

## A SPEEDED ITEM RESPONSE MODEL: LEAVE THE HARDER TILL LATER

YU-WEI CHANG

NATIONAL TSING-HUA UNIVERSITY

RUNG-CHING TSAI

NATIONAL TAIWAN NORMAL UNIVERSITY

NAN-JUNG HSU

NATIONAL TSING-HUA UNIVERSITY

A speeded item response model is proposed. We consider the situation where examinees may retain the harder items to a later test period in a time limit test. With such a strategy, examinees may not finish answering some of the harder items within the allocated time. In the proposed model, we try to describe such a mechanism by incorporating a speeded-effect term into the two-parameter logistic item response model. A Bayesian estimation procedure of the current model using Markov chain Monte Carlo is presented, and its performance over the two-parameter logistic item response model in a speeded test is demonstrated through simulations. The methodology is applied to physics examination data of the Department Required Test for college entrance in Taiwan for illustration.

Key words: item response model, Markov chain Monte Carlo, test speededness.

### 1. Introduction

Most achievement tests are administered within an allocated time. Given the time limit, test takers might not have enough time to finish answering all the test items. As a result, examinees might be forced to guess or skip some items, or their test performance might be affected by the feeling of time pressure. Test speededness is said to occur under the above situations in achievement tests (van der Linden, 2011). For simplicity, such time constraint is usually ignored in the traditional item response theory (IRT) models in which it is assumed that examinees have enough time to answer all the items. If the test performance of some examinees are affected by the time limit and the data are analyzed using traditional IRT models, the location, scale, and ability parameter estimates will be biased (Evans & Reilly, 1972; Oshima, 1994). Such bias might further impair the efficiency or optimality of some procedures utilized in adaptive testing, multistage testing, or test equating (for example, Bridgeman & Cline, 2004; Kingston & Dorans, 1984; van der Linden, Breithaupt, Chuah, & Zhang, 2007; Wollack, Cohen, & Wells, 2003). In addition, the local independence assumption of IRT models will often be violated if ability is the only latent trait considered to affect the test performance in a speeded test (Yamamoto & Everson, 1997; Yen, 1993). More complex IRT models which take into consideration the response process or possible mechanism adopted by examinees who do not have enough time to answer all the items would be necessary to overcome such discrepancies between the traditional IRT models and examinees' test behavior. In fact, a few parametric models have been proposed to characterize the effect of speededness in time limit tests.

Requests for reprints should be sent to Yu-Wei Chang, Institute of Statistics, National Tsing-Hua University, No. 101, Sec. 2, Kuang-Fu Road, Hsinchu 30013, Taiwan. E-mail: [ywchang1225@gmail.com](mailto:ywchang1225@gmail.com)

Bolt, Cohen, and Wollack (2002) utilize the mixture Rasch model (MRM, Rost, 1990) to reduce the bias of difficulty parameter estimates from the Rasch model for the speededness condition. They assume that some examinees are able to answer all the questions whereas the other group of examinees who belong to the so-called speeded class may not give their best performances toward the end of the test due to the time limit. Accordingly, the difficulty parameter estimates are allowed to differ for those end-of-test items between these two classes. Moreover, ordinal constraints are imposed on the difficulty parameters of the two classes to characterize the speededness effect. The MRM with ordinal constraints approach shows that using difficulty parameter estimates from the nonspeeded class, rather than from all examinees, could eliminate the bias coming from test speededness.

Based on the idea of mixture models, Cao and Stokes (2008) propose the IRT continuous guessing model (IRT-CG), a mixture of two-parameter logistic (2PL) IRT models, to characterize the differences between the motivated and the unmotivated classes in low-stakes tests, where test consequences are not crucial to test-takers. It is assumed that motivated examinees make their best effort to answer all the questions while the unmotivated ones answer items according to the item ordering and make less effort compared to their motivated counterparts. To accommodate this mechanism, the location parameters in the 2PL model are allowed to differ between the two classes. Although IRT-CG was originally designed for low-stakes tests to account for different test behavior, it can also be adopted to model the speededness effect assuming that the worse performance of examinees in the unmotivated class, which should be called the speeded class here, is caused by time pressure rather than low motivation.

