Bayesian Psychometric Scaling

Jean-Paul Fox

Stéphanie van den Berg

Bernard Veldkamp

Abstract

In educational and psychological studies, psychometric methods are involved in the measurement of constructs, and in constructing and validating measurement instruments. Assessment results are typically used to measure student proficiency levels and test characteristics. Recently, Bayesian item response models received considerable attention to analyze test data and to measure latent variables. Bayesian psychometric modeling allows to include prior information about the assessment in addition to information available in the observed response data. An introduction is given to Bayesian psychometric modeling, and it is shown that this approach is very flexible, provides direct estimates of student proficiencies, and depends less on asymptotic results. Various Bayesian item response models are discussed to provide insight in Bayesian psychometric scaling and the Bayesian way of making psychometric inferences. This is done according to a general multilevel modeling approach, where observations are nested in students and items, and students are nested in schools. Different examples are given to illustrate the influence of prior information, the effects of clustered response data following a PISA study, and Bayesian methods for scale construction.

Keywords

Bayesian, Bayesian psychometric scaling, Bayesian scale construction, IRT, multilevel IRT, item selection, plausible values.

Introduction

Within the scope of psychometrics, measurement models are used to assess student proficiency levels and test characteristics given assessment results. The assessment results are most often response data, which are usually categorical observations. Therefore, a non-linear or a generalized linear modeling approach is necessary to define a relation between the item-response observations and the test and student characteristics.

Such item–based response models received much attention, where student proficiency is considered to be a latent variable. Multiple items are used to measure the latent variable, and a scale measurement is done using an item response (measurement) model. Within the framework of item response theory, response models are defined that describe the probabilistic relationship between a student's response and the student's latent variable level being measured.

Recently, Bayesian item response models have been developed to analyze test data and to measure latent variables. In Bayesian psychometric modeling, it is possible to include genuine prior information about the assessment in addition to information available in the observed response data. Inferences can be made based on the observed data and additional information, which can be important in non-standardized settings, complex experimental designs, or complex surveys, where common modeling assumptions do not hold. As prior information, uncertainties can be explicitly quantified which might relate to different models, parameters, or hypotheses.

The Bayesian modeling approach is very flexible and provides direct estimates of student proficiencies. In addition, by modeling the raw observations, a direct interpretation can be given of the model parameters and their priors. Furthermore, for small sample sizes, the flexible Bayesian approach depends less on asymptotic results and is therefore very useful for small samples (e.g., Lee and Song, 2004).

In this chapter, Bayesian psychometric modeling is introduced. It is shown that an item response modeling framework can be defined that can adapt to real case scenarios. Prior information can be incorporated in the analysis, which adds to the data information, and Bayesian inferences are based on the total information available. The Bayesian response modeling framework comes with powerful simulation-based estimation methods, which supports the joint estimation of all model parameters given the observed data. The work of Albert (1992), Albert and Chib (1993), Bradlow et al. (1999), Fox (2010), Fox and Glas (2001), Patz and Junker (1999a, 1999b), and Patz et al., (2002), among others, advocated simulation-based estimation methods for psychometric models and stimulated their use.

In this introduction, Bayesian item response models are discussed to provide insight in the various measurement models and the Bayesian way of making psychometric inferences. A hierarchical item response modeling approach is discussed, where observations are nested in subjects and items (e.g., Johnson et al., 2007). Priors are used to define the nested structure of the data by modeling item and student parameters in a hierarchical way. Dependencies between observations within the same cluster are taken into account by modeling the nested structure of the response data. The development of

Bayesian item response models started with the work of Mislevy (1986), Novick (1973), Swaminathan (1982, 1985), Tsutakawa (1986). In the 1990s, this approach was further developed by Kim et al., (1994) and Bradlow et al., (1999), among others.

After introducing unidimensional item response models for dichotomous responses, it is shown how to make inferences from posterior distributions. Then, specific attention will be given to the influence of priors on the statistical inferences. It is shown that informative priors will influence the posterior results more than uninformative priors, where the influence is estimated by considering the amount of shrinkage of the posterior estimate to the prior estimate. In estimating student proficiencies, different simulation studies are done to illustrate the prior influence.

Then, an additional clustering of students is discussed, where performances from students in the same cluster are more alike than from different clusters. This leads to an extension of the Bayesian item response model to more than two levels (i.e., item response defines level 1, and student level 2) is considered a multilevel IRT model. A classic example is educational survey data collected through multistage sampling, where the primary sampling units are schools, and students are sampled conditional on the school unit. Following Fox (2010), Fox and Glas (2001), and Aitkin and Aitkin (2011), among others, a complete hierarchical modeling framework is defined by integrating the item response model with the survey population distribution. Besides item-specific differences, this multilevel item response model takes the survey design into account, the backgrounds of the respondents, and clusters in which respondents are located.

Finally, advantages of Bayesian item response models in computer based testing are discussed. The introduction of empirical prior information in item parameter estimation reduces the costs of item bank development. Besides, in Bayesian computerized (adaptive) testing, test length can be reduced considerably when prior information about the ability level of the candidates is available. Methods to elicit empirical priors are presented, and an overview of Bayesian item selection is given.

Bayesian Item Response Models Using Latent Variables

In educational measurement, student proficiency cannot be measured directly with one observed variable. More observations are necessary to measure the construct. Usually a range of items are required to assess the construct. In item response theory and structural equation modeling, constructs are represented by latent variables, which are defined by observations from multiple test items.

Item response theory (IRT) models are very popular for modeling relationships among latent variables and observed variables. Basic IRT models can be found in van der Linden and Hambleton (2000), which provide a wide range of models for different standard test situations.

In common assessments, students are assessed and their performances are observed using a test. Each response observation is considered to be an outcome of a random response variable, denoted as Y_{ik}, where index *i* refers to the student and *k* to the item. The observations are influenced by item *k*'s characteristics and student *i's* proficiency level. The psychometric measurement of student proficiencies is usually done using an item response theory model. This model is based on a probabilistic relationship between the performance and the proficiency. A probabilistic relationship is defined to reflect uncertainty about the observed responses.

Several factors (e.g., subject, group, and item characteristics) influence the observed response. Student observations cannot be interpreted as fixed entities, since a high-proficiency student can make a mistake once in a while and a low-proficiency student may guess an item correctly. Furthermore, the limited number of assessment items can be used to measure students proficiencies up to some level of accuracy. And the uncertainty due to the limited number of items will be captured by the stochastic modeling of the item responses. The limited number of students in the assessment will provide information about the test characteristics but up to some level of accuracy.

The uncertainty induced by the sampling of students is modeled through a population distribution. This population distribution characterizes the distribution of students in the population from which students are sampled. In a straightforward way, students are sampled through simple random sampling, such students are assumed to be independently selected, where each member of the population has equal probability of being selected.

Although students are commonly assumed to be sampled through simple random sampling, observations are said to be nested within students. In practice, observations from one student are more alike than observations from different students. The response observations of each student are correlated but given the proficiency level, they are assumed to be conditionally independently distributed. The proficiency is explicitly modeled as a latent variable since it is the primary object of measurement and/or to model the within-student dependencies. The latent variable representing student performance is measured using multiple items.

Let θ_i define the proficiency level of student *i.* When considering items with two response options, each item has specific item characteristics, for item *k* denoted as a_k and b_k, representing the level of discrimination and difficulty. Then, the success probability of student *i* for item *k*, or the probability of endorsing item *k*, according to the two-parameter IRT model is given by,

$$P\left(Y_{ik}=1 \mid \theta_i, a_k, b_k\right)=\begin{cases} \dfrac{\exp\left(Da_k\left(\theta_i-b_k\right)\right)}{1+\exp\left(Da_k\left(\theta_i-b_k\right)\right)} & \text{Logistic Model} \\ \Phi\left(a_k\left(\theta_i-b_k\right)\right) & \text{Probit Model} \end{cases} \tag{1.1}$$

The observed value for random variable Y_{ik} is one and coded a success. A zero observed value would indicate a failure, which is modeled by one minus the success probability. The function $\Phi(.)$ denotes the cumulative normal distribution function. For the logistic model, when *D=1.7,* a metric close to the normal-ogive (Probit) model is specified. A typical special case is the Rasch model or one-parameter model, where al item discriminations are equal to one.

