Bayesian IRT guessing models for partial guessing behaviors

Jing Cao and S. Lynne Stokes

Department of Statistical Science, Southern Methodist University

Abstract

According to the recent Nation's Report Card, 12th-graders failed to produce gains on the 2005 National Assessment of Educational Progress (NAEP), while they are earning better grades on average. One possible explanation is 12th-graders were not motivated taking the NAEP which is a low-stakes test. We develop three Bayesian IRT mixture models to describe the results from a group of examinees including both non-guessers and partial guessers. The first assumes that the guesser answers questions based on his knowledge up to a certain test item, and guesses thereafter. The second model assumes that the guesser answers relatively easy questions based on his knowledge and guesses randomly on the remaining items. The third is constructed to describe more general low motivation behavior. It assumes that the guesser gives less and less effort as he proceeds through the test. The models can provide not only consistent estimates of IRT parameters but also estimates of each examinee's nonguesser/guesser status and degree of guessing behavior. Results of a simulation study comparing the performance of the three guessing models to the 2PL-IRT model are shown. Finally, an analysis of real data from a low-stakes test administered to university students is presented.

KEY WORDS: Bayesian mixture model, IRT model; guessing behavior; low motivation; item location; low-stakes test.

1 Introduction

Sometimes examinees taking a test may guess at the answers. This kind of test-taking behavior is especially prevalent in a low-stakes test, where students are asked to take a test for which they receive neither grades nor academic credit and thus may be unmotivated to do well. If guessing is not accounted for in estimation, standard IRT (item response theory) models will underestimate the true levels of the examinees' ability. In fact, according to the recent Nation's Report Card (Grigg et al., 2007), 12th-graders failed to produce gains on the 2005 National Assessment of Educational Progress (NAEP), while they are taking more advanced courses and earning better grades on average. One possible explanation is 12th-graders were not motivated taking the NAEP which is a low-stakes test. Wise and DeMars (2005) found that motivated examinees tend to outscore their unmotivated counterparts by over a half standard deviation on average. Guessing also occurs in high-stakes tests with time constraints. In that case, examinees may switch to random guessing when time is running out in an attempt to increase their scores.

It would be desirable to have a model to accommodate guessing behavior so that estimation of item parameters and examinee abilities would not be compromised. One common approach is to include a guessing component for all test items, as in the 3-parameter logistic (3PL) IRT model. The 3PL model assumes that the examinee has a chance to answer each item correctly either from his own knowledge or, if he doesn't have the knowledge, by guessing. The model thus describes guessing behavior as an item property that applies to all the examinees. A more realistic model would allow individual differences among examinees' guessing strategies. Martin, del Pino, and De Boeck (2006) do this by extending the 3PL model to let the guessing parameter depend on the ability of the examinee. But both models assume that all the examinees have the same 'knowledge-plus-guessing' strategy on all the test items. A consequence is that the presence of such guessing behavior increases each examinee's probability of answering each item correctly, so the 3PL-IRT curve (the correct-response rate) could be higher than the nonguessing 2PL-IRT model. In reality, especially in a low-stakes test, unmotivated examinees may randomly guess the answers without trying to think it over. In this case, the guessing IRT curve is a horizontal line mostly below the

nonguessing IRT curve.

Wise and Kong (2005) propose to use response time to distinguish solution behavior and rapid-guessing behavior. They suggest that for each item, there is a threshold which is the response time boundary between solution behavior and rapid-guessing behavior. Based on this hypothesis, Wise and DeMars (2006) developed the effort-moderated model. If an examinee's response time is longer than the threshold, the model reduces to the 3PL IRT model describing solution behavior. Otherwise, the model reduces to a constant probability model with the guessing probability being the reciprocal of the number of response options. This model has a more realistic assumption that examinees include both guessers and nonguessers. The authors plot the response time distribution, and if there are two modes, they visually choose the threshold between the two modes. However, if the distribution is unimodal, the strategy is hard to apply. Another concern is that the randomness of the visual determination is not accommodated by the model. The effort-moderated model requires the response time on each item. It is tailored for computer-based exams. It won't apply to paper-based exams where response time is not available.

The effect of random guessing can be accommodated in the framework of the mixture linear logistic test model (Mislevy and Verhelst, 1990), which is a special case of the IRT model with a parameter-driven process for change (Rijmen et al., 2005). Those two approaches model the different solution strategies employed in a test. They use the marginal maximum likelihood estimates based on the EM algorithm. In this paper, we will make more specific assumptions about guessing behaviors and propose a full Bayesian estimation procedure.

The above models need only binary data (correct or incorrect) for estimation. Another guessing model for multiple choice data is the Nedelsky model (Bechger et al., 2003), which requires the additional information of which option is selected for each item by each examinee. It is based upon the idea that a person responds to a multiple choice question by first eliminating the incorrect answers (distractors) he recognizes as wrong and then guessing at random from the remaining answers. The model is hierarchical with the first level being a Bernoulli trial describing the random guess. At the second level, the probability that a wrong answer is recognized as wrong is modeled by the 2PL-IRT model. Because of the nested logistic structure, the model requires a very large sample size to get reliable estimates.

The IRT guessing models proposed in this paper were motivated by the response pattern we observed in a test administered to our classes when the students participated in a national statistics literacy assessment study. The multiple choice test was administered to 265 students and consisted of 40 items, which took 30 to 45 minutes to complete. Because the test did not affect the course grade, it was a low-stakes test. The students could choose to leave the items blank, so there were nonresponses. Figure 1 shows the number of nonresponses as a function of item location. The distinctive feature of the data is that the nearer the item was located to the end of the test, the greater the nonresponse rate. In most cases, once there was a nonresponse, none of the subsequent items were answered neither. This type of behavior seems reasonable for students who may be curious or motivated by the test at first. Gradually, they lose interest and begin guessing because their performance will not have any consequence.

We propose three IRT models to accommodate different guessing behavior. One model assumes that some examinees answer questions based on their knowledge up to a certain test item, and guess randomly thereafter. For this model, there is an item location threshold for each examinee, specifying the item number at which guessing commences. The second model assumes that some examinees answer relatively easy questions based on their knowledge and guess randomly on the rest, regardless of their location within the test. This model was motivated by our attempts to describe the behavior of our students who skipped difficult problems and attempted easier ones later in the statistical literacy assessment. The third model is constructed to accommodate behavior indicative of low motivation, which we consider to be a generalization of guessing behavior. It assumes that the unmotivated examinees (guessers) give less effort to answer the problems as the exam progresses. For convenience, we refer to this model as a guessing model also. All three can be thought of as mixture models, where one component of the mixture model is non-guessers and the remainder describe various degrees of guessing. Thus our models do not require all examinees to behave the same with respect to their guessing behavior. The three models can be estimated using binary (correct/incorrect) data only.

We use Bayesian methods for estimating our models. In Section 2, we present the three guessing models, discuss the choice of priors and hyperparameters, and describe how the

models can be fit via Gibbs sampling. In Section 3, we present results from a simulation study designed to compare the three guessing models with the non-guessing 2PL-IRT model. A method for model selection is proposed. In Section 4, our method is applied to the statistics literacy assessment test data. Some discussion follows in Section 5.

2 Three Bayesian IRT guessing models

In this section, we propose three Bayesian IRT guessing models. All of them can be considered as mixture models extended from the 2PL-IRT model. Define the binary response data, x_{ij}, with index $i = 1, \cdots, n$ for persons, and index $j = 1, \cdots, J$ for items, and

$$x_{ij} = \begin{cases} 1, & \text{if the response from person } i \text{ to item } j \text{ is correct,} \\ 0, & \text{otherwise.} \end{cases}$$

In the 2PL-IRT model, the probability of a correct response from examinee i to item j is

$$P(x_{ij} = 1 | \, \theta_i, \delta_j, \gamma_j) = \frac{\exp(\gamma_j(\theta_i - \delta_j))}{1 + \exp(\gamma_j(\theta_i - \delta_j))},$$

where θ_i is examinee i's ability parameter, δ_j is item j's difficulty parameter and γ_j is item j's discrimination parameter.

2.1 The IRT threshold guessing model

The IRT threshold guessing model (IRT-TG) is constructed under the assumption that both guessers and nonguessers take the test, where the guessers answer questions based on their knowledge up to a certain test item, and guess randomly thereafter. Our model includes an item location parameter that specifies the threshold individually for each examinee. Then the probability of a correct response from examinee i to item j is given by

$$P(x_{ij} = 1 | \, \theta_i, \delta_j, \gamma_j, c_j, \alpha_i) = \frac{\exp[\gamma_j(\theta_i - \delta_j) - I(j > \alpha_i)(\gamma_j(\theta_i - \delta_j) - c_j)]}{1 + \exp[\gamma_j(\theta_i - \delta_j) - I(j > \alpha_i)(\gamma_j(\theta_i - \delta_j) - c_j)]}, \tag{1}$$

where α_i is the ith examinee's item location threshold parameter, and c_j is the jth item's guessing parameter. Parameter α_i can be any integer from 1 to J. Indicator function

$I(j > \alpha_i)$ is defined as

$$I(j > \alpha_i) \;\; = \;\; \begin{cases} 1, & j > \alpha_i \\ 0, & j \leq \alpha_i. \end{cases}$$

The IRT-TG model can be partitioned into two parts. It is assumed that examinee i is motivated and actively seeks the answers based on his ability for the first α_i items. So the model reduces to the 2PL IRT model when $j \leq \alpha_i$,

$$P(x_{ij} = 1 \,|\, \theta_i, \delta_j, \gamma_j, c_j, j \leq \alpha_i) = \frac{\exp(\gamma_j(\theta_i - \delta_j))}{1 + \exp(\gamma_j(\theta_i - \delta_j))}.$$

Note that if $\alpha_i = J$, then examinee i has answered all the items actively and is not a guesser. After item α_i, examinee i responds to the subsequent items by random guessing, and the model has the form

$$P(x_{ij} = 1 \,|\, \theta_i, \delta_j, \gamma_j, c_j, j > \alpha_i) = \frac{\exp(c_j)}{1 + \exp(c_j)},$$

for $j > \alpha_i$. Thus the parameter c_j determines the probability of a correct response when an examinee guesses at the item. Define the guessing probability as $g_j = \frac{\exp(c_j)}{1+\exp(c_j)}$. If it is completely random guessing, then $g_j = \frac{1}{N_j}$, where N_j is the number of options for item j. In this paper, we assume that g_j is an unknown parameter so that the model is more flexible.

The IRT-TG model is the same as the HYBRID model proposed by Yamamoto (1995), which was motivated by the behavior of examinees on speeded tests, such as TOEFL or GRE. He noted that they would "switch from a strategy of thoughtful response to a strategy of patterned or random response" (Yamamoto, 1995). Yamamoto implemented the marginal maximum likelihood method to estimate the parameters of this model, while our estimation method is Bayesian. The advantage of our method is that it provides estimates of some additional parameters, such as the probability that each examinee is a guesser. It is also easy to implement using WinBUGS, a free Bayesian estimation software.

