otherwise. Upon observing the latent responses, and using normal priors for the person and item parameters, the posterior distribution of the person and item parameters can be derived using normal linear model results. The latent

responses are of course unknown. However, conditional on the observed item responses they follow a truncated normal distribution and enables sampling from the joint posterior distribution of the person and item parameters using the Gibbs sampler.

The data augmentation approach marked the introduction of MCMC to Bayesian inference in IRT modeling. Since its inception, this approach has been extended to more elaborate normal ogive models, such as multidimensional models where it is assumed that the response process is governed by multiple abilities (Béguin, 2000) and multilevel models that assume multistage cluster sampling of persons (Fox and Glas, 2001).

The logistic counterpart of the normal ogive model known as the Rasch model has recently been addressed by Polson et al. (2013), who propose a Gibbs sampling procedure for logistic models analog to the data augmentation procedure. The latent responses follow a Polya-Gamma distribution, and if the prior distributions $p(\theta)$ and $p(b)$ are normal distributions, it follows that the posterior distributions of the person and item parameters can again be derived using normal linear model results.

Both data augmentation procedures can be used to estimate most commonly used Bayesian item response theory models for categorical data. The downside is that their approaches are only suited for the normal and logistic error models, and that data augmentation approaches increase the computational burden by increasing the amount of auto-correlation in the chain. This means that the Markov chain will converge slower and longer runs are required before a preset number of iid draws are obtained for posterior inference in comparison to procedures without data augmentation. However, the ease of implementation may outweigh the burden of running longer chains.

3.2. *Metropolis Algorithm*

Another approach to sample from intractable distributions is the Metropolis algorithm. Using a Metropolis algorithm within a Gibbs sequence has been proposed by Patz and Junker (1999), and is widely used to estimate models that could not fit in the data augmentation framework. Note that to use the Metropolis algorithm to sample from the posterior distribution of a parameter, we need not have a conjugate prior to obtain tractability, nor do we need augmented variables. However, efficient implementation of the Metropolis-within-Gibbs algorithm requires the formulation of proposal densities that generate few rejected samples. In practice, this requires model-specific fine-tuning to influence the step rate and size in the parameter space for each parameter in the problem.

Marsman et al. (in press) describe two previously published algorithms that can be used to sample from a single conditional distribution; a rejection algorithm mentioned by Rubin (1984) that was applied in the *European Survey of Language Competences* (ESLC; Maris, 2012) and a Metropolis-type algorithm known as the *Single-Variable Exchange* (SVE) algorithm developed by Murray et al. (2006). The similarity between the algorithms is that both are based on the observation that a sample from a conditional distribution, say $p(\theta|\mathbf{y})$, can be obtained from samples from the joint distribution $p(\theta, \mathbf{y}) = p(\mathbf{y}|\theta)p(\theta)$ using, for instance, composition sampling.

The algorithms differ in the way they select samples $\{\theta^*, \mathbf{y}^*\}$ from the joint distribution to obtain a draw from the conditional distribution of interest: the rejection algorithm requires that there is an exact match between the observed data vector $\mathbf{y}$ and the generated data vector $\mathbf{y}^*$, whereas the SVE algorithm uses a Metropolis algorithm instead.

Marsman et al. (in press) show how the algorithms can be made suitable for large-scale applications, that is, applications where samples are required from not one, but many conditional distributions.

4. Bayesian Model Comparison

In a Bayesian framework, model assumptions and hypotheses can be tested using model comparison criteria. In this section, we briefly discuss how Bayes factors, posterior model probabilities, and the Deviance Information Criterion (DIC) can be used for testing measurement invariance (also known as differential item functioning), an important topic in educational testing (Millsap, 2011; Vandenberg and Lance, 2000). These Bayesian criteria are very flexible because they can be used for testing multiple nonnested hypotheses, which cannot be done using classical p-values. To illustrate the Bayesian model comparison approach, the one-parameter IRT model, which assumes equal item discriminations (i.e., $a_k = a_j$ for all k, j), will be considered. When taking into account the variability in item discriminations, the measurement invariance comparison procedure becomes more complicated, which is beyond the purpose of the present application.