Prior Specification of Bayesian IRT Models

The observed responses are modeled at the lowest level (i.e., observational level, level 1). At this level, a likelihood function can quantify how well the model fits the data, given values for the student and item parameters. This likelihood component describes the distribution of the data given the lower level model parameters. In Bayesian modeling, the lower level model parameters are modeled using so-called prior distributions.

The prior distributions are used to specify information about the student and item parameters. This information is not based on the data modeled at level one, but typically include information about the parameter region, the survey design, relationships with other model parameters, and so forth. Any information about the assessment can be included and it can lead to more accurate inferences since the prior information simply adds to the data information.

Assume a simple random sampling design for the students, and let students be sampled from a normal distribution such that

$$\theta_i \sim N\left(\mu_\theta, \sigma_\theta^2\right), \tag{1.2}$$

where the population mean represents the average level and the population variance the variability between students. The normal distribution is a symmetric distribution, which means that students of above and below-average performance are expected to be selected with equal probability. The normal distribution does not have wide tails, which means that students with extreme proficiency levels are rare and unlikely to be sampled.

The test items are assumed to measure a single construct. A variety in item difficulties are necessary to measure accurately construct levels at different positions of the latent scale. The items are nested in the test and they reflect the subject matter of the test. Depending on the content of the test, items can show strong inter-item correlations. The test is supposed to cover the content of the domain to be assessed. Subsequently, test scores are generalized to the test domain by assuming that test items are a random sample from an item bank. The item bank is supposed to contain many items which cover a specific test domain. Therefore, a prior for the item parameters can be interpreted from a sampling perspective, where items are sampled from an item population. The correlation among items in the test can be modeled by a hierarchical prior, where the item characteristics are assumed to vary from the general test characteristics. In that case, item difficulty parameters are assumed to be normally distributed,

$$b_k \sim N\left(\mu_b, \sigma_b^2\right), \tag{1.3}$$

where μ_b and σ_b^2 are the average test difficulty and the variability in item difficulties in the test. A comparable prior can be defined for discrimination parameters, except that they are assumed to be positive such that a log-normal prior distribution is defined,

$$\log\left(a_k\right) \sim N\left(\mu_a, \sigma_a^2\right). \tag{1.4}$$

On the logarithmic scale, the average level of discrimination and the between-item variability in test discrimination is given by μ_a and σ_a^2, respectively.

The prior distributions in Equation (1.3) and (1.4) define random item parameters. The item parameters are no longer assumed to be fixed as in traditional (frequentist) IRT modeling. However, this sampling interpretation does not always hold, since it may not always be possible to sample items that cover a specific test domain. Only when the test domain is specifically defined a representative sample can be drawn. For example, to design a spelling test, a dictionary would cover the domain to sample from. The sampling perspective can be generalized to include item generation or item cloning, where each item generated is a draw (e.g., Geerlings, van der Linden, and Glas, 2011). In this case, each item can be generated or cloned and will show some variation in characteristics compared to the original test item.

Besides the sampling argument, random item parameters can also be defined from an uncertainty perspective, since the prior expresses the uncertainty in item characteristics before seeing the data. Variation in item difficulties is expected with respect to the general test difficulty, and this variation in item difficulties is specified through the variance term. Furthermore, the prior mean and variance parameter can be modeled themselves to express uncertainty about the specific test difficulty and the between-item variability. The modeling of prior parameters is done using hyper priors. More thorough discussion about random item parameters can be found in Glas and van der Linden (2003), Fox (2010), and De Boeck (2008).

Bayesian Parameter Estimation

Inferences are based on posterior distributions, which are constructed from the data (i.e., sampling distribution) and prior information (i.e., prior distribution). The posterior distribution contains all relevant information. To introduce this approach consider the latent variable θ_i, representing the proficiency level of student *i*. Assume that the item characteristics are known. The posterior distribution is derived from the data information according to an IRT model *M*, denoted as $p\left(\mathbf{Y}_i \mid \theta_i, M\right)$, and the prior distribution, denoted as $p\left(\theta_i; \mu_\theta, \sigma_\theta^2\right)$. According to Bayes' theorem, it follows that

$$\begin{aligned} p\left(\theta_i \mid \mathbf{Y}_i, M\right) &= \frac{p\left(\mathbf{Y}_i \mid \theta_i, M\right) p\left(\theta_i; \mu_\theta, \sigma_\theta^2\right)}{p\left(\mathbf{Y}_i \mid M\right)} \\ &\propto \underbrace{p\left(\mathbf{Y}_i \mid \theta_i, M\right)}_{\text{Likelihood}} \underbrace{p\left(\theta_i; \mu_\theta, \sigma_\theta^2\right)}_{\text{Prior}} \end{aligned} \qquad (1.5)$$

where the term $p\left(\mathbf{Y}_i \mid M\right)$ does not depend on the proficiency variable and can be treated as a constant. The posterior is proportional to the likelihood function times the prior distribution. The semicolon notation in the prior distribution states that the remaining parameters are fixed and known, instead of treating them as random and conditioning on some specific values. It follows that the sample information enters the posterior information through the likelihood, which only depends on the sample size.

When increasing the number of assessed items, the sample information will dominate the prior information, and the prior information will play a less important role in the estimation of the student proficiency. When the number of assessed items is small or moderate, the prior distribution becomes more important, and the sample information will play a less important role. In this case, the posterior distribution depends less on asymptotic theory and given accurate prior information reliable results can be obtained given small samples. Informative priors are required for moderate to small sample sizes, while non-informative priors are often used when there is sufficient data information. In practical settings, good prior information might be available from experts, similar analyses, or past data samples.

Posterior-based measurement of student proficiency

To give a simple example, consider a test, where all items have equal difficulty and discrimination characteristics, which are zero and one respectively. Interest is focused on the ability measurement of a student.

Let the probability of a correct response be specified by the Probit model for a student with ability parameter θ; that is, the success probability is modeled by $\Phi(\theta)$. When a response pattern of 10 items is observed with first six successes and then four failures, the posterior distribution of θ is given by,

$$p(\theta|\mathbf{Y}) \propto \underbrace{(\Phi(\theta))^6 (1-\Phi(\theta))^4}_{\text{Likelihood}} \underbrace{p(\theta)}_{\text{Prior}}, \tag{1.6}$$

where $p(\theta)$ is the prior. The maximum likelihood estimate is the parameter value that maximizes the likelihood[1], which is $\Phi(\hat{\theta}) = 6/10, \hat{\theta} \approx .25$.

Often a conjugate prior is used such that the posterior and prior distribution are of the same parametric form. In this example, consider the beta distribution, with hyperparameters α and β, as a prior for the success rate. With hyperparameters α and β the number of successes and failures, respectively, in a sample of $\alpha+\beta-2$ independent Bernoulli trials can be specified. The posterior distribution in Equation (1.6) can be expressed as

$$\begin{aligned} p(\theta|\mathbf{Y}) &\propto (\Phi(\theta))^6 (1-\Phi(\theta))^4 (\Phi(\theta))^{\alpha-1} (1-\Phi(\theta))^{\beta-1} \\ &\propto (\Phi(\theta))^{6+\alpha-1} (1-\Phi(\theta))^{4+\beta-1} \end{aligned} \tag{1.7}$$

The posterior distribution of the success rate is also a beta distribution. When maximizing the posterior, it follows in the same way that $\Phi(\hat{\theta}) = \dfrac{6+\alpha-1}{10+\alpha+\beta-2}$.