Next, we specify priors for the parameters of our model. We assume that the J item difficulty parameters are independent, as are the n examinee ability parameters. We assign each a two-stage normal prior,

$$\begin{aligned} \theta_i &\sim \mathrm{N}(0, \tau_\theta), \qquad i = 1, \cdots, n, \\ \delta_j &\sim \mathrm{N}(0, \tau_\delta), \qquad j = 1, \cdots, J. \end{aligned}$$

where τ_θ and τ_δ both follow the conjugate inverse gamma prior,

$$\begin{aligned} \tau_\theta &\sim \text{IG}(a_\theta, b_\theta), \\ \tau_\delta &\sim \text{IG}(a_\delta, b_\delta), \end{aligned}$$

where a_θ and b_θ, a_δ and b_δ are fixed values. For the computations in this paper, the hyperparameters are assigned to produce vague priors. From Berger (1985), Bayesian estimators are often robust to changes of hyperparameters when noninformative or vague priors are used. We let $a_\theta = a_\delta = 2$ and $b_\theta = b_\delta = 1$, which means the priors for τ_θ and τ_δ each have an infinite variance.

Discrimination parameter γ_j is positive, we assume a gamma prior $\text{G}(a_\gamma, b_\gamma)$, where we choose $a_\gamma = b_\gamma = 1$. Parameter g_j is a one-to-one transformation of parameter c_j. It is more convenient to update g_j in the MCMC. We assume that g_j has a uniform prior $U(0, 0.5)$. We use this prior to include "distracted guessing", where the distractors are more likely to be chosen and g_j is relatively low, and "informed guessing", where examinees still make some effort trying to select the right choice and g_j is relatively high.

The item location parameter α_i can take any integer from 1 to J, and it follows a discrete distribution with probability $(p_1, p_2, \cdots, p_J)$ where $p_j = P(\alpha_i = j)$. A natural choice of prior for $(p_1, p_2, \cdots, p_{J-1}, p_J)$ is a Dirichlet distribution with J hyperparameters. In a typical IRT data analysis, the parameter-to-data ratio is usually large. To reduce the number of parameters, we propose a different prior. Recall that examinee i is identified as a nonguesser when $\alpha_i = J$, so p_J is the probability that an examinee is a nonguesser. Probability p_j $(j < J)$ is the probability that a guesser switches to random guessing after item j. We specify the priors for these two parts of the vector separately. We first specify a beta prior for p_J,

$$p_J \sim \text{Beta}(b_1, b_2),$$

where b_1 and b_2 are constants. We set $b_1 = b_2 = 1$, so that it is a uniform prior.

The nonresponse pattern in Figure 1 shows the shape of the cumulative probability function (cpf) of α_i for our data. It is reasonable to assume that the cpf curve of α_i is a smooth increasing curve. This is because though examinees switch to random guessing at

different item locations, it occurs gradually. Thus we assume that

$$P(\alpha_i \leq j \,|\, j < J) \;=\; \frac{j^\omega}{(J-1)^\omega}, \qquad j = 1, \cdots, J-1, \tag{2}$$

where the hyperparameter ω is positive but unknown. This curve is flexible in the sense that when ω is less than one, the cpf is a concave increasing curve; when ω equals one, the cpf is a linear increasing curve; and when ω is greater than one, the cpf is a convex increasing curve. Based on the cpf,

$$p_j \;=\; \frac{j^\omega - (j-1)^\omega}{(J-1)^\omega}(1 - p_J), \qquad j = 1, \cdots, J-1. \tag{3}$$

To complete our model, we assume

$$\omega \;\sim\; \text{Gamma}(a_\omega, b_\omega),$$

where a_ω and b_ω are fixed values. In our implementation, we set $a_\omega = b_\omega = 1.0$.

In Bayesian computation, an unknown parameter is often estimated by its posterior mean. We use the Gibbs sampler to get samples for an individual parameter from its full conditional distribution. The average of these samples yields the posterior mean. Because the item location parameter α_i can only take integer values, and more importantly, the posterior distribution is skewed to the left for the nonguessers, we use the posterior median instead of the posterior mean for α_i. The full conditional distributions in the model, listed in the Appendix, either have closed forms which can be sampled from standard distributions, or are log-concave which can be efficiently updated by the adaptive rejection sampling method (Gilks and Wild, 1992).

The full conditional distributions for examinee ability parameters θ_i and item parameters δ_j and γ_j depend only on the responses up to the item threshold location α_i. That is, it is the responses based on solution behavior that contribute to the estimation of those parameters. If we fit the 2PL-IRT model with the guessing data, the low probability of correct response resulting from random guessing usually will underestimate guesser's ability and overestimate item's difficulty. On the other hand, if the guesser's ability is extremely low or the item is extremely difficult, the probability of correct response based on the solution behavior is even lower than that of guessing behavior. In this case, the guesser's ability could be overestimated and the item's difficulty could be underestimated.

2.2 The IRT difficulty-based guessing model

The IRT difficulty-based guessing model (IRT-DG) is constructed under the assumption that both guessers and nonguessers take the test, where guessers answer only the relatively easy test items and guess on the remainder. This model is based on the theory of test-taking motivation described by Wise and DeMars (2005). They suggest that the amount of effort an examinee will expend to answer an item in a low-stakes test decreases as the task becomes more difficult. Our model assumes that a guesser resorts to guessing behavior for items that are difficult for them, where that difficulty threshold is related to the examinee's own ability parameter. Thus we assume that the probability of a correct response from examinee i to item j is given by

$$P(x_{ij} = 1 | \, \theta_i, \delta_j, \gamma_j, \eta, c_j) = \frac{\exp[\gamma_j(\theta_i - \delta_j) - \beta_i I(\delta_j - \theta_i - \eta)(\gamma_j(\theta_i - \delta_j) - c_j)]}{1 + \exp[\gamma_j(\theta_i - \delta_j) - \beta_i I(\delta_j - \theta_i - \eta)(\gamma_j(\theta_i - \delta_j) - c_j)]}, \qquad (4)$$

where $\beta_i = 1$ if person i is a guesser and 0 otherwise, and η is a parameter that measures the difficulty differential threshold that would entice a guesser to guess. If the relative difficulty of item j $(\delta_j - \theta_i)$ is higher than η, the guesser will take a random guess on the item. Otherwise, the answer to the item reflects the student's true ability. The remaining parameters are defined as in (1).

Unlike the "switch-only-once" strategy in the IRT-TG model, the IRT-DG model allows that the guessers can switch multiple times between solution behavior and guessing behavior in the test. Thus, the IRT-DG guessing pattern is more difficult to detect. To reduce the burden of estimation, we assume that $g_j = \frac{\exp(c_j)}{1+\exp(c_j)}$ is known and set it to be $1/N_j$, where N_j is the number of item choices. However, if the test is relatively long and sample size is large, we can use the same prior specification on g_j as that in the IRT-TG model.

The priors for parameters θ_i's, δ_j's, γ_j's, τ_θ, and τ_δ are the same as those specified in the IRT-TG model. As for the indicator β_i, we assign a Bernoulli prior with hyperparameter p_β. The probability p_β is assumed to have the conjugate beta distribution,

$$p_\beta \sim \text{Beta}(a_p, b_p).$$

In our implementation, we set $a_p = b_p = 1$. As for the parameter η, we assume a noninfor-

mative normal prior,

$$\begin{aligned} \eta &\sim \mathrm{N}(0, \tau_\eta), \\ \tau_\eta &\sim \mathrm{IG}(a_\eta, b_\eta), \end{aligned}$$

where a_η and b_η are fixed values. We set $a_\eta = 2$ and $b_\eta = 1$ to have an inverse gamma prior with an infinite variance.

This prior specification is simple. However, because of the indicator function $I(\delta_j - \theta_i - \eta)$, the log-concavity of the full conditionals on δ_j, θ_i, and η does not hold. Note that their full conditional expressions (see the Appendix) consist of two functions, one is a standard density easy to sample and the other is a complex function difficult to sample. We use the slice sampling (Neal, 2003) to update these parameters. Slice sampling has an advantage over the Metropolis-Hastings algorithm in that it always samples from the exact full conditional distribution. The other full conditional distributions have closed forms which can be sampled from standard distributions.

2.3 The IRT continuous guessing model

Random guessing is a special case of low motivation. Some examinees with low motivation may try to answer the test items, but use less effort than their motivated counterparts and are thus less likely to answer items correctly. Failure to accommodate low motivation can result in biased estimates of ability and item parameters. Wise and DeMars (2005) suggest a variety of methods to mitigate this problem. They suggest several methods that statistically adjust scores by using a measure of motivation obtained from each examinee via a questionnaire. One approach suggested is that this motivation measure be used as an independent variable within an IRT model. Another approach, dubbed motivation filtering, is to use data in estimation from only those examinees whose motivation score is sufficiently high.

Our third model is meant to provide an alternative method that does not require the motivation measure to statistically adjust estimates of ability and item parameters. We refer to our model as the IRT continuous guessing model (IRT-CG). It assumes that there

are motivated examinees and unmotivated examinees taking the test, where the motivated examinees answer all items using their knowledge and the unmotivated examinees expend less effort as the test progresses so that their probability of answering items correctly decreases gradually over the course of the test. If the model fits well, the estimates of ability and item parameters it provides will have smaller bias than those estimated from the 2PL-IRT model.

For convenience, we do not distinguish the IRT-CG model as a nonguessing model and we call the unmotivated examinees guessers. Actually, the three models can be described as partial guessing models. The term partial guessing has three meanings. One is that only some examinees exhibit guessing behavior. The second is that the guessers may actively answer some test items with their knowledge and guess on the rest. The third is that the unmotivated examinees may still try a bit with some effort, but not as much as they would if it were a test for a course grade.

We assume that the probability of a correct response from examinee i to item j in the IRT-CG model is given by

$$P(x_{ij} = 1 \,|\, \theta_i, \delta_j, \gamma_j, \beta_i, \phi_j) = \frac{\exp(\gamma_j(\theta_i - \delta_j - \beta_i\phi_j))}{1 + \exp(\gamma_j(\theta_i - \delta_j - \beta_i\phi_j))}, \tag{5}$$

where β_i equals 1 if person i is a guesser and 0 otherwise, ϕ_j is the motivation (guessing) factor associated with item j, which describes the extent of effort held back on item j. Because we assume a gradual change in the motivation factor, parameter ϕ_j can be considered as a smooth function of item location j. Note that if $\beta_i = 0$ (the ith examinee is a nonguesser), the model reduces to the 2PL-IRT model.

The priors for parameters θ_i's, δ_j's, γ_j's, τ_θ, τ_δ, and indicator β_i's are the same as those specified in the IRT-DG model. Denote the guessing vector as $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_J)'$. A simple prior would be an iid normal prior on each ϕ_j. Such a prior is flexible enough to accommodate any guessing pattern, but the estimates are not smooth. Thus for our computation, we use the second-order difference intrinsic auto-regressive (IAR(2)) prior (Speckman and Sun, 2003). The IAR(2) prior assumes there is an unknown smooth function describing the guessing curve. It is the second order random walk smoothness prior, $\phi_j = 2\phi_{j-1} - \phi_{j-2} + \varepsilon_j$, $j = 3, \cdots, J$, with iid Gaussian errors ε_j and diffuse prior $\phi_1 \propto 1$ and $\phi_2 \propto 1$. Note that ϕ_j depends on its two immediate neighbors and thus the estimation

of the ϕ_i's can borrow strength from each other. It is easy to show that the IAR(2) prior is a discrete version of the cubic smoothing spline.