The assumption, in both the MRM with ordinal constraints approach and IRT-CG, that examinees answer items according to item ordering, and consequently that speededness only affects the end-of-test items can be traced back to the HYBRID model (Yamamoto, 1995). The HYBRID model hypothesizes that some examinees answer questions according to item ordering up to an examinee-specific threshold and guess the remainders due to the time limit, while others answer all the items. Responses to the answered items for examinees of either case are characterized by 2PL models. In contrast to those mixture IRT model approaches where the speededness effect is assumed to be the same for all examinees within the speeded class, the examinee-specific threshold in the HYBRID model allows the number of items affected by test speededness to differ from person to person. Models capable of dealing with individual differences in the degree of speededness are considered to provide more information on understanding examinees' test behavior.

Following the idea of examinee-specific thresholds and assuming items are answered by their orderings, Goegebeur, De Boeck, Wollack, and Cohen (2008) propose a speeded IRT model with gradual process change. For brevity and simplicity, their model is referred to as IRT-GPC here. It is presumed that examinees answer items from the beginning. Once they feel that there is not enough time left to answer the rest of the items, they choose to answer only some of the remaining ones. The probability of an item ending up being picked and answered is modeled in IRT-GPC, and the gradual process change refers to the feature that the later the item ordering is, the smaller the probability of being answered will be. In other words, every item with its ordering beyond the examinee's threshold has some probability of not being in the solving process due to the time limit. The idea of modeling the probability of getting into the solving process for an item to capture the effect of speededness is very different from the previous studies in which speededness effect is directly built into the probability of correctly answering the item.

Instead of comparing the threshold to item ordering such as that in HYBRID and IRT-GPC, the IRT difficulty-based guessing model (IRT-DG; Cao & Stokes, 2008) compares the examinee-specific threshold to location parameters to determine the set of items that might be affected by low motivation in low-stakes tests. In fact, this model can be further adopted to model speededness effect in high-stakes tests via the mechanism that examinees in one class answer the easier

items but guess those harder ones due to insufficient test time; whereas examinees in the other class answer faster, and thus are not affected by the time limit. In IRT-DG, it is assumed that the difference between the threshold and ability parameter for each examinee is the same, and this difference is described by a parameter. This specification is very simple but may be restrictive for some real cases.

A concept similar to the threshold in IRT-DG has been suggested even earlier in Bejar (1985). Bejar considers the mechanism that once a test taker does not exactly know the answer to an item, he or she first leaves the item until later. At the second round, he or she answers items with more certainty from those left ones. The process is repeated until all the items are answered or test time is run out. Bejar (1985) uses this mechanism to explain certain discrepancies between the data and the traditional IRT model fittings, but does not formulate this mechanism by a parametric model. In contrast to comparing the examinee-specific threshold to the location parameters in IRT-DG, Bejar (1985) compares it to the degree of certainty to each item. These two approaches coincide if the degree of certainty can be exactly characterized by the difference between the ability parameter and the location parameter.

In the present study, a new speeded IRT model is proposed based on the mechanism that examinees answer the easier items first and retain the harder ones to a later test period in order to achieve higher scores. The idea of leaving some items till later is similar to Bejar's story. We compare our examinee-specific thresholds to the location parameters, and these thresholds are all free to be estimated in a more flexible fashion than that in IRT-DG. In addition, the concepts of gradual process and modeling the probability of getting into the solving process for an item in Goegebeur et al. (2008) are adopted in the proposed model. Details for the proposed model are presented in the next section. The differences between our model and previous studies are also discussed. A Bayesian analysis for model estimation through Markov chain Monte Carlo (MCMC) is described in Section 3. Some simulations are conducted to demonstrate the validation of the Bayesian estimation procedure in Section 4. An application to the entrance examination data is illustrated in Section 5, followed by some concluding remarks in Section 6.

## 2. Leave-the-Harder-till-Later Speeded Item Response Model

In this section, we introduce the leave-the-harder-till-later speeded Rasch model (abbreviated as LHL-Rasch) and its 2PL extension (denoted as LHL-2PL). The LHL-Rasch model, as an extension of the Rasch model, is designed for the situation in which some of examinees may not have enough time to answer all the items within a given time. We hypothesize that, while facing tests within an allocated time, an examinee tends to answer easier items first and leave items with a certain level of difficulty till later. After answering easier items, the examinee would attempt some of the first-skipped items, and eventually some of the first-skipped and attempted ones may be correctly answered. Probabilities of the first-skipped items being attended to later are modeled, and the harder the item is, the smaller the probability will be. If an examinee does not have enough time to attend to all the first-skipped items due to the time limit, some items will be left blank at the end of the test. Whenever an item, regardless of first-skipped or not, is put into the solving process, its response is characterized by the Rasch model. Formulations of the LHL-2PL and LHL-Rasch models are described in Section 2.1. In general, we refer to our speeded model as LHL-IRT, indicating either case. An alternative representation of the proposed model is given in Section 2.2 to provide more insight into the model. Moreover, the connections between our proposed model and related models are also discussed.