[1] When maximizing the log-likelihood, $\underset{\theta}{\text{Max}}\, 6\log\Phi(\theta) + 4\log(1-\Phi(\theta))$, set the first derivative equal to zero, $\dfrac{6}{\Phi(\hat{\theta})} - \dfrac{4}{1-\Phi(\hat{\theta})} = 0$. It follows that $\Phi(\hat{\theta}) = \dfrac{6}{10}$.

The hyperparameters are assumed to be equal ($\alpha=\beta$) when there is no reason to assume that a priori successes are more likely than failures. Furthermore, assume that this prior believe is as strong as the total data information (i.e., $2\alpha-2=10$). This means that the precision of the prior mode is equal to the precision of the maximum likelihood estimate. Then, $\Phi\left(\hat{\theta}\right)=11/20, \text{ and } \hat{\theta}\approx .13,$ which is much lower than the maximum likelihood estimate. The prior mode of parameter θ is zero, which corresponds with a prior success rate of .50. As a result, the posterior mode is shrunk towards the prior mode of zero. Since the amount of prior information is equal to the amount of data information, the posterior mode is located almost exactly between the prior mode of zero and the maximum likelihood estimate of .25. The posterior mode is simply the weighted average of two point estimates, and both estimates have equal precision.

A less informative prior can still assume that successes and failures are equally likely, but its prior mode has simply a lower precision compared to the data precision. The minimum informative beta prior consists of one success and one failure out of two independent Bernoulli trials such $\alpha=\beta=2,$ which is one fifth of the total amount of data. Subsequently, the posterior mode is $\Phi\left(\hat{\theta}\right)=7/12,\ \hat{\theta}\approx .21,$ and this posterior mode is shrunk towards zero with almost one fifth of the maximum likelihood estimate.

In most situations, the posterior distribution is not easily maximized and more parameters are involved. Integration is required to compute the posterior mode, but the integration does not have a closed form. When it is possible to draw samples from the posterior distribution, the posterior mean and variance can be approximated using the simulated values.

In simulation-based methods, latent variables are handled as missing data. By drawing values for the latent variables a complete data set is constructed, which is used to draw samples from the other posterior distributions. In particularly, Markov chain Monte Carlo (MCMC) methods can be used to obtain samples from the posterior distributions by drawing samples in iterative way from the full conditional posterior distributions.

For most Bayesian IRT models, directly simulating observations from the posterior distributions is difficult. The posterior distributions of the model parameters have an unknown complicated form and/or it is difficult to simulate from them. Tanner and Wong (1987) introduced the idea of data augmentation method, which greatly stimulated the use of posterior simulation methods.

This idea is illustrated using the example above. Assume that the student is sampled from a standard normal population distribution. This normal prior for the student's proficiency level is a non-conjugate prior, given the likelihood specified in Equation (1.6). Simulating directly from the posterior distribution is difficult, since this distribution has an unknown form. However, augmented data can be defined, which are assumed to be underlying normally distributed latent responses with a mean of θ and a variance of one, and restricted to be positive (negative) when the response is correct (incorrect).

The normal prior is a conjugate prior for the normally distributed latent responses, and subsequently, the posterior distribution of θ is normal. The mean and variance of the posterior distribution can be derived from properties of the normal distribution (e.g., Albert, 1992; Fox, 2010), and the posterior mean equals,

$$E\left(\theta \middle| Z, Y\right) = \frac{10\overline{Z}}{11} + \frac{\mu_\theta}{11}, \tag{1.8}$$

where $\overline{Z}$ is the average augmented latent response. In Appendix A, R-code is given to draw samples from the posterior distribution of the student proficiency parameter. The mean value of the sampled values will approximate the posterior mean of the posterior distribution, which can be used as an estimate of the student proficiency. The posterior mean in Equation (1.8) has a precision-weighted form, where the precision of the maximum likelihood estimate is 10 and the prior precision is 1. The posterior precision is the sum of the prior and the data precision.

In Figure 1 the posterior distribution of the student proficiency is plotted using simulated draws according to the R code in Appendix A. It can be seen that the posterior is shrunk towards the prior mean of zero, but the amount of shrinkage is small since the data precision is ten times higher than the prior precision. The posterior is slightly higher peaked than the likelihood since the precision of the posterior mean is the sum of the data and the prior precision.

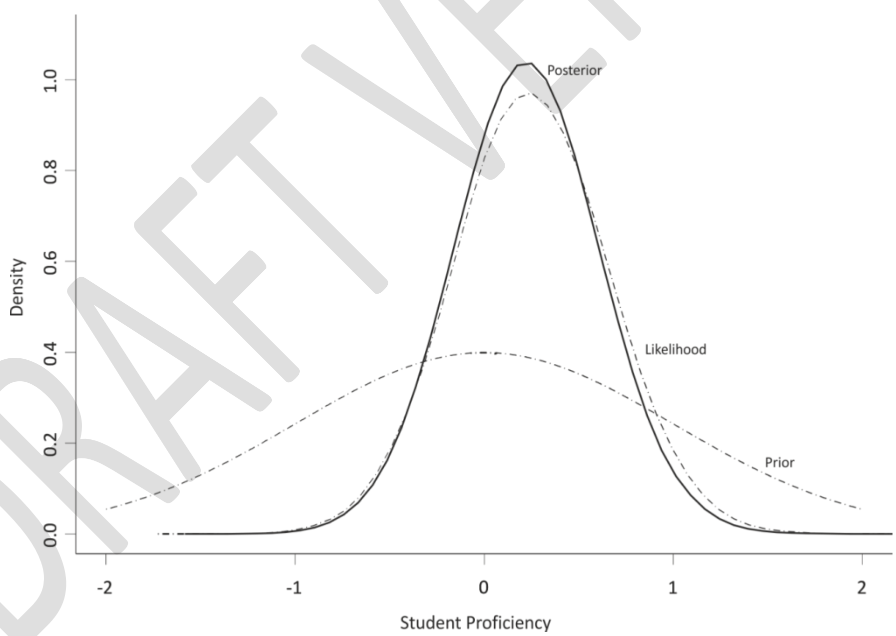

Figure 1: Prior, likelihood, and posterior distribution of student proficiency.

A simulation study using WinBUGS

For those who do not want to derive all full conditional posterior distributions and implement an MCMC method to sample from them, various statistical software programs facilitate simulation-based estimation techniques. A popular program is WinBUGS (Lunn et al., 2000), which supports the estimation of Bayesian models using MCMC techniques. The program can handle a wide variety of Bayesian IRT models. The program will generate simulated samples from the joint posterior, which can be used to estimate parameters, latent variables, and functions of them.

When dealing with multiple students and items with different item difficulties and discriminations, a more complex MCMC method is needed to draw samples from all posteriors. In Appendix A, Listing 1 gives the WinBUGS code for the Bayesian two-parameter IRT model as described in Equations (1.1) to (1.4).

Responses were generated, for 1,000 students and 10 items, according to a two-parameter logistic IRT model. For prior specification I, the precision of the prior distributions were not a priori specified but modeled using hyperpriors. As discussed in the example above, the prior precision influenced the amount of shrinkage of the posterior mean to the prior mean. By modeling the precision parameters of the prior distributions, these parameters are estimated using the data. For prior specification II, the parameters of the prior distributions were a priori fixed.