Because of the two diffuse priors $\phi_1 \propto 1$ and $\phi_2 \propto 1$, the IAR(2) prior is improper where the $J \times J$ precision matrix of $\boldsymbol{\phi}$ is singular with rank $J-2$. In this case, we should check whether the posterior distribution is proper. It is reasonable to assume that there is little guessing behavior at the beginning of the test. Based on the assumption, we can set $\phi_1 = 0$ and $\phi_2 = 0$, making the IAR(2) prior proper and the Bayesian computation more efficient.

Written in vector format, the adjusted IAR(2) prior has density,

$$[\,\boldsymbol{\phi_3}\,|\,\tau_\phi] \propto \frac{1}{\left(\tau_\phi^{(J-2)/2}\right)} \exp\left(-\frac{1}{2\tau_\phi}\boldsymbol{\phi_3}'\mathbf{V}_\phi\boldsymbol{\phi_3}\right), \tag{6}$$

where $\boldsymbol{\phi_3} = (\phi_{3,}\phi_{4,}\cdots,\phi_J)'$ and $\mathbf{V}_\phi/\tau_\phi$ is the precision matrix. The full rank $(J-2)\times(J-2)$ matrix $\mathbf{V}_\phi$ has the form

$$\mathbf{V}_\phi = \begin{bmatrix} 6 & -4 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ -4 & 6 & -4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -4 & 5 & 2 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

The variance parameter τ_ϕ is assumed to follow a conjugate inverse gamma prior,

$$\tau_\phi \sim \mathrm{IG}(a_\phi, b_\phi),$$

where the hyperparameters are fixed at $a_\phi = 2$ and $b_\phi = 1$. The full conditional distributions of the IRT-CG model, listed in the Appendix, either have closed forms or are log-concave, so that they can be efficiently updated.

3 Simulation Study

We conducted a simulation study in order to examine the performance of our models and estimation procedures. Ten replications of a sample of 2000 examinees for a test of 40 items

were generated under each of four models: the three guessing models (IRT-TG, IRT-DG, IRT-CG), and the 2PL-IRT model. Then each of the forty datasets was analyzed by the four models. Because of the relatively large sample size and the number of candidate models, only ten replications were generated under each model to save computing time. Nonetheless, the conclusion is the same based on the estimates from individual samples and the pooled estimates.

The true values of the θ_i's and δ_j's were drawn from the standard Normal distribution. The true γ's were drawn from the uniform distribution $U(0.5, 2)$. Under the IRT-TG model, we assumed that 60% of the examinees are guessers, the item location threshold parameter α_j's are simulated with probability p_j from formula (3) where parameter ω is set to be 3.0. Under the IRT-DG model, we assumed 60% of the examinees are guessers, and the relative-difference threshold parameter η is set to be -1.0. For these two models, the guessing probability g_j is set to be 0.25, assuming each multiple choice question has 4 options. Under the IRT-CG model, we assumed that 60% of the examinees are unmotivated examinees (guessers), and the motivation vector $\boldsymbol{\phi_3} = (\phi_{3,}\phi_{4,} \cdots, \phi_J)'$ comes from an exponential curve.

With the data generated from the 2PL-IRT model, we can investigate whether the estimates from the guessing models are undermined by fitting the guessing models when there is no guessing behavior. With the data generated from the guessing models, we can examine how much improvement each guessing model can provide compared to the other models.

Table 1 presents the Pearson correlation coefficients between the true values of the parameters and their estimates. With the data generated by the 2PL-IRT model (first column of Table 1), all three guessing models are successful in identifying all the examinees as nonguessers for each of the ten datasets. More specifically, all of the item location threshold parameter α_i's from the IRT-TG model take 40 as their estimate, indicating all the examinees are nonguesssers. Figure 2 shows the posterior probabilities of examinees being guessers ($\beta_i = 1$) from the IRT-CG model (top panel) and from the IRT-DG model (bottom panel). The probabilities are all below 0.10, which indicates that the examinees are unlikely to be guessers. The correct estimation of examinees' nonguesser/guesser status reduces the guessing models to the true 2PL-IRT model, and this explains why all the four models yield the same correlations.

Table 1: Simulation Results (r)

Model\Data	2PL-IRT	IRT-TG	IRT-DG	IRT-CG
2PL-IRT	$r_\theta = 0.993$ $r_\delta = 1.000$ $r_\gamma = 0.999$	$r_\theta = 0.939$ $r_\delta = 0.925$ $r_\gamma = 0.799$	$r_\theta = 0.913$ $r_\delta = 0.973$ $r_\gamma = 0.632$	$r_\theta = 0.917$ $r_\delta = 0.948$ $r_\gamma = 0.962$
IRT-TG	$r_\theta = 0.993$ $r_\delta = 1.000$ $r_\gamma = 0.999$	$r_\theta = 0.987$ $r_\delta = 1.000$ $r_\gamma = 0.996$	$r_\theta = 0.903$ $r_\delta = 0.977$ $r_\gamma = 0.681$	$r_\theta = 0.918$ $r_\delta = 0.949$ $r_\gamma = 0.960$
IRT-DG	$r_\theta = 0.993$ $r_\delta = 1.000$ $r_\gamma = 0.999$	$r_\theta = 0.932$ $r_\delta = 0.927$ $r_\gamma = 0.811$	$r_\theta = 0.953$ $r_\delta = 0.999$ $r_\gamma = 0.952$	$r_\theta = 0.917$ $r_\delta = 0.948$ $r_\gamma = 0.961$
IRT-CG	$r_\theta = 0.993$ $r_\delta = 1.000$ $r_\gamma = 0.999$	$r_\theta = 0.970$ $r_\delta = 0.977$ $r_\gamma = 0.813$	$r_\theta = 0.851$ $r_\delta = 0.990$ $r_\gamma = 0.758$	$r_\theta = 0.977$ $r_\delta = 0.999$ $r_\gamma = 0.998$

With the data generated by the three guessing models, the correlation coefficients are higher based on the true model than those from the other models. By comparison, the IRT-TG model provides the best alignment between the estimates and the true values. This is because of its relatively stronger assumption and parsimonious specification of the parameters. It assumes that after a certain item guessers switch from the solution behavior to the random guessing behavior. It is a switch-only-once strategy, which is different from the IRT-DG model where guessers can switch back and forth based the item relative-difference. Also there is no transition between the solution behavior and the guessing behavior, in contrast to the IRT-CG model which assumes a gradual guessing effect. From this perspective, the assumption of the IRT-TG model provides more information about the guessing pattern. Furthermore, though the item location threshold parameter α_i for a guesser can take any integer from 1 to $J-1$, based on the prior specification, there is only one hyperparameter ω determines the discrete distribution (2). The parsimonious prior specification brings more power in the estimation.

Table 2 presents the mean of $\sqrt{mse}$ over the estimates. These measurements evaluate the

Table 2: Simulation Results ($\sqrt{mse}$)

Model\Data	2PL-IRT	IRT-TG	IRT-DG	IRT-CG
2PL-IRT	$\sqrt{mse}_\theta = 0.303$	$\sqrt{mse}_\theta = 0.405$	$\sqrt{mse}_\theta = 0.467$	$\sqrt{mse}_\theta = 0.464$
	$\sqrt{mse}_\delta = 0.74$	$\sqrt{mse}_\delta = 0.338$	$\sqrt{mse}_\delta = 0.612$	$\sqrt{mse}_\delta = 0.424$
	$\sqrt{mse}_\gamma = 0.073$	$\sqrt{mse}_\gamma = 0.245$	$\sqrt{mse}_\gamma = 0.424$	$\sqrt{mse}_\gamma = 0.167$
IRT-TG	$\sqrt{mse}_\theta = 0.305$	$\sqrt{mse}_\theta = 0.357$	$\sqrt{mse}_\theta = 0.488$	$\sqrt{mse}_\theta = 0.495$
	$\sqrt{mse}_\delta = 0.083$	$\sqrt{mse}_\delta = 0.096$	$\sqrt{mse}_\delta = 0.580$	$\sqrt{mse}_\delta = 0.395$
	$\sqrt{mse}_\gamma = 0.083$	$\sqrt{mse}_\gamma = 0.111$	$\sqrt{mse}_\gamma = 0.404$	$\sqrt{mse}_\gamma = 0.214$
IRT-DG	$\sqrt{mse}_\theta = 0.305$	$\sqrt{mse}_\theta = 0.413$	$\sqrt{mse}_\theta = 0.407$	$\sqrt{mse}_\theta = 0.496$
	$\sqrt{mse}_\delta = 0.081$	$\sqrt{mse}_\delta = 0.312$	$\sqrt{mse}_\delta = 0.248$	$\sqrt{mse}_\delta = 0.396$
	$\sqrt{mse}_\gamma = 0.084$	$\sqrt{mse}_\gamma = 0.218$	$\sqrt{mse}_\gamma = 0.333$	$\sqrt{mse}_\gamma = 0.214$
IRT-CG	$\sqrt{mse}_\theta = 0.306$	$\sqrt{mse}_\theta = 0.376$	$\sqrt{mse}_\theta = 0.571$	$\sqrt{mse}_\theta = 0.399$
	$\sqrt{mse}_\delta = 0.084$	$\sqrt{mse}_\delta = 0.167$	$\sqrt{mse}_\delta = 0.522$	$\sqrt{mse}_\delta = 0.131$
	$\sqrt{mse}_\gamma = 0.084$	$\sqrt{mse}_\gamma = 0.224$	$\sqrt{mse}_\gamma = 0.468$	$\sqrt{mse}_\gamma = 0.125$

estimates from the magnitude perspective. The estimates from the true guessing model have the smallest $\sqrt{mse}$ than those from other three models. Meanwhile, the estimates have the similar $\sqrt{mse}$ from the four models for data generated from the 2PL IRT model because each of the three guessing models can identify all the examinees as nonguessers correctly. The conclusion is consistent with that based on the Pearson correlation coefficient from Table 1. So the estimates from the true model preserve not only the relative ranking but also the absolute magnitude of the parameters.

To visualize the improvement, we also present the following figures. Figure 3 shows the estimates of item difficulty parameters δ_j's versus the true values based on the data generated from the IRT-TG model. The estimates from the IRT-TG model (top panel) have a perfect alignment with the true values, while those from the 2PL-IRT model (bottom panel) tend to overestimate the item difficulty. The closer the item located near the end of test, the more serious the overestimation (see the circles for the last ten items in Figure 3). Once examinees switch to guessing, they randomly guess on the test items regardless of the item difficulty, so easy items located near the end of test are most severely affected. The low correct-response

rate of these items is interpreted as the items being difficult by the 2PL-IRT model. For example, the circle farthest from the line in Figure 3 is the item with the true difficulty -0.52. It is the 39th item in the test. The estimate from the 2PL-IRT model is 1.09.

Figure 4 plots the estimates of item discrimination parameters γ_j's versus the true values based on the data generated from the IRT-TG model. The estimates from the IRT-TG model (top panel) are unbiased, while those from the 2PL-IRT model tend to underestimate item discrimination parameters. This is because when examinees switch to guessing, the probability of correct response is pulled towards the guessing probability g_j, regardless of the ability. The 2PL-IRT model would interpret this as the item being less discriminating. The closer the item placed near the end of test, the more guessing behavior involved, and the more obvious the underestimation is (see the circles for the last ten items in Figure 4).