Measurement invariance will be investigated for $K = 11$ geometry items of the College Basic Academic Subjects Examination (CBASE) for males ($j = 1$, $N_1 = 1,034$) and females ($j = 2$, $N_2 = 4,452$) (Millsap, 2011). CBASE is an exam intended for students enrolled in college, assessing knowledge and skills in mathematics, English, science, and social studies. A 1-PL BIRT model was specified where the group-specific item difficulty parameters, b_{kj}, for item $k = 1, \ldots, 11$, and group $j = 1$ or 2, are treated as fixed (e.g., De Boeck, 2008; Frederickx et al., 2010; Verhagen, 2012). An item k is measurement invariant when a male or female with the same latent trait θ has the same probability of a certain response (Mellenbergh, 1989; Millsap and Everson, 1993), that is, $b_{k1} = b_{k2}$.

First, we consider the following multiple hypothesis test $M_{1k} : b_{k1} = b_{k2}$ versus $M_{2k} : b_{k1} < b_{k2}$ versus $M_{3k} : b_{k1} > b_{k2}$ (i.e., "item k is measurement invariant" versus "item k is easier for males" versus "item k is easier for females," respectively), which will be tested for all $k = 1, \ldots, 11$ items. Bayes factors between the constrained models M_{1k}, M_{2k}, and M_{3k} can be computed relatively easily using the encompassing prior approach (Klugkist et al., 2005). To apply this methodology, a proper "encompassing prior" must be specified under an encompassing model, say, $M_e : (b_{k1}, b_{k2})' \in \mathbb{R}^2$, where $M_{tk} \subset M_e$, for $t = 1, 2, 3$ and $k = 1, \ldots, 11$. Two different encompassing priors will be considered to check prior sensitivity. First, a bivariate Student's t prior will be considered for $\mathbf{b}_k$, with mean vector $\mathbf{0}$, scale matrix $\mathbf{I}_2$, and degrees of freedom 1, that is, $p(\mathbf{b}_k|M_e) = t(\mathbf{0}, \mathbf{I}_2, 1)$, for $k = 1, \ldots, K$. Second, a bivariate normal prior will be considered for $\mathbf{b}_k$ with zero means and fixed identity covariance matrix $\mathbf{I}_2$. Note that both priors seem reasonable because item difficulty parameters will be 0 "on average" and standard deviations from zero not larger than 1 because the latent traits are assumed to be standard normally distributed.

Subsequently, the priors under the constrained models M_{tk} are proportional to the encompassing prior in their constrained spaces, that is, $p(b_k|M_{1k}) \propto p(\mathbf{b}_k = b_k\mathbf{1}|M_e)$, $p(\mathbf{b}_k|M_{2k}) = 2p(\mathbf{b}_k|M_e)I_{\{b_{k1}<b_{k2}\}}$, and $p(\mathbf{b}_k|M_{3k}) = 2p(\mathbf{b}_k|M_e)I_{\{b_{k1}>b_{k2}\}}$, where $I()$ is the indicator function (Mulder, 2014). When denoting $\delta_k = b_{k2} - b_{k1}$ to be the difference between the item difficulty in the groups, the Bayes factor between each constrained model against the encompassing model can then be expressed as

$$B(M_{1k}, M_e) = \frac{p(\delta_k = 0|\mathbf{y})}{p(\delta_k = 0)}, \tag{9}$$

$$B(M_{2k}, M_e) = \frac{Pr(\delta_k > 0|\mathbf{y})}{Pr(\delta_k > 0)}, \tag{10}$$

$$B(M_{3k}, M_e) = \frac{Pr(\delta_k < 0|\mathbf{y})}{Pr(\delta_k < 0)}, \tag{11}$$