In Table 1: Posterior estimates of hyperparameters under different item priors, for Prior I, the estimated prior and hyperprior parameters are stated. It follows that the average discriminating level of the items is around .121 (on the logarithmic scale) and the between-item variability in discrimination is around .265. The average item difficulty is .200, and the variability in difficulty across items is estimated to be .438. For Prior II, the hyperprior parameters were fixed at specific values, as given in Table 1. As a result, the model with Prior II does not provide information about the posterior population item characteristics, since they are restricted by the prior to specific values.

To compare the fit of both models, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) were computed (Fox, 2010, pp. 57-61). The AIC and the BIC are given in Table 1. The model with Prior I fits the data better. Under Prior II, the discriminations parameters are shrunk towards the average discriminations and difficulty values due to their relatively high precision values. Prior II allows less variation in item characteristic across items. This leads to a less optimal fit of the model, since the AIC and BIC are both higher for the model with Prior II. Note that the AIC and BIC model selection indices usually perform well when data were generated using the one-parameter or two-parameter model.

Table 1: Posterior estimates of hyperparameters under different item priors

	Prior I		Prior II
	Mean	SD	Mean
Discrimination			
μ_a	.121	.165	.50
σ_a^2	.265	.142	.10
Difficulty			
μ_b	.200	.206	.00
σ_b^2	.438	.228	.10
Information Criterion			
BIC	9192	47	9283
AIC	9094	47	9185

The relevance of modeling the prior parameters becomes even more important when assuming that students might guess answers correctly. To account for randomly guessing, the three-parameter IRT model extends the two-parameter model by introducing a guessing or pseudo-chance parameter, which represents the probability that a student guesses the item correctly.
Let the three-parameter model be defined as

$$P\left(Y_{ik}=1 \middle| \theta_i, a_k, b_k, c_k\right)=\begin{cases} c_k+\left(1-c_k\right)\dfrac{\exp\left(Da_k\left(\theta_i-b_k\right)\right)}{1+\exp\left(Da_k\left(\theta_i-b_k\right)\right)} & \text{Logistic Model} \\ c_k+\left(1-c_k\right)\Phi\left(a_k\left(\theta_i-b_k\right)\right) & \text{Probit Model} \end{cases}$$

where c_k denotes the probability of guessing the item correctly.

In the Bayesian modeling approach, a prior distribution is required to define the prior information about random guessing behavior in the test. For a multiple choice item, it is reasonable to assume that the probability of guessing an item correctly is one divided by the number of response categories. This will not control for educated guesses, when one or more incorrect response options are easily identified. Other response formats can lead to more discussion about the appropriateness of the prior, since it will influence the estimate of the guessing parameter and the student's ability parameter. When overestimating the pseudo-chance parameter, student's abilities are underestimated since they obtain less credit for correctly scored items. Chiu and Camilli (2013) showed that better performing students, opposed to less-performing students, obtain more credit according to the three-parameter scoring rule. This difference in scores become larger when the guessing probability increases.

As for the other IRT parameters, the prior distribution for the guessing parameter influences the posterior estimates, where the prior parameters define the amount of shrinkage of the posterior mean to the prior mean. The amount of shrinkage can be severe, when the response data do not contain much information about the random guessing behavior. Consequently, student scores can be highly influenced by the prior information.

To illustrate this consider an artificial data set, generated using the three-parameter logistic model (3PL), of 500 students responding to ten items. The uncommented code of the 3PL in Listing 1 was used to estimate the parameters of the 3PL, using the same priors for the discrimination and difficulty parameters. For the guessing parameter, a beta distribution was defined with parameters α and β (i.e., b11 and b12 in Listing 1). The beta distribution restricts the guessing parameter to take values between zero and one. As in the example of modeling success rates, hyperparameters α and β define the number of successes (correctly guesses) and failures (incorrectly guesses), respectively, in a sample of $\alpha+\beta-2$ independent Bernoulli trials.

At a third level, the hyperparameters α and β were modeled according to a uniform distribution. Different boundary values of the uniform distribution were used to explore the effects of the hyperprior specification. For prior I, the hyperparameters were uniformly distributed between two and ten. The estimated average guessing probability was around .29 (3.30/(3.30+8.23)). The posterior mean estimates are given in Table 2. There was a moderate variation in guessing probability between items with a posterior standard deviation of .13. For Prior II, the hyperparameters were uniformly distributed between 2 and 500, which allowed the estimation of a much tighter prior for the guessing parameters compared to the restriction in Prior I. Using this prior, the average guessing probability was around .27 (122.20/(122.20+336.30)), and the variability in guessing across items almost zero.

It follows that the estimated beta prior parameters are very high under prior II. As a result the posterior information is highly peaked around .27. The more flexible uniform Prior II leads to much higher hyperparameter estimates. Under Prior II, the posterior information about guessing is much more centered than under Prior I. The data do not support between-item variability in guessing. The beta prior accumulates the evidence for guessing by shrinking all item guessing estimates to a general level of guessing. The AIC and BIC are based on the same likelihood and do not indicate that one model fits the data better.

Table 2: Posterior mean estimates of the hyperparameters for different guessing priors

	Prior I		Prior II	
	Mean	SD	Mean	SD
Discrimination				
μ_a	.63	.26	.66	.23
σ_a^2	.39	.25	.33	.18
Difficulty				
μ_b	.17	.32	.19	.33
σ_b^2	1.07	.60	1.11	.65
Guessing				
α	3.30	.89	122.20	42.91
β	8.23	1.37	336.30	97.24
Information Criterion				
BIC	5315	38	5317	37
AIC	5231	38	5233	37

Following the scoring rule of the 3PL model as stated in Chiu and Camilli (2013),

$$
\begin{aligned}
T_i &= \sum_{k=1}^{K} y_{ik} a_k \left(1+\lambda_{ik}^{-1}\right) \\
\lambda_{ik} &= c_k \exp\left(-Da_k\left(\theta_i - b_k\right)\right),
\end{aligned}
\tag{1.9}
$$

student ability-scores were estimated for the different priors. The sum of the logarithmic scores of students were computed under both prior specifications, since otherwise small sampled values of λ_{ik} led to numerical problems. In Figure 2, the sum scores are plotted against the difference in scores for all students. It follows that score differences become larger for higher scoring students. This difference can be around four points when making just one item incorrect. The differences are most negative, which means that under the more flexible Prior II students obtained relatively lower scores. In that situation guessing seemed to be more prominent, which led to an underestimate of student's performances.

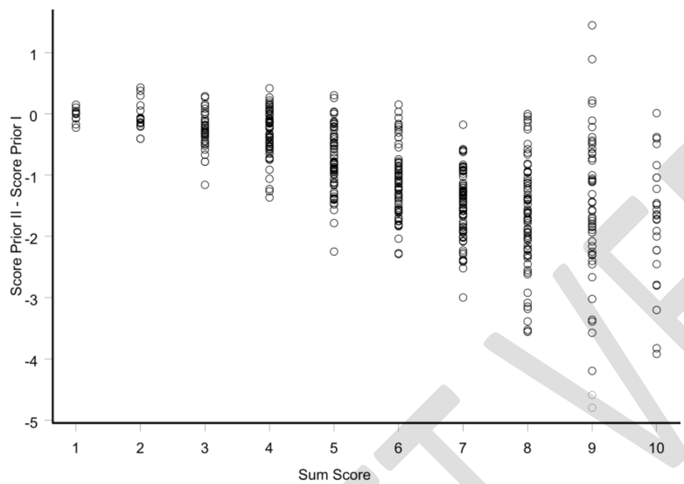

Figure 2: Differences between scores under different guessing priors versus the sum scores

Multistage Sampling Design: Clustering Students

In many data collection designs, individuals are sampled from different subgroups, for instance sampled from different schools. Such a multistage design leads to correlated observations, since students from the same school are more similar in their performances than students from different schools. A multilevel modeling approach is required to account for the (intraclass) correlation induced by the clustering of students in schools, where for example students from the same school have the same education program (Fox & Glas, 2001).