Figure 5 shows the estimates of ability parameters θ_i's versus the true values based on the data generated from the IRT-TG model. There are 1200 guessers who start guessing from different items. To have a clear comparison, we only show the estimates for guessers with $\alpha_i = 22$, which means their responses after item 22 are randomly chosen. The estimates from the IRT-TG model (top panel) are unbiased, while those from the 2PL-IRT model (bottom panel) tend to underestimate ability. In the 2PL-IRT model, the number of correct answers determines the ability estimate. Guessers with high ability would have fewer correct responses due to guessing, and their abilities would be underestimated while those with low ability would be less affected since their correct response rate would be lower even without guessing. Also, the earlier the guessing behavior starts, the more serious the underestimation with the 2PL-IRT model would be. As for the IRT-TG model, the estimates of parameters δ_j's, γ_j's and θ_i's are unbiased because only the responses based on the solution behaviors are used.

Figure 6 (IRT-CG) and Figure 7 (IRT-DG) are the histograms showing the posterior probability of examinee being a guesser ($\beta_i = 1$). Using 0.5 as the threshold (the vertical line in both figures), the two guessing models have classified the majority of examinees into the right category. The simulation study indicates that, for the sample size considered, the estimates will not suffer by fitting the guessing models when there is no guessing behavior. This is because all three guessing models classify the examinees as nonguessers correctly,

which reduces the guessing models to the true 2PL-IRT model. On the other hand, if guessing behavior is present, the true guessing model can adjust the bias from the 2PL-IRT model. Furthermore, the true guessing model can produce reliable estimates on examinees' nonguesser/guesser status and estimates on the degree of guessing.

We have proposed three guessing models under different assumptions. A natural question to ask is which one to use? It is often impossible to know which kind of guessing behavior is prevalent in a test. We resort to a Bayesian model selection criterion. The Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002), is an extension from the AIC that can be calculated directly from the MCMC chain. It is a diagnostic that combines model fit with model complexity. Here we introduce the outline of the DIC. For model M_i, let $\boldsymbol{\theta}_i$ denote the vector of parameters, $f_i(\mathbf{y}|M_i, \boldsymbol{\theta}_i)$ the likelihood function, and $\bar{\boldsymbol{\theta}}_i$ the posterior mean of $\boldsymbol{\theta}_i$. The DIC for model M_i is given by

$$\text{DIC}_i = \bar{D}_i + p_{D_i}, \tag{7}$$

where

$$\begin{aligned} \bar{D}_i &= \text{E}_{\boldsymbol{\theta}_i|\mathbf{y}}[-2\log f_i(\mathbf{y}|M_i, \boldsymbol{\theta}_i)] \\ p_{D_i} &= \text{E}_{\boldsymbol{\theta}_i|\mathbf{y}}[-2\log f_i(\mathbf{y}|M_i, \boldsymbol{\theta}_i)] + 2\log f_i(\mathbf{y}|M_i, \bar{\boldsymbol{\theta}}_i). \end{aligned}$$

The first term $\bar{D}_i$ in (7) is the the posterior expectation of deviance. It can be treated as a Bayesian measure of model adequacy. The second term p_{D_i} is termed as "the effective number of parameters" , which is the difference between the posterior mean deviance and the deviance of the posterior mean. It serves as a penalty term that measures the complexity of the model. The effective number of parameters p_{D_i} is more appropriate in Bayesian hierarchical models compared to the nominal number of parameters. It is because the parameters in Bayesian hierarchical models are not independent, for example, some parameters may share the same prior distribution. As with AIC, a small value of DIC is preferred.

Figure 8 is the DIC plot. In the top panel, DIC scores are shown for the three guessing models under each of the ten datasets generated from the IRT-TG model. Ten out of the ten, DIC selects the true model. In the next two plots, the data are generated from the IRT-DG model and the IRT-CG model respectively. Again for each of the ten datasets, DIC

selects the true model. Note that there is a clear distinction between the DIC score of the true model and those from the other two models. So DIC can be used as a guidance to decide which guessing model is better.

4 Application

The ARTIST project (https://app.gen.umn.edu/artist/index.html), funded by a grant from the National Science Foundation, provides a variety of assessment tools for first courses in Statistics. For two semesters, the undergraduate students from our classes who took Business Statistics were asked to take the Comprehensive Assessment of Outcomes in a first Statistics Course (CAOS) test. The multiple choice test consists of 40 items and usually takes 30 to 45 minutes to complete. Because the test result did not affect the course grade, it was a low-stakes test. There were 265 students taking the test, among whom 111 had nonresponses in their answers. However, the response rate was 100% for the first four questions of the assessment, which supports our assumption that the guessing parameters on the first two items (ϕ_1 and ϕ_2) are zero in the IRT-CG model.

The majority of the nonresponses follow the threshold guessing pattern; that is, once there was a nonresponse, the subsequent items were not answered. There were a few tests with nonresponses between responses. To have a fair comparison of the three guessing models and the 2PL-IRT model, we impute the missing data, easily using MCMC, that is consistent with the model being fitted. For example, when the model IRT-DG is fitted, the missing data are imputed with the parameters from the IRT-DG model. So the models are not being fitted based on the same set of data, but rather on a complete dataset with the missing part imputed based on the model-specific assumption. Table 3 lists the DIC scores, which are evaluated on the observed data under each of the four models. The IRT-TG model has the smallest DIC score which indicates it is a better model compared to the other models. Note that the majority of the students have finished all the items in the exam. The winning IRT-TG model indicates that there is significant guessing behavior present in the test and the dominant guessing pattern from the observed data is the threshold guessing behavior. Because the test is low-stakes, the students were showing less and less motivation, which

explains why the IRT-DG model is less favored.

Table 3: DIC for CAOS Test

Model	2PL-IRT	IRT-TG	IRT-CG	IRT-DG
DIC	11734.029	11414.296	11619.222	11674.149

Next we use the Bayesian χ^2 test to evaluate the fit of the IRT-TG model. The test was proposed by Johnson (2004). The essential idea is to evaluate Pearson's goodness-of-fit statistic at parameter values drawn from the posterior distribution. Johnson shows that the statistic asymptotically follows a χ^2 distribution. Johnson also provides a rule of thumb to choose the number of cells which is to take $n^{0.4}$ equiprobable cells. In our case, with $n = 265$, that is 10 cells. For the convenience, we group every four items into one cell. Specifically, let $\boldsymbol{\vartheta}$ be the parameter vector, $\tilde{\boldsymbol{\vartheta}}$ be a sampled value $\boldsymbol{\vartheta}$ from the posterior. Then, define

$$R^B(\tilde{\boldsymbol{\vartheta}}) = \sum_{k=1}^{10} \frac{(n_k - m_k(\tilde{\boldsymbol{\vartheta}}))^2}{m_k(\tilde{\boldsymbol{\vartheta}})},$$

where $n_k = \sum_{i=1}^{n} \sum_{j=(k-1)+1}^{(k-1)+4} x_{ij}$ which is the total number of correct responses for the kth group of items, and $m_k(\tilde{\boldsymbol{\vartheta}}) = \sum_{i=1}^{n} \sum_{j=(k-1)+1}^{(k-1)+4} P(x_{ij} = 1|\tilde{\boldsymbol{\vartheta}})$ which is the sum of the probabilities of giving correct answers for the kth group of items from all the students. The statistic $R^B(\tilde{\boldsymbol{\vartheta}})$ has an asymptotic χ^2 distribution with 9 degrees of freedom. One advantage of the test is that values of $R^B(\tilde{\boldsymbol{\vartheta}})$ can be directly computed with the updates of the parameters within the MCMC schemes. In our study, the proportion of $R^B(\tilde{\boldsymbol{\vartheta}})$ values exceeding the 95th percentile from the χ^2_9 distribution is 0.0063. It shows that the IRT-TG model provides an adequate fit of the data.

We might expect to find that the faster the test was completed, the larger the number of items the examinee was guessing to answer. Since the ARTIST website provided the time each student spent on the test, this hypothesis could be examined. The time spent varied from about 4 to 30 minutes. If the hypothesis is true, we would expect a high correlation between the two variables, the time spent on the test and the item location threshold. This is because the less the time, the earlier the guessing begins, and the smaller the item location parameter. Actually the correlation between the two variables is only 0.3, showing a rather weak linear relationship. Our explanation is that there may be some motivated examinees

who are good at logic thinking and are capable of finishing the test in a short amount of time. On the other hand, some guessers may take their time to solve the first couple of problems where they have spent a considerable amount of time, and then they lost interest. At least based on this dataset, time spent on the test is not a very good indicator of guessing behavior.

5 Discussion

In this paper, we propose three Bayesian IRT guessing models to accommodate different kinds of guessing behavior for a multiple choice test. The 3PL-IRT model assumes all the students behave exactly the same way on all the items. The Bayesian guessing models are constructed under a more realistic assumption that examinees include both nonguessers and guessers, and the guessers could show different degree of guessing behavior. Adjusted estimates of the IRT parameters and the inference on the degree of guessing are achieved by fitting the models. We also propose to use the DIC to determine which kind of guessing behavior is dominant.

Our three models include two in which the likelihood of guessing or low effort is related to the location of the item in the test. The first of these, the IRT threshold guessing model (IRT-TG), may be appropriate for a low-stakes test in which the examinees are motivated at first by curiosity or interest in the test, and then abruptly abandon effort and begin to select responses at random after a certain point. This model may also be appropriate for speeded tests. The second model, the IRT continuous guessing model (IRT-CG), assumes a gradual change of examinee motivation that does not result (necessarily) in guessing, but just low effort that results in a smaller probability of correct response. This can occur on a low-stakes test under the same conditions as discussed above for the IRT-TG model, or due to fatigue. Both of these patterns result in more distortion to items late in the test, especially low-difficulty ones. This observation can be applied in the design of a test. If a researcher wants to investigate the degree of this item-location-related guessing behavior in the test, he should put more easy test items near the end. The discrepancy of the correct response rates on these easy test items will help identify the presence and type of guessing

or low motivation behavior.

The third model we consider is the IRT difficulty-based guessing model (IRT-DG). It assumes the guessers only answer the relatively easy test items and guess on the remaining items. It differs from the other two in that the chance of guessing on an item is not related to its location. Unlike the "switch-only-once" strategy in the IRT-TG model, the guessers can switch multiple times between solution behavior and guessing behavior in the test.

All the three guessing models assume that the guessers employ a certain homogeneous guessing strategy. So they do not accommodate the possibility that different guessers exhibit different strategies. Presumably that would be possible, but it would take a very long test, which would be unlikely for a low-stakes test. Our goal is to apply these models to find the most dominant guessing pattern, as shown in the real data analysis in the last section. We also need to point out that these models won't help us correct scores for people who start out with low motivation or guessing, though the models are capable of identify them as guessers. The estimates of those guessers' ability will be around the mean of the group ability. This is because we won't have much information on which to estimate their true state. Our leverage in these models comes from having some period of time when examinees exhibit their own natural ability, and then change over the course of the test.

In low-stakes tests where different groups of examinees have participated, researchers may be more interested in estimates of subgroup mean abilities rather than estimates of individual abilities. Our guessing models can not only provide better estimates on group mean ability, they can also help identify specific guessing patterns for each group. Thus they may shed light on the evaluation of effectiveness of the test and provide useful information on future test design.