### 2.1. Leave-the-Harder-till-Later Speeded 2PL Model

Let  $Y_{pj}$  be the dichotomous response of examinee  $p$  on item  $j$ , where  $p = 1, 2, \dots, P$ , and  $j = 1, 2, \dots, J$ . We denote  $b_j$  and  $a_j$  as the location and scale parameters, respectively,

for item  $j$ , and  $\theta_p$  as the ability parameter for examinee  $p$  in the 2PL model. In the following, we introduce LHL-2PL directly and regard LHL-Rasch as its special case. Under LHL-2PL, the probability of examinee  $p$  obtaining a correct response on item  $j$  is a product of two terms, the probability of getting a correct response under the 2PL model and the probability of getting into the solving process affected by the test speededness, that is,

$$P(Y_{pj} = 1|a_j, b_j, \theta_p, \tau_p, \lambda) = \pi_{pj} \quad (1)$$

with

$$\pi_{pj} = \frac{1}{1 + e^{-a_j(\theta_p - b_j)}} \cdot e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}, \quad (2)$$

where  $\tau_p$  is the examinee-specific threshold parameter for the speededness effect of examinee  $p$ ;  $\lambda$ , larger than zero, is the overall speededness rate; and  $I\{\cdot\}$  is the indicator function. For convenience, the term  $e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}$  is referred to as the *speeded term* in this article. Though the location parameter is no longer a direct indication of item difficulty in the 2PL model, the term LHL is preserved for our 2PL version of the speeded model. For brevity, the term *difficult/easy* is used instead of large/small location parameter as we describe the scenario in the 2PL settings. LHL-Rasch is simply a special case with  $a_j = 1$ , for  $j = 1, 2, \dots, J$ , in LHL-2PL.

Parameter  $\tau_p$  represents, for examinee  $p$ , the location threshold. Items with location parameters exceeding  $\tau_p$  will be regarded as not-so-easy and requiring considerable amount of time to be solved, or as so difficult and simply having no idea how to solve at the first glance. Test strategy considered in the present model is that examinees do not put such items into the problem solving process until all the easier items are answered already. In other words, examinee  $p$  would directly try to answer all the items with location parameters smaller than  $\tau_p$  and retain those harder items to a later period of the test. We refer to the examinee-specific parameter  $\tau_p$  as the speededness point for examinee  $p$ .

It is further assumed that a particular first retained item  $j$  would be attended or answered, for examinee  $p$ , with probability  $e^{-\lambda(b_j - \tau_p)}$ . That is, the harder the retained item is, the smaller the probability will be of that item being in the solving process. The decay rate of  $e^{-\lambda(b_j - \tau_p)}$  is determined by the speededness rate  $\lambda$ . When  $b_j$  exceeds  $\tau_p$ , a larger  $\lambda$  brings a smaller probability of examinee  $p$  answering item  $j$  than a smaller  $\lambda$ . To sum up, when the location parameter  $b_j$  is smaller than the examinee-specific threshold  $\tau_p$ , the probability of examinee  $p$  answering item  $j$  equals 1; otherwise, the probability equals  $e^{-\lambda(b_j - \tau_p)}$ . Hence, for examinee  $p$ , the probability of putting item  $j$  into the solving process is written as  $e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}$ , the speeded term.