The clustering of students can be modeled as an extension of Equation (1.2), where instead of a simple random sampling design, a multistage sampling design applies. Therefore, consider *J* schools randomly sampled from a population of schools, after which students are randomly sampled within each school. Let θ_{ij} denote the latent variable of student *i* in school *j*, and Y_{ijk} the response observation of this student to item *k*. The two-parameter IRT model defines the relationship between the observation and the latent variable,

$$P\left(Y_{ijk}=1 \middle| \theta_{ij}, a_k, b_k\right)=\frac{\exp\left(Da_k\left(\theta_{ij}-b_k\right)\right)}{1+\exp\left(Da_k\left(\theta_{ij}-b_k\right)\right)}. \tag{1.10}$$

The population distribution of students is generalized by assuming that students are sampled at random given the school average,

$$\theta_{ij} \sim N\left(\mu_j, \sigma_j^2\right), \tag{1.11}$$

where μ_j represents the school mean and σ_j^2 the school-specific variance of the latent variable. Schools are sampled at random from a population with μ_θ the overall expected proficiency in the population, and σ_θ^2 the variability across school means; that is,

$$\mu_j \sim N\left(\mu_\theta, \sigma_\theta^2\right). \tag{1.12}$$

Equations (1.3) and (1.4) can be used as prior distributions for the item parameters. In order to identify the model parameters, it is sufficient to constrain μ_θ and σ_θ^2 to 0 and 1, respectively. When assuming that the residual variance is the same for each school, thus $\sigma_j^2 = \sigma^2$ for all schools *j*, the intraclass correlation is given by $\sigma_\theta^2/(\sigma_\theta^2 + \sigma^2)$, which represents the proportion of variance in the latent variable explained by the clustering of students in schools.

When the intraclass correlation is substantial, the question arises where this similarity comes from. Are teaching methods different across schools and can this explain the homogeneity in performances within a school? Or are certain background variables such as socio-economic status of the parents very similar within schools and can this explain the similarity?

To study these questions, the above multilevel structure is extended to include predictor variables that might explain the similarity within schools. Two types of predictors can be recognized, a variable that says something about the individual *i* in school *j*, like for example the gross income of the parents or the sex of the student, or that says something about the school *j*, for example the number of students or the teaching methods. Assume that both types of predictors are available, where ***X*** contain student and ***Z*** school explanatory information. Then, the above population model for students and schools, Equation (1.11) and (1.12), can be extended with linear effects at the student and school level, respectively, such that,

$$\begin{aligned} \theta_{ij} &\sim N\left(\mu_j + \mathbf{X}_{ij}^t \boldsymbol{\beta}, \sigma^2\right) \\ \mu_j &\sim N\left(\mu_\theta + \mathbf{Z}_j^t \boldsymbol{\gamma}, \sigma_\theta^2\right), \end{aligned} \tag{1.13}$$

where $\boldsymbol{\beta}$ is a vector with the regression coefficients at the individual level, and $\boldsymbol{\gamma}$ is a vector of regression coefficients at the school level. Both vectors of regression coefficients can be given multivariate normal prior distributions, or alternatively, independent identical normal priors.

An additional clustering of schools in countries leads to a three-level model. Following the multistage sampling design of large international educational surveys; item data are nested in students, that are nested in schools, which are in turn nested in countries. Below, this will be illustrated in an example concerning an international comparison of reading performances.

Plausible values

In large-scale international educational surveys, three-level IRT models can be used to compare student proficiency across countries, where item data is clustered within individuals, individuals within schools and schools within countries. Such nonlinear multilevel models are difficult to estimate, given the large amount of data, the large number of model parameters, the often complicated design of the individual tests, where not all students are administered the exact same questions, and the non-random nature of sampling of schools and individuals. In that case, it is often more convenient to work with plausible values using multiple imputation, rather than working with the raw item data (Rubin, 1987).

For example, for a large-scale PISA study (OECD, 2009) concerning reading proficiency, a multilevel model is specified with the latent reading proficiencies as outcomes. Plausible values for reading proficiency are defined as posterior samples, conditioning on the student's observed item data and a large number of covariates. The general idea is to take at least three different plausible values for each student. The multilevel analysis can be carried out for each set of plausible values, and then results are summarized. The formulas for multiple imputation in Rubin (1987) can be used to summarize results.

Multilevel IRT using plausible values

For this example, a three-level IRT model was applied to the 2009 PISA data on reading proficiency taking plausible values as provided by PISA (http://pisa2009.acer.edu.au). The data set consists of 13 countries, with a varying number of schools for each country, and in turn a varying number of students for each school. The model of interest had two predictors at the country level (age of first selection and number of school types), two predictors at the school level (autonomy regarding resources and autonomy regarding curriculum), and one predictor at the individual level (socio-economic status). Consider one plausible value for reading proficiency, called PV1, it follows that

$$PV1_{ijc} \sim N\left(E\left(PV1_{ijc}\right), \sigma^2\right), \tag{1.14}$$

with

$$\begin{aligned}
&E\left(PV1_{ijc}\right) = \mu + \mathbf{z}_c^t\boldsymbol{\gamma} + \mathbf{x}_{jc}^t\boldsymbol{\beta} + SES_{ijc}\alpha + r_c + e_{jc} \\
&r_c \sim N\left(0, \sigma^2_{res.country}\right) \\
&e_{jc} \sim N\left(0, \sigma^2_{res.school}\right),
\end{aligned}$$

where $PV1_{ijc}$ is a standardized plausible value for individual *i* in school *j* in country *c*, $\boldsymbol{z}_c$ is the vector of standardized country covariates, multiplied by regression coefficients $\boldsymbol{\gamma}$, $\boldsymbol{x}_{jc}$ is the vector of standardized school covariates, multiplied by school level regression parameters $\boldsymbol{\beta}$, and α is the regression coefficient for standardized socio-economic status (SES). Parameter r_c is the residual at the country level (i.e., a country effect that is not explained by covariates, be it at country, school or individual level) with variance $\sigma^2_{res.country}$, and e_{jc} the residual at the school level (i.e., a school effect not accounted for by the covariates in the model) with variance $\sigma^2_{res.school}$. As priors for the individual, school and country level residual variances, inverse gamma priors can be specified,

$$\sigma^2_{res.country} \sim InvGamma(.1,.1)$$
$$\sigma^2_{res.school} \sim InvGamma(.1,.1)$$
$$\sigma^2 \sim InvGamma(.1,.1)$$

For the remaining regression parameters a normal prior was specified with mean zero and variance ten, which seems reasonable since most predictor variables are standardized. The model was run using the WinBUGS script in Listing 2. In Table presents the means and standard deviations of the posterior distributions for each model parameter.

Table 3: Posterior mean estimates of multilevel IRT model parameters using plausible values.

	Mean	SD
m	0.09	0.03
s^2	0.52	0.00
$s^2_{res.school}$	0.43	0.01
$s^2_{res.country}$	0.08	0.04
α	0.15	0.00
b_1	-0.07	0.09
b_2	0.02	0.05
g_1	-0.07	0.06
g_2	-0.03	0.03

Given that the posterior mean for α is more than 2 standard deviations away from zero, it is concluded that there is a clear effect of socio-economic status on individual proficiency in reading. The regression parameters $\boldsymbol{\beta}$ are not clearly different from zero, it is concluded that autonomy in resources and curriculum do not explain variation in reading proficiency across schools. Also, the country level predictors do not explain much variance. Furthermore, school effects are clearly present within countries but there is a lot of unexplained variance at the individual level (PV1 was standardized, so 52% of the variance in individual differences is left unexplained). Note however, that these conclusions are based on only one set of plausible values.