using the prior and posterior functions under the encompassing model M_e. Note that $B(M_{1k}, M_e)$ corresponds to the Savage-Dickey density ratio (Dickey, 1971; Wetzels et al., 2010). Further note that $p(\delta_k = 0|M_e)$ have analytic expressions that can be derived from the Student's t and bivariate normal encompassing priors. The posterior density of δ_k at 0 can be estimated using a numerical estimate of the posterior density based on the S posterior draws of $\delta_k^{(s)} = b_{k2}^{(s)} - b_{k1}^{(s)}$ (e.g., using the function "logspline" in R). Furthermore, $\Pr(\delta_k < 0|M_e) = \Pr(\delta_k > 0|M_e) = .5$ and the corresponding posterior probabilities can be estimated as the proportion of posterior draws satisfying the constraints. Subsequently, Bayes factors can be computed between the constrained models using the transitive relation of the Bayes factor, for example, $B(M_{1k}, M_{2k}) = B(M_{1k}, M_e)/B(M_{2k}, M_e)$. For example, if $B(M_{1k}, M_{2k}) = 100$, this would imply that M_{1k} receives 100 times more support from the data than M_{2k}, which can be qualified as "strong" evidence for M_{1k} (Kass and Raftery, 1995).

For this multiple hypothesis test, it is easier to interpret posterior model probabilities (PMPs) instead of Bayes factors, because PMPs sum up to one. Equal prior model probabilities are chosen, which implies that all models are assumed to be equally likely a priori, that is, $P(M_{1k}) = P(M_{2k}) = P(M_{3k}) = \frac{1}{3}$. Posterior model odds can then be obtained by updating the prior model odds (which are equal to 1) with the observed Bayes factors according to

$$\frac{P(M_{2k}|\mathbf{y})}{P(M_{1k}|\mathbf{y})} = B(M_{2k}, M_{1k}) \times \frac{P(M_{2k})}{P(M_{1k})} = B(M_{2k}, M_{1k}). \tag{12}$$

Subsequently, for our three constrained models, the PMPs can be computed from the Bayes factors according to $P(M_{tk}|\mathbf{y}) = \frac{B(M_{tk}, M_{1k})}{1+B(M_{2k}, M_{1k})+B(M_{3k}, M_{1k})}$, for $t = 1$, 2, or 3.

The posterior model probabilities computed from the CBASE data using the two different prior choices for $\mathbf{b}_k$ can be found in Table 1. As can be seen the results are not very sensitive to the choice of the prior, and therefore we shall focus on the Student's t prior. Hence, there is positive evidence that items 2, 3, 4, and 9 are non-invariant ($P(M_{1k}|\mathbf{y}) > 0.75$, for $k = 2$, 3, 4, and 9); there is positive and very strong evidence that items 1 and 7, respectively, are easier for males ($P(M_{12}|\mathbf{y}) = 0.775$ and $P(M_{27}|\mathbf{y}) = 0.991$); and also strong evidence that item 6 is easier for females ($P(M_{36}|\mathbf{y}) = 0.976$). Finally, there is no clear evidence for or against invariance for items 5, 8, 10, 11 (i.e., all PMPs for these items are smaller than .75).

Another Bayesian criterion that can be used for Bayesian model comparison is the DIC. Unlike the Bayes factor, the DIC cannot be computed for different models based on the output of the encompassing model. Instead, each model must be separately fit to the data. Furthermore, the DIC is not recommendable for evaluating models with inequality constraints (Mulder et al., 2009). The DIC was computed for two models: $M_0 : b_{k1} = b_{k2} \in \mathbb{R}^1$, for all $k = 1, \ldots, K$ (all items are measurement invariant) and $M_1 : (b_{k1}, b_{k2})' \in \mathbb{R}^2$, for all $k = 1, \ldots, K$ (none of the items are invariant). The DICs can be computed relatively easy from the MCMC output. The DIC of a model M_t is defined as

$$\text{DIC}(M_t) = \overline{D(\boldsymbol{\xi}_t)} + d_t, \tag{13}$$

where $\boldsymbol{\xi}_t$ contains all free parameters under model M_t, $\overline{D(\boldsymbol{\xi}_t)}$ is a goodness-of-fit statistic,

Table 1
Posterior model probabilities (PMPs) of three models based on an encompassing prior for $\mathbf{b}_k$ with a bivariate Student's t distribution and a bivariate normal distribution, for items $k = 1, \ldots, 11$, obtained from the CBASE data using equal prior model probabilities. The largest PMP is typed in boldface