Let  $\mathbf{b} = (b_1, \dots, b_J)$ ,  $\mathbf{a} = (a_1, \dots, a_J)$ ,  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_P)$ ,  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_P)$ ,  $\mathbf{y}_p = (y_{p1}, \dots, y_{pJ})$  and  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_P)$ . We assume that, given  $\mathbf{b}, \mathbf{a}, \boldsymbol{\theta}, \boldsymbol{\tau}$ , and  $\lambda$ , responses over all items for examinee  $p$  are conditionally independent, i.e.,

$$f(\mathbf{y}_p | \mathbf{a}, \mathbf{b}, \theta_p, \tau_p, \lambda) = \prod_{j=1}^J f(y_{pj} | a_j, b_j, \theta_p, \tau_p, \lambda) = \prod_{j=1}^J \pi_{pj}^{y_{pj}} (1 - \pi_{pj})^{1 - y_{pj}}. \quad (3)$$

Furthermore, responses from different examinees are assumed to be independent, conditional on  $\mathbf{b}, \mathbf{a}, \boldsymbol{\theta}, \boldsymbol{\tau}$ , and  $\lambda$ .

Both  $\theta_p$  and  $\tau_p$  are parameters about examinees, and they are further assumed to be bivariate normally distributed:

$$\begin{pmatrix} \theta_p \\ \tau_p \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho\sigma_\theta\sigma_\tau \\ \rho\sigma_\theta\sigma_\tau & \sigma_\tau^2 \end{pmatrix} \right), \quad (4)$$

where  $\theta_p$  and  $\tau_p$  are allowed to be correlated with correlation  $\rho$ . For model identification purpose, the marginal distribution of  $\theta_p$  is set to be  $N(0, 1)$  under LHL-2PL and set to be  $N(0, \sigma_\theta^2)$  under LHL-Rasch.

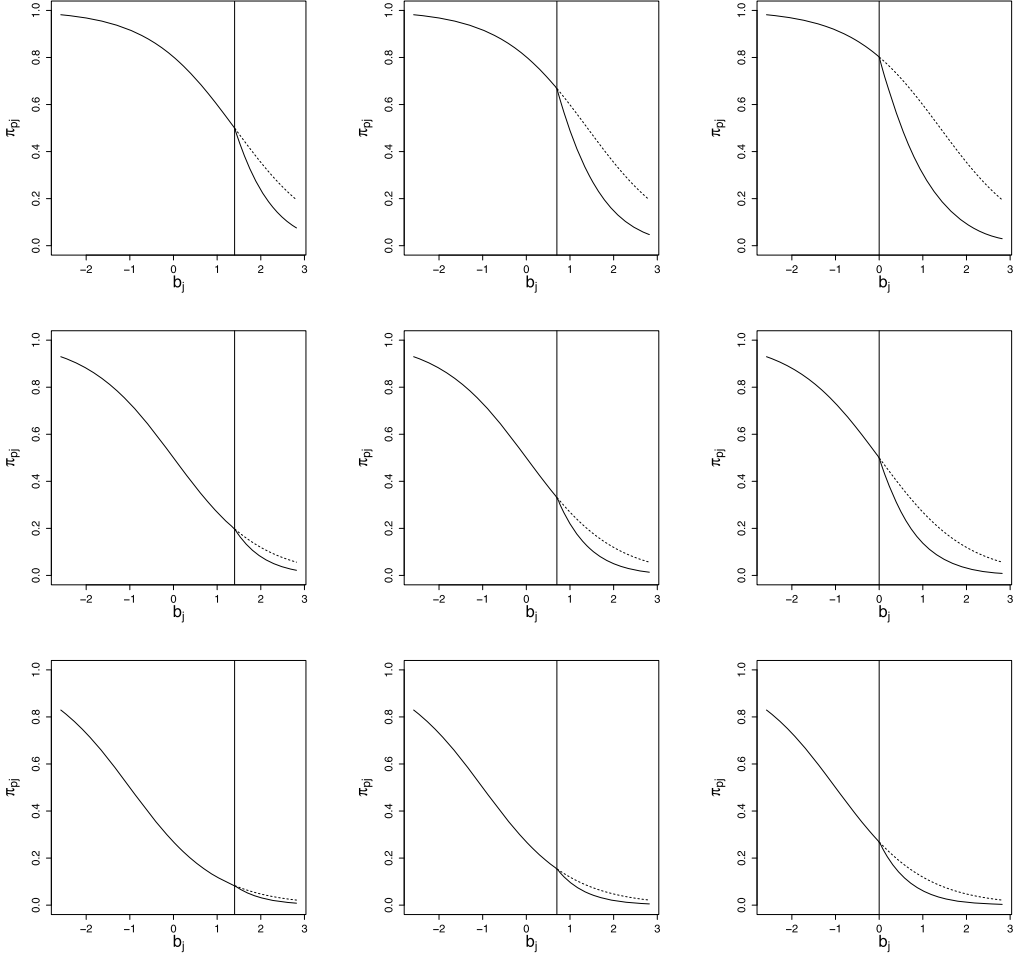


FIGURE 1.