Possible model extensions include the possibility of different regression coefficients across countries, for example different effects of SES across countries, and different school-level and individual-level variances across countries, for example more variation in school quality within a particular country. Note that in all such models all students belong to only one school, and all schools belong to only one country.

A further possible extension is to allow individuals to belong to multiple groups. For example, in genetic models, individuals can be correlated because they either share two parents (siblings) or only one parent (half-siblings): half-siblings are correlated because they belong to the same group of people that are the offspring of one particular parent (sharing on average 25% of the genetic variance, see Falconer & MacKay, 1996), and full siblings are correlated because they belong to two such separate groups (thus sharing on average 50% of the genetic variance). Such genetic IRT models for item data are available for twin data (Van den Berg, Glas & Boomsma, 2007) and for pedigree data (Van den Berg, Fikse, Arvelius, Glas & Strandberg, 2010).

Bayesian Scale Construction

In Bayesian Scale Construction (BSC), a Bayesian IRT model is applied in the process of constructing a scale, also referred to as a test. The starting point in BSC is a collection of items that have been pre-tested and for which posterior-based measurement has been applied to estimate the item parameters. All of these items and their item parameters are stored in an item bank. Especially in the area of educational measurement, large item banks have been developed. Item selection algorithms can be applied to construct a scale based on specifications like, for example, specifications related to measurement properties of the scale, to the content, or to the time available.

In a typical BSC problem, the goal is to select those items that maximize measurement precision, while a list of constraints related to various specifications of the scale have to be met. Various classes of BSC problems can be distinguished based on the formats of the scales. The first class of BSC problems is related to the paper-and-pencil (P&P) scales. Scales in this class have a fixed format. All respondents answer the same set of questions. Nowadays, these scales could be administered on a computer as well, but the basic format is still comparable to a P&P scale.

A second class, mainly used in the area of educational measurement, is related to multi-stage scales or tests. These tests consist of a number of stages. After completing a stage, the ability level of the respondent is estimated and the respondent is directed to an easier module, a more difficult module, or a module of comparable difficulty. Both the number of stages, and the number of modules might vary.

The third class of BSC problems are related to Computerized Adaptive Tests (CATs). The general procedure of CAT is the following. After administering an item, the ability of the respondent is estimated and the subsequent item is selected that is most informative at the estimated ability level. Respondents with high ability estimates get more difficult items, while respondents with low ability estimates get easier ones. The CAT stops after a fixed number of items or when a certain level of measurement precision for the ability estimate has been obtained. The main advantage of BSC is that collateral information about the respondents can be taken into account during test assembly. This advantage holds for all classes of BSC problems, but is most prominent in CAT.

Item selection in BSC

Van der Linden (2005) describes how BSC problems can be formulated as mathematical programming models. Mathematical programming models are general models for solving optimization problems. They have been applied in business and economics, but also for some engineering problems. Areas that use mathematical programming models include transportation, energy, telecommunications, and manufacturing. These models have proved to be useful in modeling diverse types of problems in planning, routing, scheduling, assignment, and design.

Theunissen (1985) was among the first to apply these models to scale construction (SC). Decision variables x_i can be introduced that denote whether an item is selected $(x_i = 1)$ or not $(x_i = 0)$. Test specifications can be modeled as either categorical, quantitative, or logical constraints, where categorical constraints are related to item attributes that classify items in various categories, quantitative constraints are about numerical attributes of items, and logical constraints are related to inter item relationships. A generic model for SC can be formulated as:

$$
\begin{aligned}
&\max \sum_{i=1}^{I} J_i(\theta) x_i \\
&\sum_{i \in c} x_i \le n_c \qquad c = 1,\ldots,C, \\
&\sum_{i=1}^{I} q_i x_i \le b_q \qquad q = 1,\ldots,Q, \\
&\sum_{i \in l} x_i \le 1 \qquad l = 1,\ldots,L, \\
&\sum_{i=1}^{I} x_i = n, \\
&x_i \in \{0,1\}.
\end{aligned}
\tag{1.15}
$$

where $J_i(\theta)$ denotes the contribution of item $i = 1, \ldots, I$, to the measurement precision of the test, *c* denotes a category, n_c the number of items for category *c*, q_i a numerical attribute of item *i*, b_q the bound of a numerical constraint *q*, *l* is an index for the various logical constraints, and *n* denotes the test length.

Computer programs, like CPLEX or LPSolve, can be used to generate tests that perform optimal with respect to the objective function and meet all the constraints. For the classes of P&P and the multi-stage SC problems, a single SC model has to be solved. For CAT problems, the shadow test approach (van der Linden & Reese, 1998) can be applied.

When Bayesian IRT is used to measure the ability parameters, the measurement precision is related to the posterior distribution of the ability parameter. In BSC, those items have to be selected that contribute most to the measurement precision. There are several ways to deal with the relationship between the posterior and the measurement precision.

Therefore, several item selection criteria have been proposed. Owen (1975) proposed to select items with a difficulty level closest to the estimated ability. Van der Linden (1998) introduced Maximum

Posterior Weighted Information, Maximum Expected Information, Minimum Expected Posterior Variance, and Maximum Expected Posterior Weighed Information as item selection criteria. Chang and Ying (1996) introduced the Maximum Posterior Weighted Kullback-Leibler Information criterion. This is not an exhaustive list, but all of these criteria have in common that they are posterior-based, where some of the employ Fisher Information and others are based on Kullback-Leibler information. Veldkamp (2010) describes how all of these criteria can be implemented in the shadow test approach. For example, for Owen's criterion, the model for the selection of the g-th item is given by

$$\begin{aligned} &\min \sum_{i=1}^{I} |b_i x_i - \hat{\theta}| \\ &\sum_{i \in V_{g-1}} x_i = g-1, \\ &\sum_{i=1}^{I} x_i = n, \\ &x_i \in \{0,1\}. \end{aligned} \tag{1.16}$$

where the set V_{g-1} denotes the items that have been selected in the previous (g-1) steps of CAT.

The posterior distribution of the latent variable contains information from the prior and the response data. When an uninformative prior distribution is used, the measurement precision solely depends on the response data. When an informative prior is used, the information from the response data is combined with prior beliefs. For some applications, like licensure exams, relying on prior beliefs might be undesirable. In other cases, there might be quite a strong argument in favor of incorporating prior beliefs. Imagine the case were a lot of collateral information about the respondent is available. This information could come from earlier tests of the same topic (as in progress testing), or from other subtests that correlate highly with the test at hand. Following Mislevy (1987) and Zwinderman (1991), Matteucci and Veldkamp (2012) elaborated the framework for dealing with collateral information.

Bayesian Dutch intelligence scale construction

The methodology of BSC was applied to a computerized adaptive Dutch intelligence scale (Maij – de Meij, et al, 2008). The scale consisted of three subscales (Number Series, Figure Series, and Matrices). First, the Matrices subtest is administered, after that the Number Series subtest. The correlation between the scores on the Number Series subtest and the Matrices subtest is equal to ρ=0.394. This information was used to elicit an empirical prior for the number series ability, based on the estimated ability for the Matrices subtest:

$$\theta_{NS} \sim N\left(-.243 + .394\hat{\theta}_M, .414\right) \tag{1.17}$$

where θ_{NS} and θ_M represent the latent scores on the Number Series and the Matrices subscales, respectively.

In order to demonstrate the attributed value of BSC, the use of this empirical prior was compared to the standard normal prior $\theta_{NS} \sim N(0,1)$. For this computerized adaptive scale, an item bank was available that consisted of 499 items calibrated with the Probit model described in Equation (1.1). The intelligence test is a variable length CAT where a stopping rule is formulated based on the measurement precision.