	Student t prior			Normal prior		
Item k	$P(M_{1k}\|\mathbf{y})$	$P(M_{2k}\|\mathbf{y})$	$P(M_{3k}\|\mathbf{y})$	$P(M_{1k}\|\mathbf{y})$	$P(M_{2k}\|\mathbf{y})$	$P(M_{3k}\|\mathbf{y})$
1	0.223	**0.775**	0.002	0.142	**0.856**	0.002
2	**0.802**	0.016	0.182	**0.743**	0.019	0.239
3	**0.839**	0.021	0.140	**0.781**	0.026	0.193
4	**0.905**	0.024	0.071	**0.879**	0.0340	0.087
5	**0.718**	0.016	0.266	**0.651**	0.020	0.329
6	0.024	0.000	**0.976**	0.029	0.000	**0.971**
7	0.001	**0.999**	0.000	0.009	**0.991**	0.000
8	**0.521**	0.468	0.010	0.438	**0.552**	0.010
9	**0.869**	0.032	0.099	**0.857**	0.032	0.111
10	**0.531**	0.009	0.459	0.451	0.012	**0.538**
11	0.346	**0.648**	0.006	0.284	**0.712**	0.004

$\overline{D(\boldsymbol{\xi}_t)} = E_{p(\boldsymbol{\xi}_t|\mathbf{y},M_t)}\{-2\log p(\mathbf{y}|\boldsymbol{\xi}_t, M_t)\}$, and $d_k = \overline{D(\boldsymbol{\xi}_t)} + 2\log p(\mathbf{y}|\hat{\boldsymbol{\xi}}_t, M_t)\}$ denotes the effective number of free parameters in M_t, where $\hat{\boldsymbol{\xi}}_t$ is the posterior mean of $\hat{\boldsymbol{\xi}}$. The expectation in the above expressions can be estimated using a Monte Carlo estimate: $\overline{D(\boldsymbol{\xi}_t)} \approx -\frac{2}{S}\sum_{s=1}^{S}\log p(\mathbf{y}|\boldsymbol{\xi}_t^{(s)}, M_t)$. The second part of d_k can be obtained by estimating $\hat{\boldsymbol{\xi}}$ as the arithmetic mean of the posterior draws $\boldsymbol{\xi}_t^{(s)}$, and plug this into $\log p(\mathbf{y}|\hat{\boldsymbol{\xi}}_t, M_t)$. Because $\overline{D(\boldsymbol{\xi}_t)}$ and d_t are computed from posterior draws of $\boldsymbol{\xi}$, the DIC is not very sensitive to the prior when using relatively vague priors for $\boldsymbol{\xi}$. This can be seen as an advantage of the DIC.

The DIC was computed using the WinBUGS program (Lunn et al., 2000). The DIC was computed for the two priors that were also used for computing Bayes factors. The Student's t prior yielded $\text{DIC}(M_1) = 63528.4$ and $\text{DIC}(M_0) = 63571.1$ and the normal prior yielded $\text{DIC}(M_1) = 63530.8$ and $\text{DIC}(M_0) = 63560.5$. The DICs for both priors favor M_1 over M_0, hence the model in which none of the items are invariant is preferred by this criterion.

5. Final Remarks

A general class of BIRT models is discussed, which represents flexible latent variable models to analyze item response data. Following a Bayesian approach, several modeling advantages can be given. Prior information can be easily included in the analysis, and more accurate inferences can be obtained for small sample sizes. The procedure of making posterior inferences can also be applied to latent variables, missing data, and other unknown quantities. Furthermore, the Bayesian approach comes with powerful simulation-based methods for estimation. It is shown that the two well-known sampling techniques (data-augmentation, Metropolis-type algorithms) can be used to estimate normal ogive and logistic IRT models, which cover a general class of IRT models.

The illustration concerning measurement invariance of CBASE geometry items shows how to test hypothesis using the DIC and Bayes factor. The Bayesian procedure for testing hypothesis can deal with multiple hypothesis, where MCMC can be used to compute posterior model probabilities, Bayes factors, and DICs.