The  $\pi_{pj}$  (solid line) and  $\pi_{pj}$  with the speeded term excluded (broken line) against  $b_j$ , for 9 hypothesized examinees. The  $\theta_p$  equals to 1.4 (the first row), 0 (the second row) and  $-1$  (the third row). The  $\tau_p$  equals to 1.4 (the first column), 0.7 (the second column) and 0 (the third column), and the value of  $\tau_p$  is indicated by the vertical line.

Another identification issue is regarding the specification of  $\lambda$ . Whenever  $b_j$  is smaller than  $\tau_p$  for all examinees  $p$  and all items  $j$ , it implies that no test speededness occurs, and the LHL-IRT model reduces to the IRT model. On the other hand, LHL-IRT with  $\lambda = 0$  also reduces to the IRT model. Therefore, zero is excluded from the range of  $\lambda$  under LHL-IRT to prevent model unidentifiability.

To visualize the effect of the speeded term, the  $\pi_{pj}$  with or without the speeded term under LHL-Rasch is plotted against  $b_j$  in Figure 1, under different levels of  $\theta_p$  and  $\tau_p$ . A larger  $\theta_p$  or a smaller  $\tau_p$  makes a more significant difference in  $\pi_{pj}$  between LHL-Rasch and Rasch, and consequently a greater bias in estimation may occur if we mis-specify LHL-Rasch as Rasch.

## 2.2. Some Notes on the LHL-IRT Model

LHL-IRT is compared to several existing models in this subsection. First of all, a two-stage representation for LHL-IRT via Bernoulli random variables is formulated. Based on this representation, LHL-IRT is further compared to IRT-GPC. Secondly, some connections between

LHL-IRT, IRT-GPC and models capable of dealing with nonresponse data are illustrated. At last, the differences between the present model, IRT-DG and the story of Bejar (1985) are explained.

Similar to the idea in Goegebeur et al. (2008), getting a correct response in the present model can be decomposed into a two-stage procedure. At the first stage, a Bernoulli trial with success probability  $e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}$  determines, for examinee  $p$ , whether to solve or leave blank item  $j$  due to test speededness. In the case of problem solving, the probability of a correct response is modeled by a 2PL model. In contrast, leaving blank directly leads to a wrong answer. Formally speaking, let  $Z_{pj} = 1$  denote the event that examinee  $p$  attempts to answer item  $j$ , and let  $Z_{pj} = 0$  be the event of not attempting item  $j$ . The two stages are stated as

$$\begin{aligned} Z_{pj} | b_j, \tau_p, \lambda &\sim \text{Bernoulli}(e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}), \\ Y_{pj} | a_j, b_j, \theta_p, Z_{pj} &\sim \text{Bernoulli}\left(\frac{1}{1 + e^{-a_j(\theta_p - b_j)}} \cdot Z_{pj}\right). \end{aligned}$$

This two-stage formulation is equivalent to (1) and (2) for the dichotomous response  $Y_{pj}$ , due to

$$\begin{aligned} E(Y_{pj} | a_j, b_j, \theta_p, \tau_p, \lambda) &= E(E(Y_{pj} | a_j, b_j, \theta_p, \tau_p, \lambda, Z_{pj})) \\ &= E\left(\frac{1}{1 + e^{-a_j(\theta_p - b_j)}} \cdot Z_{pj} \mid a_j, b_j, \theta_p, \tau_p, \lambda\right) \\ &= \frac{1}{1 + e^{-a_j(\theta_p - b_j)}} \cdot e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}. \end{aligned}$$

Though the two-stage point of view in LHL-IRT is similar to that in IRT-GPC (Goegebeur et al., 2008), the mechanism described in the present model is quite different from that in IRT-GPC. In IRT-GPC, it is assumed that items are answered in their orderings, and the affected or skipped items are those with greater item ordering. However, in the present model, the situation that more *difficult* items would be skipped first while all the *easier* ones are answered is accommodated. These different assumptions in the underlying mechanism result in different speeded terms in the two models. Speededness point in IRT-GPC is compared to item ordering whereas in LHL-IRT it is compared to ordered location parameters. The strategy of introducing a two-stage process to a traditional IRT model is similar to that in nonignorable nonresponse modeling, while the variable  $Z_{pj}$  indicating whether an item undergoes the problem solving process is unobserved in our work, in contrast to being observed in nonignorable nonresponse modeling (Holman & Glas, 2005; O’Muircheartaigh & Moustaki, 1999). Glas and Pimentel (2008) further utilize such a nonresponse model for a sequence of missing responses in the last few items due to test speededness.