According to Nation's Report Card, 12th-graders' performance on the NAEP reading assessment has been declining over the last decade: the average scale scores are 290, 287, and 286 in year 1998, 2002, and 2005, respectively. By comparison, 4th-graders' performance has been improving, where their scores for the three years are 215, 219, and 219. However, 12th-graders are taking more advanced courses and earning better grades on average in school. Brophy and Ames (2005) in their report for the National Assessment Governing Board

have suggested that "the NAEP assessment of twelfth graders faces daunting motivational obstacles". They have also discussed two alternatives to change the 12th-grade NAEP. One is to abandon the 12th-grade NAEP testing, the other is to continue or expand the test where strategies engaging motivation orientations are required. Both of the alternatives have significant impact over the NAEP program. In order to make an informed decision, we could use the guessing models to evaluate whether low motivation is serious in 12th-graders NAEP and what is the dominating pattern of low motivation.

The code for fitting the models discussed in this paper was written in Fortran. We also wrote WinBUGS programs that can be downloaded from the website at http://www.smu.edu/statistics/TechReports/tech-rpts.asp. The estimates from the WinBUGS programs are very similar to those from our Fortran programs. WinBUGS is a free software to run Bayesian data analysis. The biggest attraction of the software is that it does not require the specification of the full conditional distributions or the computation algorithm. All it needs is the model likelihood and the assignment of the priors. However the WinBUGS program cannot provide the DIC score because of the mixture hierarchical structure.

REFERENCES

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer.

Bechger, T., Maris, G., Verstralen, H., & Verhelst, N. (2003). The Nedelsky model for multiple choice items. R & D report, Arnhem: Cito.

Brophy, J., & Ames, C. (2005). NAEP testing for twelfth graders: motivational issues. A paper prepared for the National Assessment Governing Board.

Gilk, W., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.

Grigg, W., Donahue, P., & Dion, G. (2007). The nation's report card: 12th-grade reading and mathematics 2005. National Center for Education Statistics.

Johnson, V. (2004). A Bayesian χ^2 test for goodness of fit. *Annals of Statistics*, 32, 2361-2384.

Martin, E.S., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, 30, 183-203.

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215.

Neal, R.M. (2003). Slice sampling. *The Annuals of Statistics* , 31, 705-767.

Rijmen, F., De Boeck, P., & van der Maas, H.L.J. (2005). An IRT model with a parameter-driven process for change. *Psychometrika*, 70, 651-669.

Speckman, P.L., & Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90, 289-302.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Methodological*, 64, 583–616.

Wise, S.L., & DeMars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement,* 43 (1), 19-38.

Wise, S.L., & DeMars, C.E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment,* 10 (1), 1-17.

Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education,* 18 (2), 1-17.

Yamamoto K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. TOEFL technical report No. TR-10.

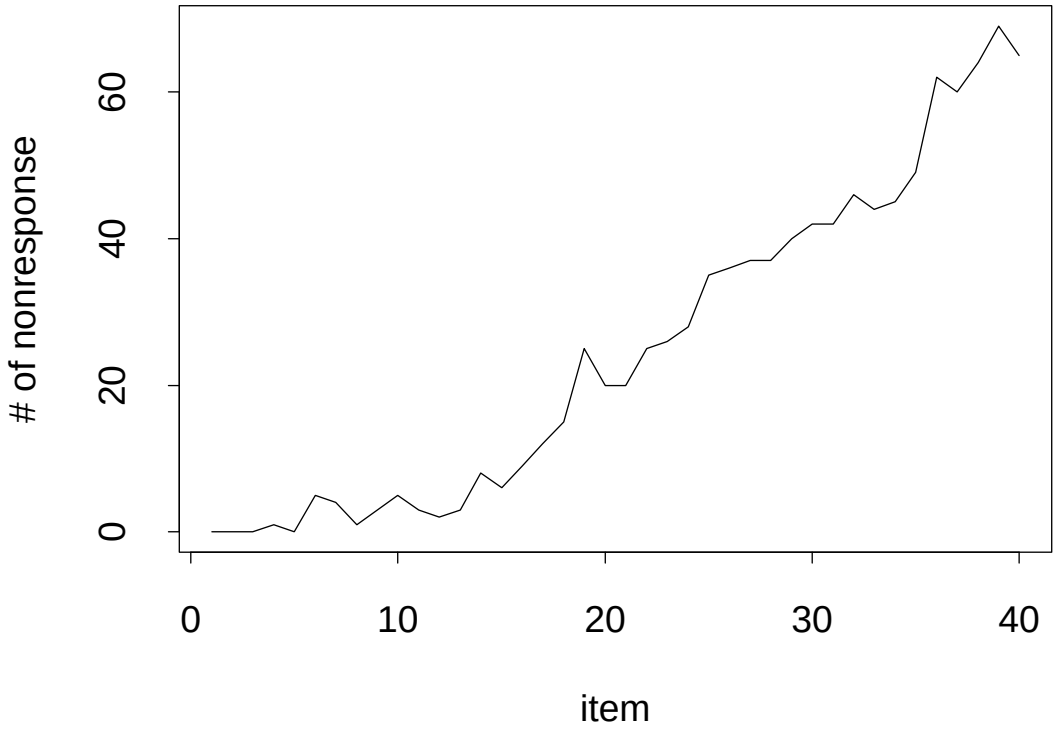

Figure 1: Number of nonresponse on each item.

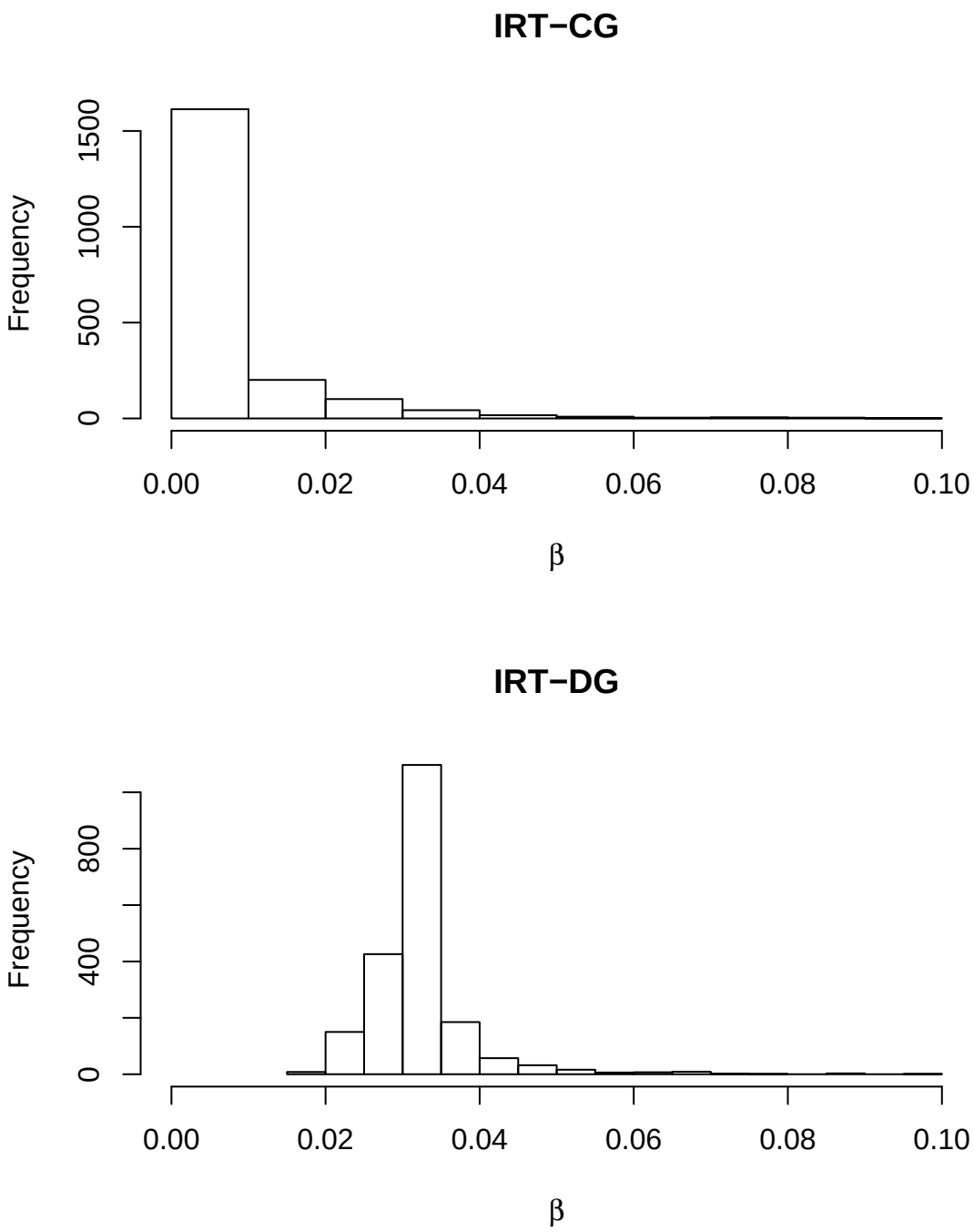

Figure 2: Posterior probability of β_i based on the 2PL-IRT data.

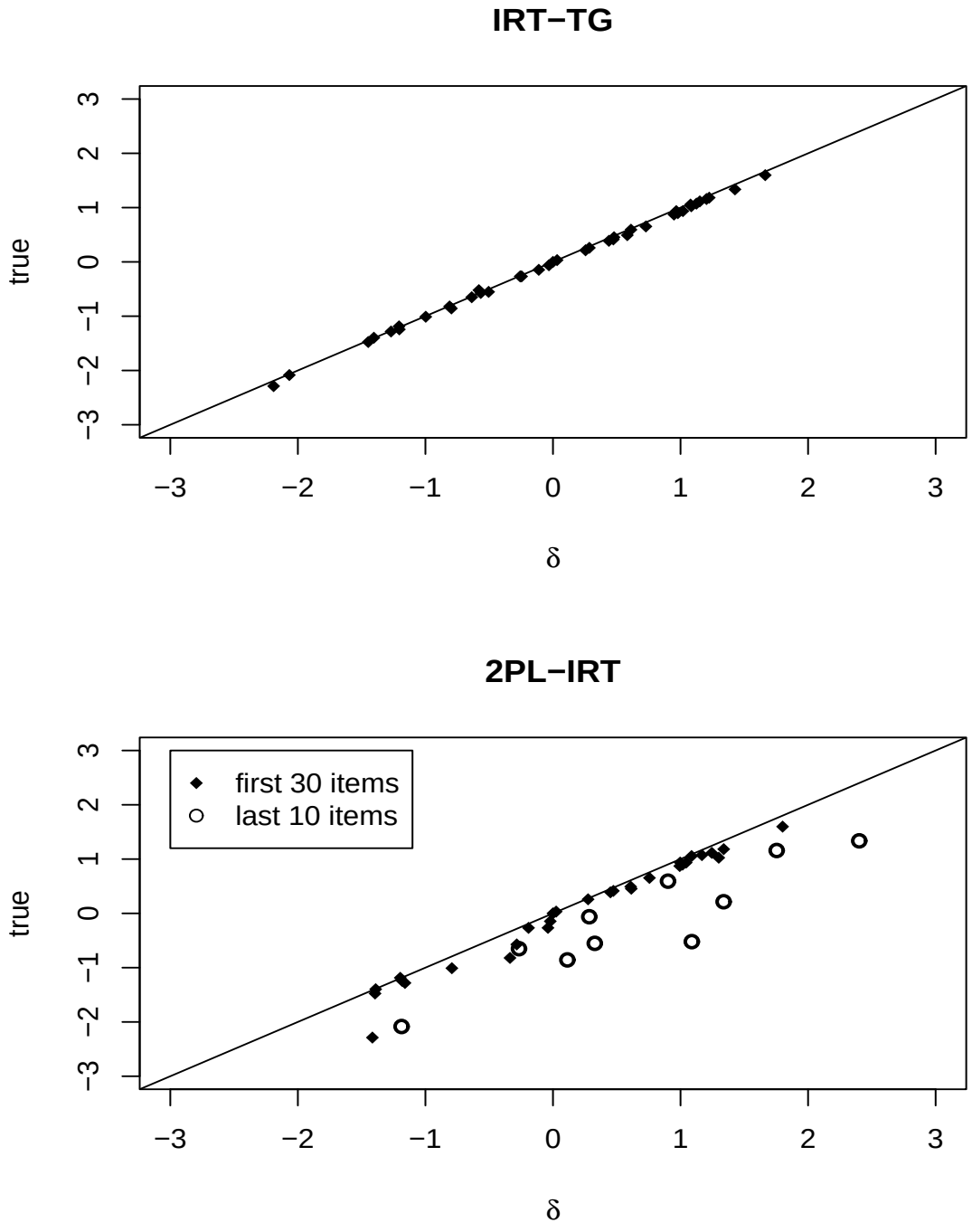

Figure 3: Estimates of δ_j's under the IRT-TG model and the 2PL-IRT model.