These different aspects of the Bayesian approach can be very useful for different applications in educational assessment. Depending on the complexity of the assessment, the Bayesian approach is attractive and is shown to be widely applicable. Matteucci et al. (2012) discussed a Bayesian approach to computerized adaptive testing (CAT), where informative priors were used to improve the parameter estimates and the efficiency of the CAT. Fox et al. (2014) developed a Bayesian multivariate measurement modeling approach to measure student performance and feedback-seeking behavior. Beside student responses to an information literacy test, information was retrieved from students using an information retrieval system, which was developed to provide students easily relevant feedback.

References

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

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

Béguin, A. (2000). Robustness of Equating High-Stakes Tests. Unpublished Ph.D. dissertation, University of Twente, Enschede, The Netherlands.

Cho, S.-J., Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics* 35:336–370.

De Boeck, P. (2008). Random item IRT models. *Psychometrika* 73:533–559.

De Boeck, P., Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

De Jong, M. G., Steenkamp, J. B. E. M., Fox, J. -P. (2007). Relaxing cross-national measurement invariance using a hierarchical IRT model. *Journal of Consumer Research* 34:260–278.

Dickey, J. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Statistics* 42:204–223.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Fox, J.-P., Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66:271–288.

Fox, J.-P., Klein Entink, R., Timmer, C. (2014). The joint multivariate modeling of multiple mixed response sources: Relating student performances with feedback behavior. *Multivariate Behavioral Research* 49:54–66.

Fox, J.-P., Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In: Davidov, E., Schmidt, P., Billiet, J., eds. *Cross-Cultural Analysis: Methods and Applications*. London: Routledge Academic, pp. 467–488.

Frederickx, S., Tuerlinckx, F., De Boeck, P., Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement* 47:432–457.

He, Y., Wolf, R. E., Normand, S.-L. T. (2010). Assessing geographical variations in hospital processes of care using multilevel item response models. *Health Services and Outcomes Research Methodology* 10:111–133.

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

Klugkist, I., Laudy, O., Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods* 10:477–493.

Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 10:325–337.

Maris, G. (2012). Analyses. In: Jones, N., et al. eds., *First European Survey on Language Competences*, European Commission, pp. 298–331. Brussels, European Commission. Available at: http://ec.europa.eu/languages/eslc/index.html

Marsman, M. (2014). Plausible values in statistical inference. (Unpublished doctoral dissertation). University of Twente, Enschede The Netherlands.

Matteucci, M., Mignani, S., Veldkamp, B. P. (2012). Prior distributions for item parameters in IRT models. *Communications in Statistics, Theory and Methods* 41:2944–2958.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research* 13:127–143.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.

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

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis* 71:448–463.

Mulder, J., Klugkist, I., Meeus, W., van de Schoot, A., Selfhout, M., Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology* 53:530–546.

Murray, I., Ghahramani, Z., MacKay, D.J.C. (2006). MCMC for doubly-intractable distributions. In: Proc. 22nd A. Conf. Uncertainty in Artificial Intelligence (eds R. Dechter and T. S. Richardson), pp. 359–366. Cambridge: Association for Uncertainty in Artificial Intelligence Press.

Patz, R., Junker, B. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics* 24:146–178.

Polson, N., Scott, J., Windle, J. (2013). Bayesian inference for logistic models using pólya-gamma latent variables. *Journal of the American Statistical Association* 108:1339–1349.

Rubin, D. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 12:1151–1172.

Vandenberg, R. J., Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 3:4–70.

Van den Berg, S. M., Glas, C. A. W., Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics* 37:604–616.

Van der Linden, W. J., Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

Verhagen, A. J. (2012). *Bayesian Item Response Theory Models for Measurement Variance*. Ph.D. dissertation, University of Twente. Enschede, The Netherlands.

Verhagen, A. J., Fox, J.-P. (2013a). Bayesian tests of measurement invariance. *The British Journal of Mathematical and Statistical Psychology* 66:383–401.

Verhagen, A. J., Fox, J.-P. (2013b). Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine* 32:2988–3005.

Vermunt, J. K. (2003). Multilevel Latent Class Models. *Sociological Methodology* 33:213–239.

Wang, X., Berger, J. O., Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics* 7:126–153.

Wetzels, R., Grasman, R. P. P. P., Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis* 38:666–690.