The idea of comparing examinee-specific threshold  $\tau_p$  to the location parameters is similar to IRT-DG. However, the threshold here is freely estimated, in contrast to being set to be a constant plus  $\theta_p$  in IRT-DG. Besides, the situations that the two models try to accommodate are quite different. In IRT-DG, examinees belonging to the unmotivated or speeded class guess whenever an item is *harder* than their thresholds. The present model characterizes that for an item with its location parameter exceeding one’s speededness point, the probability of getting to solve that item could range from 0 to 1. Once the item is in the problem solving process, the test taker tries his or her best to answer the item. The two-stage Bernoulli trail point of view for the LHL-IRT model makes this difference easier to be caught on. Moreover, the assumption that all the first-skipped items are possible to be attempted by test takers in LHL-IRT also makes the LHL scenario different from the idea of Bejar (1985).

## 3. Estimation Procedure

In the proposed model, there is a more complex likelihood structure than that in the traditional IRT models due to the inclusion of the examinee-specific LHL speeded term. We perform a Bayesian analysis for the proposed speeded model using the MCMC techniques. In particular, a two-layer hierarchical prior is assumed for the model parameters to reduce the impact of the prior settings on the posterior inference (Fox, 2010). The detailed prior settings and the estimation procedures for the LHL-2PL model are given in this section, and they can be easily modified for LHL-Rasch.

We start with the first-layer prior settings for the parameters, including  $\{b_j: j = 1, 2, \dots, J\}$ ,  $\{a_j: j = 1, 2, \dots, J\}$ ,  $\lambda$ ,  $\mu_\tau$ ,  $\sigma_\tau^2$  and  $\rho$ . The location parameters  $b_j$ 's are assumed to follow normal prior distributions (Swaminathan & Gifford, 1986). The scale parameters  $a_j$ 's and the speededness rate parameter  $\lambda$  are all positive and assumed to follow exponential prior distributions. That is,

$$\begin{aligned} b_j &\sim N(\mu_b, \sigma_b^2), \\ a_j &\sim \text{Exp}(\beta_a), \\ \lambda &\sim \text{Exp}(\beta_\lambda), \end{aligned}$$

where  $\mu_b$ ,  $\sigma_b^2$ ,  $\beta_a$  and  $\beta_\lambda$  are hyperparameters.

For parameters  $\mu_\tau$ ,  $\sigma_\tau^2$ , and  $\rho$ , which characterize the joint distribution of examinee's ability and speededness points, we use standard prior settings for  $(\mu_\tau, \sigma_\tau^2)$ :

$$\mu_\tau \sim N(\mu_0, \sigma_0^2), \quad \sigma_\tau^2 \sim \text{Inv-Gamma}(\alpha_0, \beta_0),$$

where  $\mu_0$ ,  $\sigma_0^2$ ,  $\alpha_0$ , and  $\beta_0$  are hyperparameters. The prior of the correlation  $\rho \in (-1, 1)$  is specified through its Fisher's  $z$  transformation to ensure good behavior at the boundary under the random walk Metropolis–Hastings scheme (used later in the Bayesian analysis), i.e.,

$$\zeta \equiv \log\left(\frac{1+\rho}{1-\rho}\right) \sim N(\mu_\zeta, \sigma_\zeta^2),$$

and equivalently,

$$\pi(\rho) = \frac{1}{\sqrt{2\pi}\sigma_\zeta} \exp\left\{-\frac{1}{2\sigma_\zeta^2} \left(\log \frac{1+\rho}{1-\rho} - \mu_\zeta\right)^2\right\} \frac{2}{(1+\rho)(1-\rho)}.$$