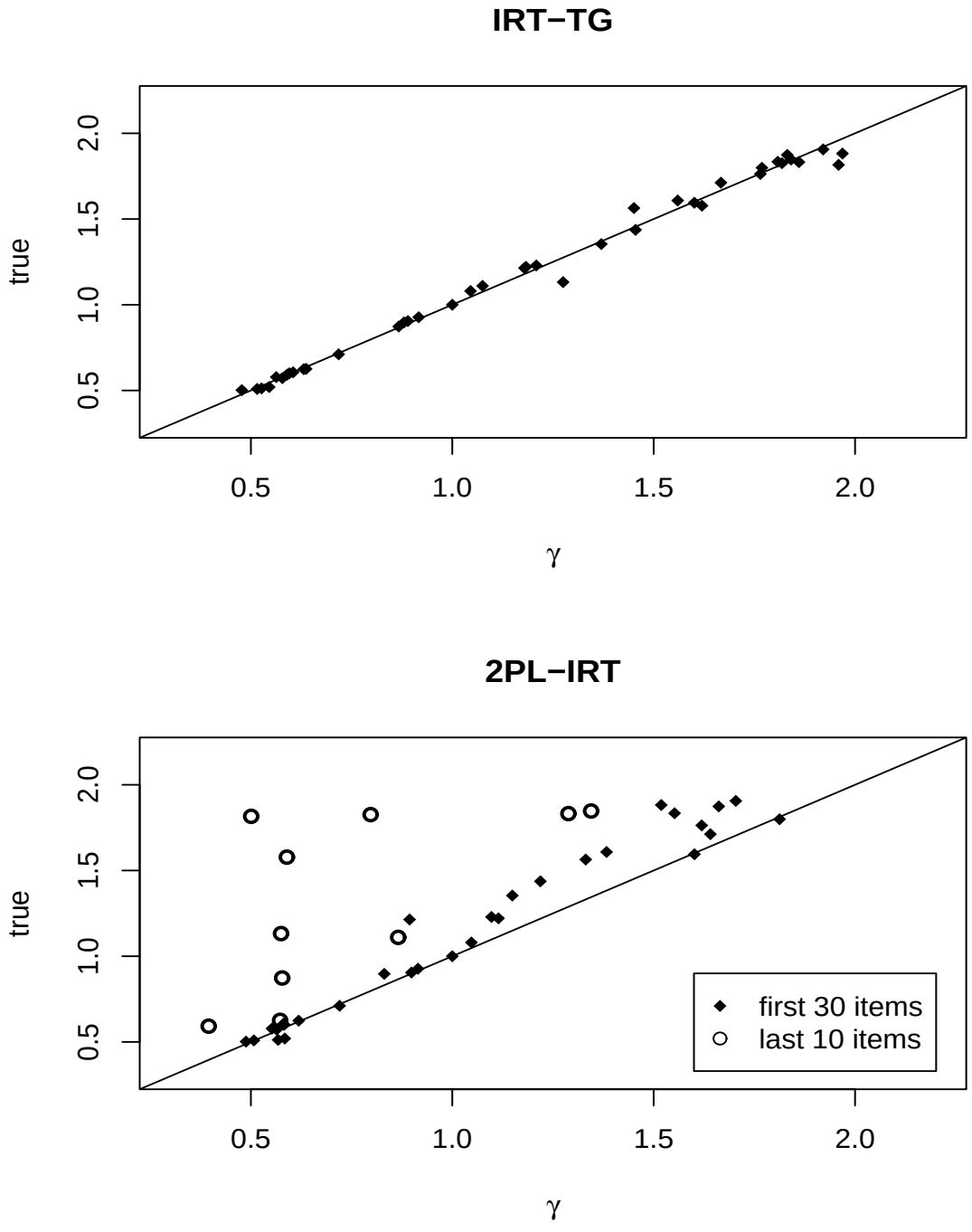

Figure 4: Estimates of γ_j's under the IRT-TG model and the 2PL-IRT model.

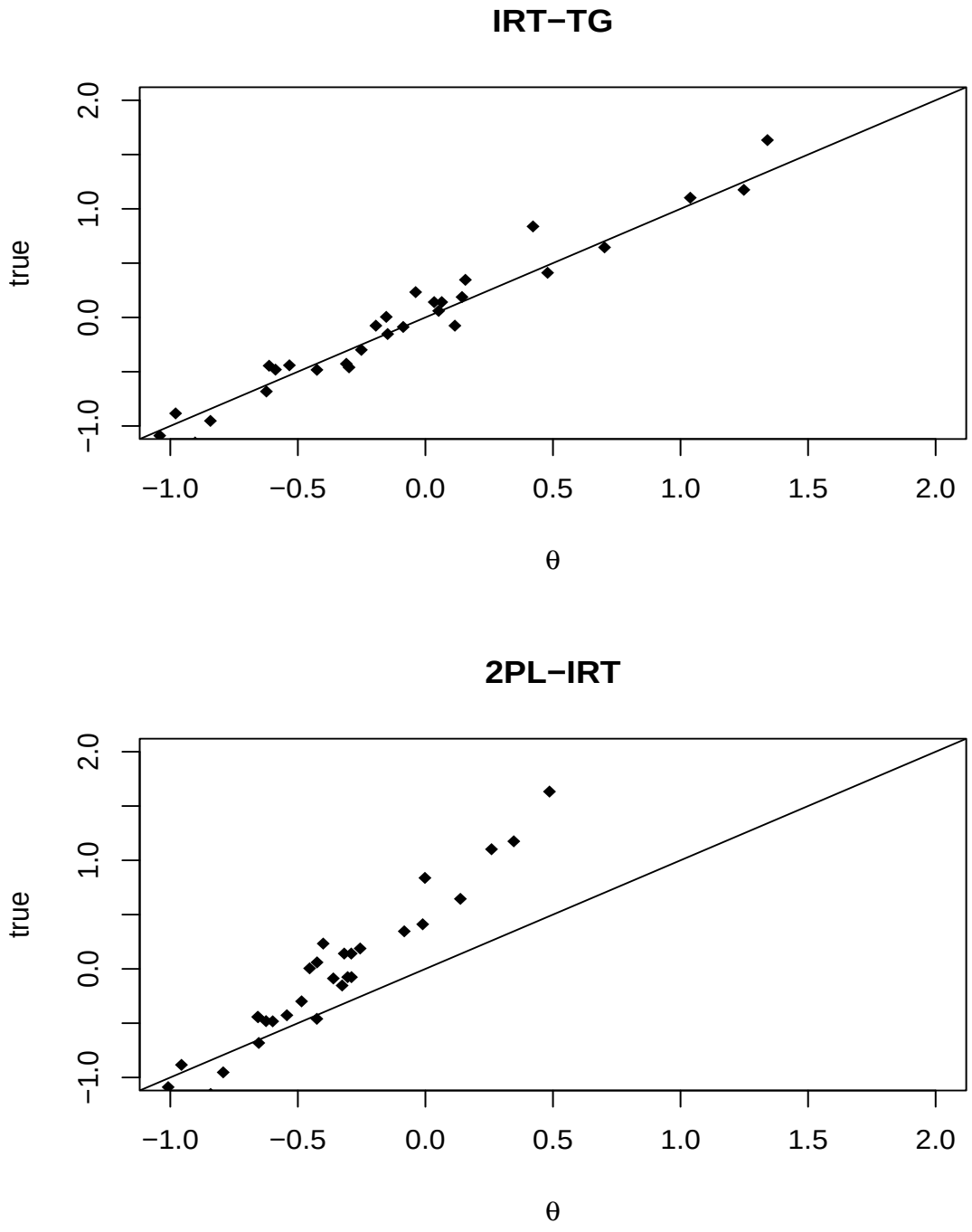

Figure 5: Estimates of θ_i's under the IRT-TG model and the 2PL-IRT model.