Based on the estimated abilities of 660 real candidates, more or less evenly distributed over the ability range, answer patterns to the variable length CAT were simulated and the person parameters were re-estimated using the Bayesian framework described above. The test length for various levels of the estimated ability is shown in Figure 3. For a more elaborate description of this example, See also Matteucci and Veldkamp (2012).

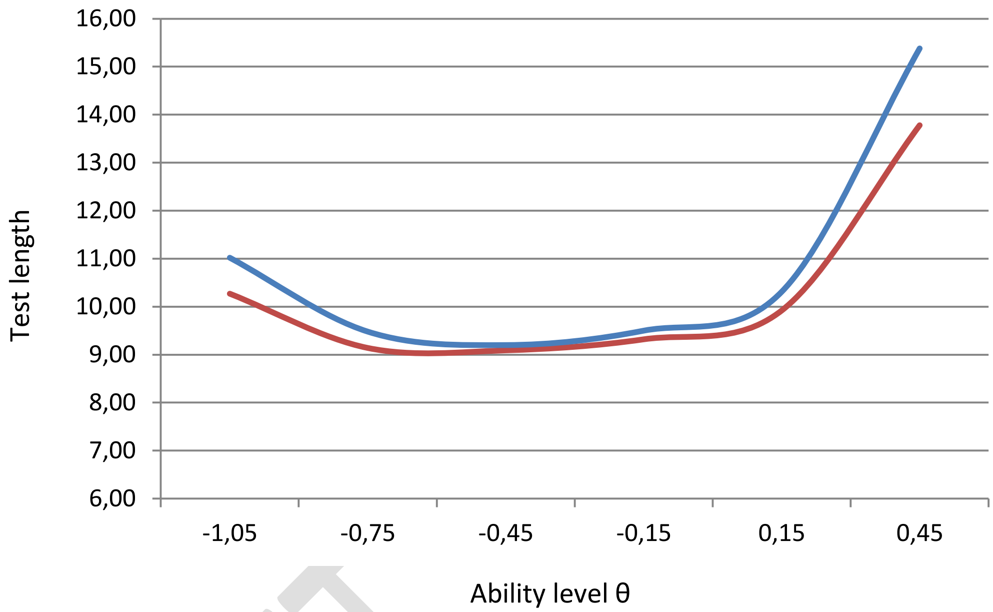

Figure 3: Test length; BSC with an informative empirical prior (red) and with a uninformative prior.

For those candidates with an ability level in the middle of the population, the informative prior resulted in slightly shorter tests, but the effect of adding collateral information was only small. For the candidates with the lowest and highest ability values however, considerable reduction in test length was obtained. In other words, empirical priors can be used successfully to reduce test length for respondents in the tails of the distribution, without any loss of measurement precision.

Here, the use of BSC was illustrated for ability estimation. Another application of BSC is related to the estimation of item parameters, see, for example, Matteucci, Mignani, & Veldkamp (2012).

Discussion

Bayesian psychometric modeling have received much attention with the introduction of simulation-based computational methods. With the introduction of these powerful computational methods, Bayesian IRT modeling became feasible, which made it possible to analyze much more complex models and to include prior information in the statistical analysis. Prior information can be used to quantify uncertainties concerning model parameters or hypotheses. Furthermore, especially in educational measurement, prior information can be useful when dealing with a non-standardized test setting, a complex experimental design, a complex survey, or any other measurement situation where common modeling assumptions do not hold.

The Bayesian approach in combination with sampling based estimation techniques is particularly powerful for hierarchically organized item response data. A multilevel modeling approach for item response data has been described, where item response observations are nested in items and students. Prior distributions are defined for the parameters of the distribution of the data. Subsequently, the parameters of these prior distributions can be described by hyper priors. Extending this multilevel modeling approach even further, the survey population distribution of students and/or items can be integrated into the item response model. Such a multilevel modeling framework takes the survey design into account, the backgrounds of the respondents, and clusters in which respondents are located. A classic example is discussed, where educational survey data are collected through multistage sampling (PISA, 2009), where the primary sampling units are schools, and students are sampled conditional on the school unit.

Advantages of Bayesian item response models have also been discussed in Bayesian scale construction. The use of empirical prior information in student and item parameter estimation can reduce the costs of item bank development. Furthermore, it is shown that the test length can be reduced considerably when prior information about the ability level of the candidates is available.

Appendix A.

Data Augmentation Scheme in R.

```
N <- 10              #sample size
Y <- sample(c(0,1), N, replace = T, prob = c(.5,.5)) #simulate response pattern
Z <- Y               #initialize vector of latent response data
mutheta <- 0         #prior mean value
XG <- 1000           #number of iterations
Mtheta <- matrix(0,ncol=2,nrow=XG) #store output
library(msm)         # use function rtnorm to sample from truncated normal distribution

for(ii in 1:XG){
 Z[Y==1] <- rtnorm(sum(Y), mean=theta, sd=1, lower=0, upper=Inf)          #sample Z>0 when Y=1
 Z[Y==0] <- rtnorm(sum(1-Y), mean=theta, sd=1, lower=-Inf, upper=0)       #sample Z<0 when Y=0

 #sample posterior values for θ given normal prior with mean mutheta
 theta1 <- rnorm(1,mean = sum(Z)/(N+1)+ mutheta/(N+1),sd=sqrt(1/(N + 1)))
 #sample posterior values for θ given uniform prior
 theta2 <- rnorm(1,mean = mean(Z),sd=sqrt(1/N))

 Mtheta[ii,1] <-theta1 # Average approximates posterior mean
 Mtheta[ii,2] <-theta2 # Average approximates posterior mean (and maximum likelihood estimate)
}
```

Listing 1: WinBUGS Code: Bayesian IRT for Binary Response Data

```
model{
   for (i in 1:N) { #Students
     for (k in 1:K) { #Items
       p[i,k] <- (exp(1.7*(a[k]*theta[i]-b[k]))/(1+exp(1.7*(a[k]*theta[i]-b[k])))) #Logistic 2PL IRT Model
       #p[i,k] <- c[k] + (1-c[k])*(exp(1.7*(a[k]*theta[i]-b[k]))/(1+exp(1.7*(a[k]*theta[i]-b[k])))) #Logistic 3PL IRT Model
       Y[i,k] ~ dbern(p[i,k])
     }
     theta[i] ~ dnorm(0,1) #Standard Normal Population Distribution for Students
  }

for (k in 1:K) {
 adummy[k] ~ dnorm(mu[2],prec[2]) #Lognormal distribution for Item Discriminations
 a[k] <- exp(adummy[k])
 b[k] ~ dnorm(mu[1],prec[1]) #Normal Distribution for Item Difficulties
#c[k] ~ dbeta(b11,b12) #Beta distribution for Guessing parameters
}

#Hyperprior specifications
 mu[1] ~ dnorm(0,1)
 mu[2] ~ dnorm(1,1)
 prec[1] ~ dgamma(1,1)
 prec[2] ~ dgamma(1,1)
 sigma[1] <- 1/prec[1]
 sigma[2] <- 1/prec[2]
##guessing hyperprior specification
#b11~dunif(2,10)
#b12~dunif(2,10)
#ec <- b11/(b11+b12)                              # posterior mean of average guessing in the test
#varc <- (b11*b12)/(pow(b11+b12,2)*(b11+b12+1))   # posterior variance of average guessing in the test
}
```