Given the first-layer prior, we specify the second-layer prior for all of the hyperparameters in the following:

$$\begin{aligned} \mu_b &\sim N(\mu_2, \sigma_2^2), \\ \sigma_b^2 &\sim \text{Inv-Gamma}(\alpha_1, \beta_1), \\ \beta_a &\sim \text{Inv-Gamma}(\alpha_2, \beta_2), \\ \beta_\lambda &\sim \text{Inv-Gamma}(\alpha_3, \beta_3), \\ \mu_0 &\sim N(\mu_1, \sigma_1^2), \\ \sigma_0^2 &\sim \text{Inv-Gamma}(\alpha_6, \beta_6), \\ \beta_0 &\sim \text{Gamma}(\alpha_4, \beta_4), \\ \mu_\zeta &\sim N(\mu_3, \sigma_3^2), \\ \sigma_\zeta^2 &\sim \text{Inv-Gamma}(\alpha_5, \beta_5). \end{aligned}$$

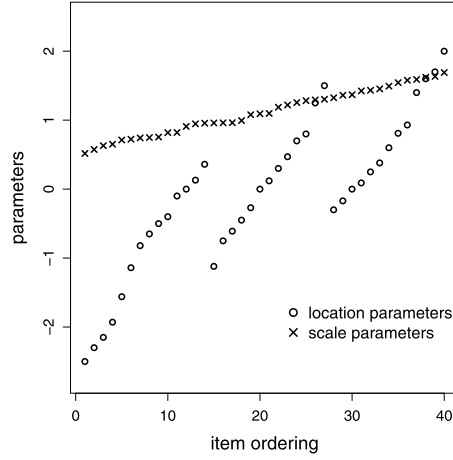


FIGURE 2.

True values for the location and scale parameters in the simulation study.

Basically, we use normal priors for location-type hyperparameters, inverse gamma priors for scale-type hyperparameters, and gamma priors for rate-type hyperparameters. Since the prior on  $\beta_0$  provides fair flexibility for the prior  $\pi(\sigma_\tau^2)$  already, the shape hyperparameter  $\alpha_0$  is fixed at 2.5 for convenience. All parameters in the second-layer priors, including  $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ ,  $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2)$ ,  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_6)$ , and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_6)$ , are assigned in a reasonable way. We further assume that all the priors are independent.

Let  $\boldsymbol{\xi}_1 = (\mathbf{a}, \mathbf{b}, \lambda, \mu_\tau, \sigma_\tau^2, \rho)$  and  $\boldsymbol{\xi}_2 = (\mu_b, \sigma_b^2, \beta_a, \beta_\lambda, \mu_0, \sigma_0^2, \beta_0, \mu_\zeta, \sigma_\zeta^2)$ . The joint posterior of  $\boldsymbol{\xi}_1$ ,  $\boldsymbol{\theta}$ ,  $\boldsymbol{\tau}$ , and  $\boldsymbol{\xi}_2$  given the data  $\mathbf{y}$  satisfies

$$f(\boldsymbol{\xi}_1, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}_2 | \mathbf{y}) \propto f(\mathbf{y} | \boldsymbol{\xi}_1, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}_2) f(\boldsymbol{\theta}, \boldsymbol{\tau} | \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \pi(\boldsymbol{\xi}_1 | \boldsymbol{\xi}_2) \pi(\boldsymbol{\xi}_2), \quad (5)$$

where the sampling distribution  $f(\mathbf{y} | \boldsymbol{\xi}_1, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}_2)$  is specified in (1), (2), and (3),  $f(\boldsymbol{\theta}, \boldsymbol{\tau} | \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$  is defined in (4),  $\pi(\boldsymbol{\xi}_1 | \boldsymbol{\xi}_2)$  is the first-layer prior, and  $\pi(\boldsymbol{\xi}_2)$  is the second-layer hyperprior. Since the joint posterior is high-dimensional, and its form is nonstandard, we approximate the posterior using the MCMC method via implementing the Metropolis–Hastings algorithm under which the samples are sequentially drawn for  $(\boldsymbol{\xi}_1, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}_2)$  from their full conditionals. The full conditionals and the detailed sampling schemes are given in the [Appendix](#). The performance of this estimation procedure is demonstrated in Section 4 through a simulation study.

#### 4. Simulation Study

To evaluate the performance of the Bayesian inference under the proposed LHL-IRT model, two simulations were conducted based on data generated from the LHL scenarios, one for LHL-Rasch and the other for LHL-2PL. Furthermore, in order to investigate the possible estimation bias under model misspecification, for each simulated realization, both LHL-IRT and IRT models were fitted.