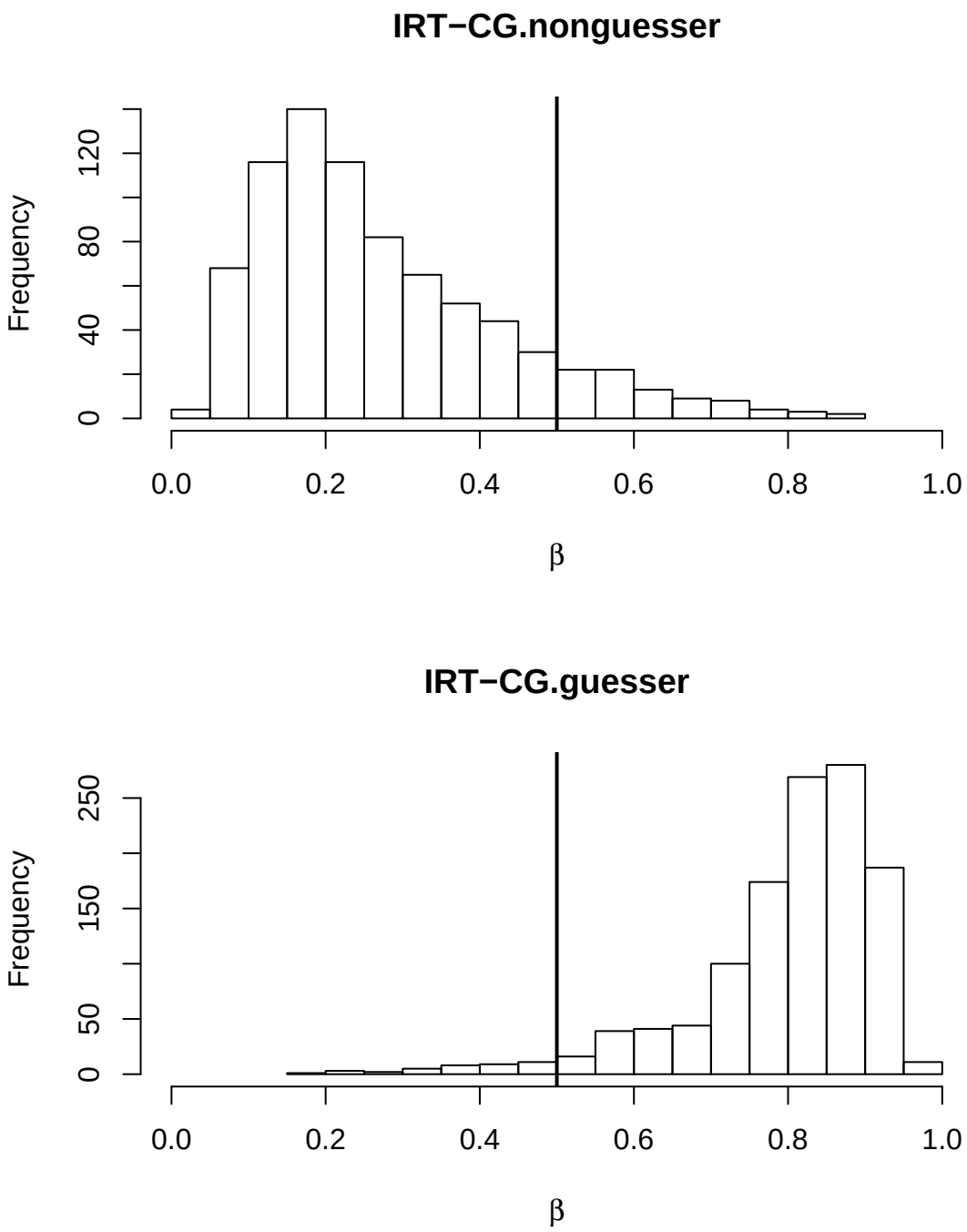

Figure 6: Posterior probability of β_i under the IRT-CG model.

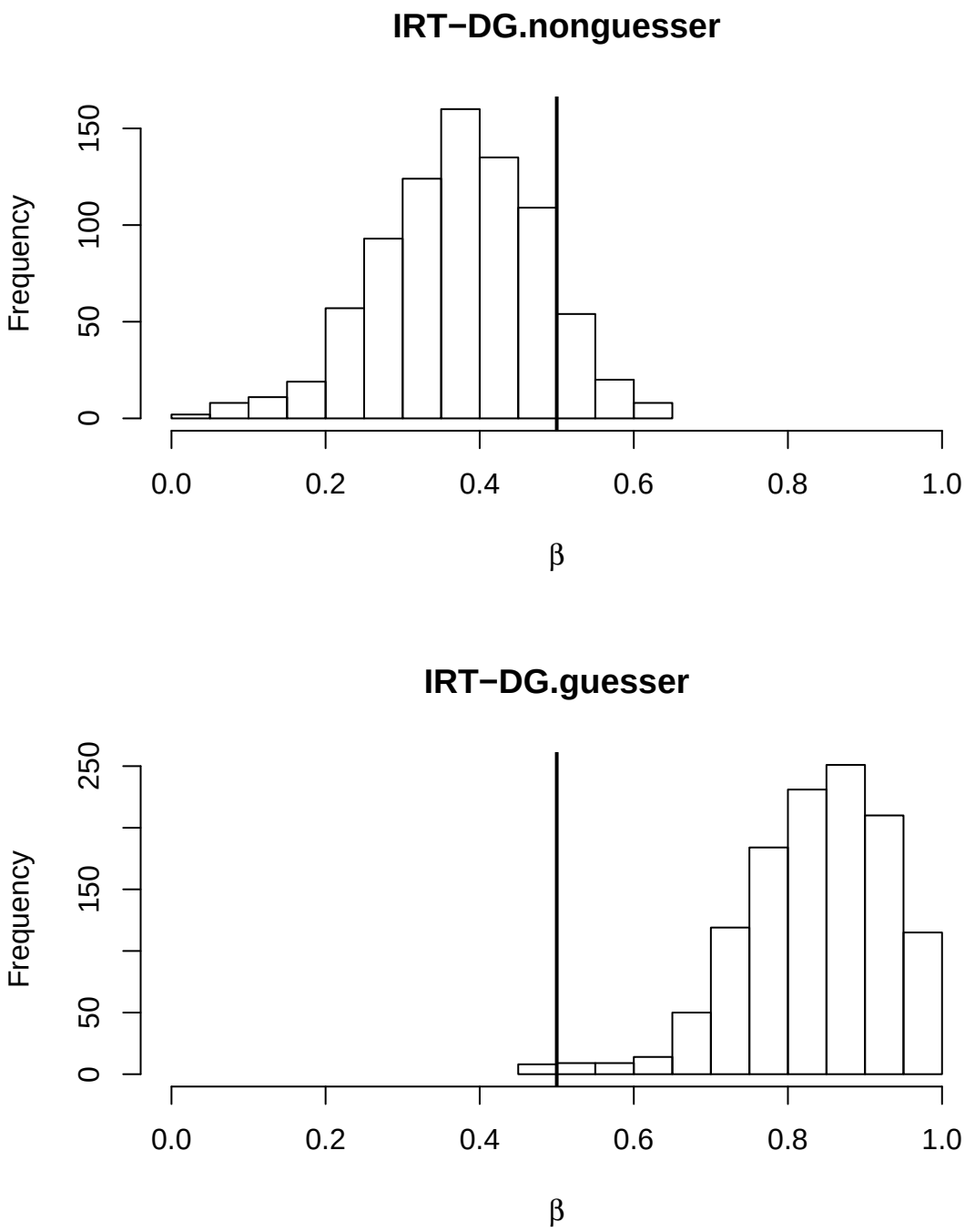

Figure 7: Posterior probability of β_i under the IRT-DG model.

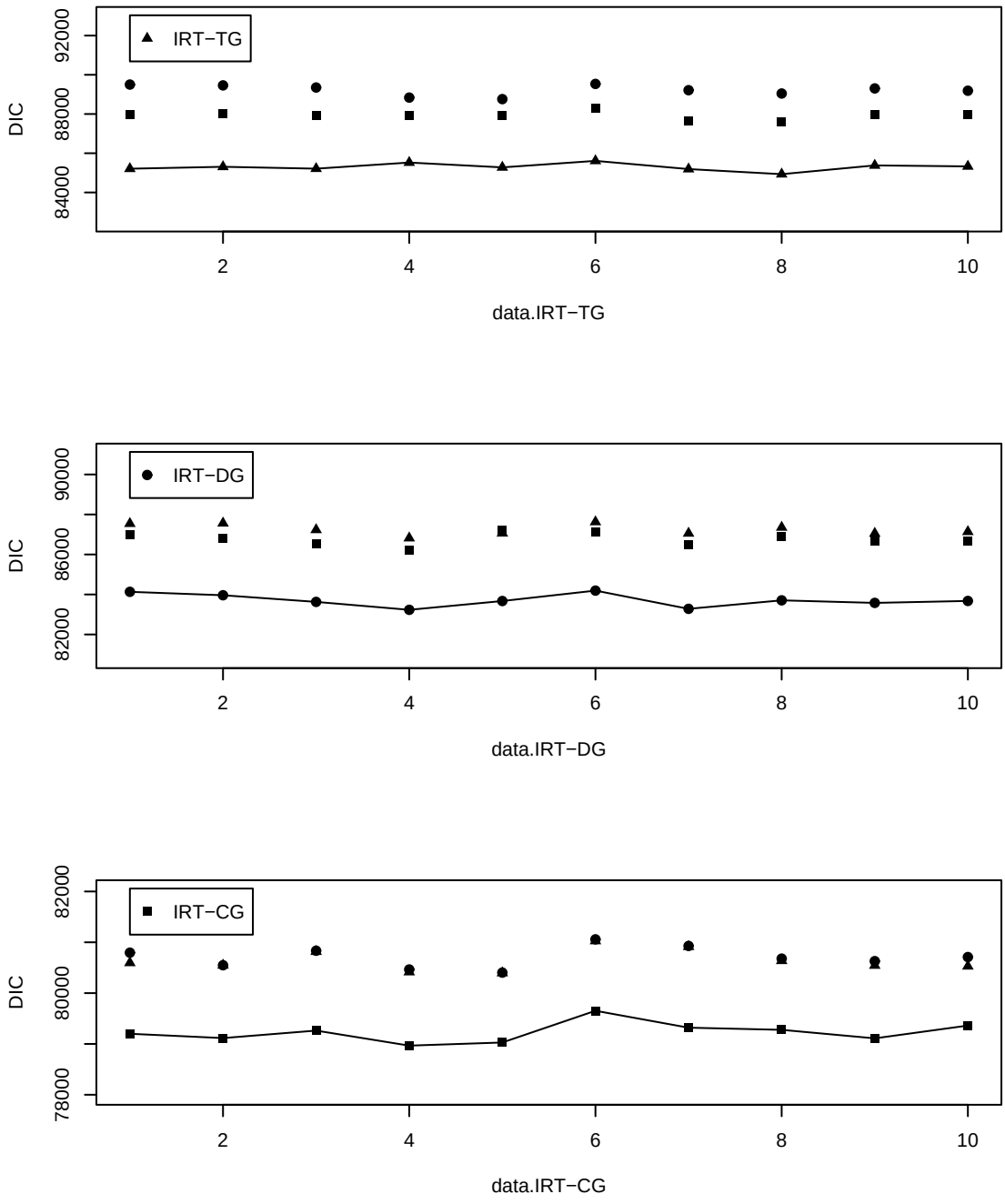

Figure 8: DIC plot for the three guessing models.

Appendix

1. Full conditional distributions for the IRT-TG model.

a) Given the other parameters and data, the full conditional distribution of θ_i is

$$(\theta_i \mid others; data) \propto \frac{\exp[\theta_i \sum_{j=1}^{\alpha_i} x_{ij}\gamma_j]}{\prod_{j=1}^{\alpha_i}[1+\exp(\gamma_j(\theta_i-\delta_j))]} \exp\Big(-\frac{\theta_i^2}{2\tau_\theta}\Big),$$

b) Given the other parameters and data, the full conditional distribution of δ_j is

$$(\delta_j \mid others; data) \propto \frac{\exp[-\gamma_j\delta_j \sum_{i:\,\alpha_i\geq j} x_{ij}]}{\prod_{i:\,\alpha_i\geq j}[1+\exp(\gamma_j(\theta_i-\delta_j))]} \exp\Big(-\frac{\delta_j^2}{2\tau_\delta}\Big).$$

c) Given the other parameters and data, the full conditional distribution of γ_j is

$$(\gamma_j \mid others; data) \propto \frac{\exp[\gamma_j \sum_{i:\,\alpha_i\geq j} x_{ij}(\theta_i-\delta_j)]}{\prod_{i:\,\alpha_i\geq j}[1+\exp(\gamma_j(\theta_i-\delta_j))]} \gamma_j^{a_\gamma-1} \exp(-b_\gamma\gamma_j).$$

d) Given the other parameters and data, the full conditional distribution of τ's is an inverse gamma distribution,

$$\begin{aligned}(\tau_\theta \mid others; data) &\sim \mathrm{IG}\Big(\frac{n}{2}+a,\ \frac{1}{2}\sum_{i=1}^{n}\theta_i^2+b\Big),\\ (\tau_\delta \mid others; data) &\sim \mathrm{IG}\Big(\frac{J}{2}+a,\ \frac{1}{2}\sum_{j=1}^{J}\delta_i^2+b\Big).\end{aligned}$$

e) Given the other parameters and data, the full conditional distribution of α_i is a discrete distribution,

$$\begin{aligned}&P(\alpha_i = l \mid others; data)\\ =\ &\frac{\prod_{j=1}^{l}\frac{\exp(x_{ij}\gamma_j(\theta_i-\delta_j))}{1+\exp(\gamma_j(\theta_i-\delta_j))}\prod_{j=l+1}^{J}(g_j)^{x_{ij}}(1-g_j)^{(1-x_{ij})}p_l}{\sum_{h=1}^{J}\Big(\prod_{j=1}^{h}\frac{\exp(x_{ij}\gamma_j(\theta_i-\delta_j))}{1+\exp(\gamma_j(\theta_i-\delta_j))}\prod_{j=h+1}^{J}(g_j)^{x_{ij}}(1-g_j)^{(1-x_{ij})}p_h\Big)}.\end{aligned}$$

f) Given the other parameters and data, the full conditional distribution of p_J is

$$(p_J \mid others; data) \sim \mathrm{beta}(\textstyle\sum_{i=1}^{n} I(\alpha_i = 40) + b_1,\ n - \sum_{i=1}^{n} I(\alpha_i = 40) + b_2),$$

where $I(\alpha_i = 40)$ equals 1 if $\alpha_i = 40$ and 0 otherwise.