Listing 2: WinBUGS Code: Multilevel model for IRT-based plausible values

```
model

{

############################
##### description of variables
############################

# school covariates
# x variables:    x[country number, school number, variable number]
#1. Autonomy resources   (simple ratio) school level variable
#2. Autonomy curriculum  (simple ratio) school level variable

# country covariates (z variables)
#    1 = Standardized(normal): Age of first selection (system level)
#   2 = Standardized(normal): Number of school types (system level)


# student SES: SES.student[country, school, student]
# plausible value 1: PV1[country, school, student]


# K : number of countries with complete data on country covariates
# M : number of schools within country
# N : number of students  per [country,school]
# which.countries: a vector describing the countries with complete data on all country covariates.
# R: number of country covariates
# Q: number of school covariates


############################################################
### the actual modelling of the plausible values, ie the likelihood
############################################################

for (k in 1:K) # for every country
        {
        z.gamma[which.countries[k]]  <- inprod(  z[which.countries[k],] , gamma[])
        v[which.countries[k]] ~ dnorm(0,tau.country)
        for (j in 1:M[which.countries[k]]) # for every school in each country
                {
                x.beta[which.countries[k],j]   <- inprod( x[which.countries[k], j, ], beta[])
                u[which.countries[k],j]  ~  dnorm(0, tau.school)
                for (i in 1:N[which.countries[k],j]) # for every individual in each school in each country
                        {
                        PV1[which.countries[k],j,i]  ~ dnorm( expTheta[which.countries[k],j,i], tau.individual)
                        expTheta[which.countries[k],j,i] <- mu + alpha*SES.student[which.countries[k],j,i] +
x.beta[which.countries[k],j]  +  z.gamma[which.countries[k]]  +  u[which.countries[k],j]+ v[which.countries[k]]
                        }
                }
        }


#######################################################
## Priors for population mean and variances
#######################################################
```

```
mu ~ dnorm(0,.1) # normal prior for the population mean

tau.individual ~ dgamma(.1, .1)  # inverse gamma prior for residual variance at student level
var.individual <- 1/tau.individual

tau.school ~  dgamma(.1, .1)  # inverse gamma prior for residual variance at school level
var.school <- 1/tau.school

tau.country~ dgamma(.1,.1) # inverse gamma prior for residual variance at country level
var.country<- 1/tau.country


#####################################################
####priors for regression coefficients alpha, beta and gamma
#####################################################

# effect of individual covariate SES:
alpha ~dnorm(0,.1)

# effects of school covariates:
beta[1] ~ dnorm(0, .1)
beta[2] ~ dnorm(0, .1)

# effects of country covariates
gamma[1] ~ dnorm(0,.1)
gamma[2] ~ dnorm(0,.1)


}
```

References

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

Albert, James. H. and Siddharta Chib. 1993. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association, 88*, 669-679.

Bradlow, Eric T., Howard H. Wainer, and Xiaohui Wang. 1999. A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

De Boeck, Paul. 2008. Random item IRT models. *Psychometrika, 73*, 533-559. DOI: 10.1007/s11336-008-9002-x.

Chang, Hua-Hua, and Zhiliang Ying. (1996). Global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229. DOI: 10.1177/014662169602000303

Chiu, Ting-Wei, and Gregory Camilli. 2013. Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement, 37*. 76-86. DOI: 10.1177/0146621612459369.

Falconer, Douglas S., and Trudy F.C. MacKay. 1996. *Introduction to Quantitative Genetics*, Ed 4. Harlow, Essex, UK: Longman.

Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Methods.* New York: Springer.

Fox, Jean-Paul, and Cees A.W. Glas. 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 271-288.

Geerlings, Hanneke, Cees A.W. Glas, and Wim J. van der Linden. 2011. Modeling rule-based item generation. *Psychometrika, 76*, 337-359. DOI: 10.1007/S11336-011-9204-X.

Glas, Cees. A.W., and Wim J. van der Linden (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 247-261. DOI: 10.1177/0146621603027004001.

Johnson, Matthew S., Sandip Sinharay, and Eric P. Bradlow. 2007. Hierarchical Item Response Theory Models. In Rao, C.R, Sinharay, S. *Handbook of Statistics, Vol. 26: Psychometrics.* Amsterdam: Elsevier, p. 587-605.

Kim, Seock-Ho, Allan S. Cohen, Frank B. Baker, Michael J. Subkoviak, and Tom Leonard. 1994. An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59,* 405-421.

Lee, Sik-Yum, and Xin-Yuan Song. 2004. Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39,* 653-686.

Lord, Frederick M., and Melvin R. Novick. 1968. *Statistical theories of mental test scores, with contributions by Allan Birnbaum*. Reading, MA: Addison-Wesley.

Lunn, David .J., Andrew Thomas, Nicky Best, and David Spiegelhalter. 2000. WinBUGS -- a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10***,** 325--337.

Matteucci, Mariagiulia, Stephania Mignani, and Bernard P. Veldkamp. 2012. Prior distributions for item parameters in IRT models. *Communications in Statistics, Theory and Methods, 41,* 2944-2958. DOI: 10.1080/03610926.2011.639973

Matteucci, Mariagiulia, and Bernard P. Veldkamp (2012). The use of MCMC CAT with empirical prior information to improve the efficiency of CAT. *Statistical Methods and Applications.* In press.

Maij- de Meij, Annette M., Lolle Schakel, Nico Smid, N. Verstappen, A. Jaganjac. 2008. *Connector Ability; Professional Manual.* Utrecht, The Netherlands: PiCompany B.V.

Mislevy, Robert J. 1986. Bayes model estimation in item response models, *Psychometrika, 51*, 177—195.

Mislevy, Robert J. 1987. Exploiting Auxiliary Information About Examinees in the Estimation of Item Parameters. *Applied Psychological Measurement, 11,* 81-91.

Novick, Melvin R., Charles Lewis, and Paul H. Jackson. 1973. The estimation of proportions in *m* groups. *Psychometrika, 38,* 19—46.

OECD (2009). PISA 2009 Assessment Framework: Key Competencies in Reading,Mathematics and Science, Paris: OECD Publishing.

Owen, Roger J. 1975. A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association, 70,* 351-356.

Patz, Richard J., and Brian W. Junker. 1999a. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, Richard J., and Brian W. Junker. 1999b. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Patz, Richard. J., Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. 2002. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341—384.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley & Sons.

Sing, Xin-Yuan, and Sik-Yum Lee. 2012. A tutorial on the Bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology, 56,* 135-148. DOI:10:1016/j.jmp.2012.02.001

Swaminathan, Hariharan, and Janice A. Gifford. 1982. Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175—192.

Swaminathan, Hariharan, and Janice A. Gifford. 1985. Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349—364.

Tanner, Martin A., and Wing H. Wong. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528-550.

Theunissen, Theodorus. J. J. M. 1985. Binary programming and test design. Psychometrika, 50, 411-420.

Tsutakawa, R. K. and Lin, H. Y. 1986. Bayesian estimation of item response curves. *Psychometrika, 51*, 251—267.

Van den Berg, Stéphanie M., Cees A.W. Glas, and Dorret I. Boomsma (2007). Variance decomposition using an IRT measurement model, *Behavior Genetics, 37*, 604—616.

Van den Berg, Stéphanie M., Freddy Fikse, Per Arvelius, Cees A.W. Glas, and Erling Strandberg (2010). Integrating phenotypic measurement models with animal models. *Proceedings of the 9th World Congress on Genetics applied to Livestock Production.* Leipzig, Germany.

van der Linden, Wim. J. 1998. Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 201-216.

van der Linden, Wim. J. 2005*. Linear Models for Optimal Test Design.* New York: Springer.

van der Linden, Wim J., and Ronald K. Hambleton eds. 1997. *Handbook of Modern Item Response Theory.* New York: Springer.

van der Linden, Wim. J., & Lynda M. Reese. 1998. A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22,* 259-270. DOI: 10.1177/01466210022031570

van der Linden, Wim J., and Ronald K. Hambleton eds. 1997. *Handbook of Modern Item Response Theory.* New York: Springer.

Veldkamp, Bernard P. 2010. Bayesian item selection in constrained adaptive testing using shadow tests. *Psicologica, 31,* 149-169.

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

Biographical Note