g) Given the other parameters and data, the full conditional distribution of ω is

$$(\omega \mid others; data) \propto \prod_{i:\,\alpha_i<J} \frac{\alpha_i^{\omega} - (\alpha_i - 1)^{\omega}}{(J-1)^{\omega}} \omega^{a_\omega - 1} \exp(-b_\omega \omega).$$

h) Given the other parameters and data, the full conditional distribution of g_j is a truncated beta distribution

$$(g_j \mid others; data) \sim \text{beta}\left(\sum_{i:\,\alpha_i<j} x_{ij} + 1,\ \sum_{i:\,\alpha_i<j} (1 - x_{ij}) + 1\right), \qquad g_j \in (0,\, 0.5).$$

2. Full conditional distributions for the IRT-DG model.

a) Given the other parameters and data, the full conditional distribution of θ_i is

$$\begin{aligned}&(\theta_i \mid others; data)\\ \propto\ &\prod_{j=1}^{J} \frac{\exp\left\{x_{ij}\left[\gamma_j(\theta_i - \delta_j) - \beta_i I(\delta_j - \theta_i - \eta)(\gamma_j(\theta_i - \delta_j) - c_j)\right]\right\}}{1 + \exp\left[\gamma_j(\theta_i - \delta_j) - \beta_i I(\delta_j - \theta_i - \eta)(\gamma_j(\theta_i - \delta_j) - c_j)\right]} \exp\Big(-\frac{\theta_i^2}{2\tau_\theta}\Big).\end{aligned}$$

b) Given the other parameters and data, the full conditional distribution of δ_j is

$$\begin{aligned}&(\delta_j \mid others; data)\\ \propto\ &\prod_{i=1}^{n} \frac{\exp\left\{x_{ij}\left[\gamma_j(\theta_i - \delta_j) - \beta_i I(\delta_j - \theta_i - \eta)(\gamma_j(\theta_i - \delta_j) - c_j)\right]\right\}}{1 + \exp\left[\gamma_j(\theta_i - \delta_j) - \beta_i I(\delta_j - \theta_i - \eta)(\gamma_j(\theta_i - \delta_j) - c_j)\right]} \exp\Big(-\frac{\delta_j^2}{2\tau_\delta}\Big).\end{aligned}$$

c) Given the other parameters and data, the full conditional distribution of γ_j is

$$(\gamma_j \mid others; data) \propto \frac{\exp[\gamma_j \sum_{i:\,\beta_i=0 \text{ or } \delta_j-\theta_i\le\eta} x_{ij}(\theta_i - \delta_j)]}{\prod_{i:\,\beta_i=0 \text{ or } \delta_j-\theta_i\le\eta}[1 + \exp(\gamma_j(\theta_i - \delta_j))]} \gamma_j^{a_\gamma - 1} \exp(-b_\gamma \gamma_j).$$

d) Given the other parameters and data, the full conditional distribution of τ's is an inverse gamma distribution,

$$\begin{aligned}(\tau_\theta \mid others; data) &\sim \text{IG}\Big(\frac{n}{2} + a_\theta,\ \frac{1}{2}\sum_{i=1}^{n} \theta_i^2 + b_\theta\Big),\\ (\tau_\delta \mid others; data) &\sim \text{IG}\Big(\frac{J}{2} + a_\delta,\ \frac{1}{2}\sum_{j=1}^{J} \delta_i^2 + b_\delta\Big),\\ (\tau_\eta \mid others; data) &\sim \text{IG}\Big(\frac{1}{2} + a_\eta,\ \frac{1}{2}\eta^2 + b_\eta\Big).\end{aligned}$$

e) Given the other parameters and data, the full conditional distribution of β_i is a discrete distribution,

$$\begin{aligned} & P(\beta_i = 1 \mid others; data) \\ = & \frac{\prod_{j=1}^{J} \frac{\exp\{x_{ij}[\gamma_j(\theta_i-\delta_j)-I(\delta_j-\theta_i-\eta)(\gamma_j(\theta_i-\delta_j)-c_j)]\}}{1+\exp[\gamma_j(\theta_i-\delta_j)-I(\delta_j-\theta_i-\eta)(\gamma_j(\theta_i-\delta_j)-c_j)]} p_\beta}{\prod_{j=1}^{J} \frac{\exp\{x_{ij}[\gamma_j(\theta_i-\delta_j)-I(\delta_j-\theta_i-\eta)(\gamma_j(\theta_i-\delta_j)-c_j)]\}}{1+\exp[\gamma_j(\theta_i-\delta_j)-I(\delta_j-\theta_i-\eta)(\gamma_j(\theta_i-\delta_j)-c_j)]} p_\beta + \prod_{j=1}^{J} \frac{\exp[x_{ij}(\gamma_j(\theta_i-\delta_j))]}{1+\exp(\gamma_j(\theta_i-\delta_j))}(1-p_\beta)}. \end{aligned}$$

f) Given the other parameters and data, the full conditional distribution of p_β is

$$(p_\beta \mid others; data) \sim \text{beta}(\textstyle\sum_{i=1}^{n} \beta_i + a_p,\ n - \sum_{i=1}^{n} \beta_i + b_p).$$

g) Given the other parameters and data, the full conditional distribution of η is

$$\begin{aligned} & (\eta \mid others; data) \\ \propto & \prod_{i:\,\beta_i=1} \prod_{j=1}^{J} \frac{\exp\{x_{ij}[\gamma_j(\theta_i-\delta_j)-I(\delta_j-\theta_i-\eta)(\gamma_j(\theta_i-\delta_j)-c_j)]\}}{1+\exp[\gamma_j(\theta_i-\delta_j)-I(\delta_j-\theta_i-\eta)(\gamma_j(\theta_i-\delta_j)-c_j)]} \exp\Big(-\frac{\eta^2}{2\tau_\eta}\Big). \end{aligned}$$

3. Full conditional distributions for the IRT-CG model.

a) Given the other parameters and data, the full conditional distribution of θ_i is

$$(\theta_i \mid others; data) \propto \frac{\exp[\theta_i \sum_{j=1}^{J} \gamma_j x_{ij}]}{\prod_{j=1}^{J}[1+\exp(\gamma_j(\theta_i-\delta_j-\beta_i\phi_j))]} \exp\Big(-\frac{\theta_i^2}{2\tau_\theta}\Big).$$

b) Given the other parameters and data, the full conditional distribution of δ_j is

$$(\delta_j \mid others; data) \propto \frac{\exp[-\delta_j\gamma_j \sum_{i=1}^{n} x_{ij}]}{\prod_{i=1}^{n}[1+\exp(\gamma_j(\theta_i-\delta_j-\beta_i\phi_j))]} \exp\Big(-\frac{\delta_j^2}{2\tau_\delta}\Big).$$

c) Given the other parameters and data, the full conditional distribution of γ_j is

$$(\gamma_j \mid others; data) \propto \frac{\exp[\gamma_j \sum_{i=1}^{n} x_{ij}(\theta_i-\delta_j-\beta_i\phi_j)]}{\prod_{i=1}^{n}[1+\exp(\gamma_j(\theta_i-\delta_j-\beta_i\phi_j))]} \gamma_j^{a_\gamma-1} \exp(-b_\gamma\gamma_j).$$

d) Given the other parameters and data, the full conditional distribution of τ's is an inverse gamma distribution,

$$\begin{aligned} (\tau_\theta \mid others; data) &\sim \text{IG}\Big(\frac{n}{2}+a_\theta,\ \frac{1}{2}\sum_{i=1}^{n}\theta_i^2+b_\theta\Big), \\ (\tau_\delta \mid others; data) &\sim \text{IG}\Big(\frac{J}{2}+a_\delta,\ \frac{1}{2}\sum_{j=1}^{J}\delta_i^2+b_\delta\Big), \\ (\tau_\phi \mid others; data) &\sim \text{IG}\Big(\frac{J-2}{2}+a_\phi,\ \frac{1}{2}(\phi_j-2\phi_{j-1}+\phi_{j-2})^2+b_\phi\Big). \end{aligned}$$

e) Given the other parameters and data, the full conditional distribution of β_i is a discrete distribution,

$$\begin{aligned} & P(\beta_i = 1 \mid others; data) \\ = & \frac{\prod_{j=1}^{J} \frac{\exp(x_{ij}\gamma_j(\theta_i - \delta_j - \phi_j))}{1+\exp(\gamma_j(\theta_i - \delta_j - \phi_j))} p_\beta}{\prod_{j=1}^{J} \frac{\exp(x_{ij}\gamma_j(\theta_i - \delta_j - \phi_j))}{1+\exp(\gamma_j(\theta_i - \delta_j - \phi_j))} p_\beta + \prod_{j=1}^{J} \frac{\exp(x_{ij}\gamma_j(\theta_i - \delta_j))}{1+\exp(\gamma_j(\theta_i - \delta_j))}(1 - p_\beta)}. \end{aligned}$$

f) Given the other parameters and data, the full conditional distribution of p_β is

$$(p_\beta \mid others; data) \sim \text{beta}(\textstyle\sum_{i=1}^{n} \beta_i + a_p,\ n - \sum_{i=1}^{n} \beta_i + b_p).$$

g) Given the other parameters and data, the full conditional distribution of ϕ_j is

$$(\phi_j \mid others; data) \propto \frac{\exp[-\phi_j \gamma_j \sum_{i=1}^{n} \beta_i x_{ij}]}{\prod_{i=1}^{n}[1 + \exp(\gamma_j(\theta_i - \delta_j - \beta_i \phi_j))]} \exp\Big(- \frac{(\phi_j - \mu_{\phi_j})^2}{2\sigma_{\phi_j}^2}\Big),$$

where $\mu_{\phi_j} = \frac{\sum_{j \neq k} v_{jk}\phi_j}{v_{jj}}$ and $\sigma_{\phi_j}^2 = \frac{1}{v_{jj}}$ $(j, k = 3, 4, \cdots, J)$, and v_{jk} is the element in the $(j-2)$th row and $(k-2)$th column of the matrix $\mathbf{V}_\phi/\tau_\phi$.