We first describe the simulation settings for LHL-2PL. There are 40 items with their location parameters  $b_j$ 's ranging from  $-3$  to  $2$  and their scale parameters  $a_j$ 's ranging from  $0.5$  to  $2$ . More specifically, the true values for the location and scale parameters are plotted in Figure 2. These values were chosen to represent the situation that there are several parts of questions in a test, and the items within each part were ordered according to their location parameters. Under this setting, some items with larger location parameters would appear before items with smaller



TABLE 1.

RMSE of estimates from LHL-Rasch fitting and Rasch fitting under data generated from the LHL-Rasch model (10 replicates).

Parameter	$P = 250$		$P = 500$		$P = 1,000$	
	LHL	Rasch	LHL	Rasch	LHL	Rasch
$\mathbf{b}$	0.2243	0.5453	0.1700	0.5539	0.1094	0.5860
$\theta$	0.3742	0.4240	0.3682	0.4243	0.3646	0.4303
$\tau$	0.8823	–	0.5877	–	0.4347	–

location parameters in the whole test. Therefore, in this case, one can distinguish between the GPC and LHL mechanisms. As for the scale parameters in the 2PL and LHL-2PL models, for items in the same part, the larger the location parameter is, the larger the scale parameter will be. Regarding the examinee related parameters, the parameters associated with  $\tau_p$  were specified as  $\mu_\tau = 0.2$  and  $\sigma_\tau^2 = 0.5$ , the threshold parameter  $\tau_p$  was set to be positively correlated with the ability parameter  $\theta_p$  with correlation  $\rho = 0.8$ , and the speededness rate  $\lambda$  was set to 1. The values of  $(\lambda, \mu_\tau, \sigma_\tau^2)$  were chosen such that the speeded effect was of moderate size. For example, the expected number of items that one examinee solves, averaged over all examinees, was about 31 items out of the total of 40 items.

The simulation was performed under three different sample sizes of examinees,  $P = 250$ , 500, and 1,000. In each case, ten sets of independent replications were simulated according to the above model settings, and the Bayesian analysis described in Section 3 was applied to each replicate with the following second-layer prior settings:  $\mu = (0, 0, 0)$ ,  $\sigma^2 = (1, 10, 1)$ ,  $\alpha = (3, 3, 3, 2.5, 3, 3)$ , and  $\beta = (3, 8, 8, 2, 10, 8)$ . In the LHL-Rasch case, all settings were identical to those for LHL-2PL, except that  $\{a_j\}$  were all fixed at 1, and  $\sigma_\theta^2$  was set to be freely estimated. For comparison, we also adopted the Bayesian estimation for the traditional IRT model using the prior settings and posterior sampling scheme used for LHL-IRT. The identifiability constraints adopted here under IRT models were the same as their counterparts in LHL-IRT models, i.e.,  $(\mu_\theta, \sigma_\theta^2) = (0, 1)$  in 2PL and  $\mu_\theta = 0$  in Rasch.

The Bayesian analysis was implemented via MCMC schemes detailed in the Appendix, and the convergence was checked by the method of  $\hat{R}^{1/2}$  (Gelman & Rubin, 1992) based on five independent chains starting from different initial values. For most of the replications, convergence was achieved after about 10,000 iterations. For each parameter, the posterior mean was calculated as our Bayes estimates, based on 30,000 MCMC draws after burn-in for each replicate. All the computations were implemented using free statistical software R, and it took about 100 hours for each replicate of the LHL-2PL case with 1,000 examinees, on a PC with Intel Core i7 and 2 GB of RAM.

We used the posterior means as the point estimates for parameters of interest, denoted as  $\hat{\mathbf{a}}$ ,  $\hat{\mathbf{b}}$ ,  $\hat{\theta}$ ,  $\hat{\tau}$  and  $\hat{\lambda}$ . The estimation performance is evaluated by the root mean squared error (RMSE) of the posterior means, summarized in Tables 1 and 2, for LHL-Rasch and LHL-2PL, respectively. For brevity, we summarize the estimation performance for groups of parameters instead of for each individual parameter, where the parameter groups are formed according to their common characteristics in the model. More specifically, the performance in estimating  $a_j$ 's and  $b_j$ 's are summarized by their respective RMSE values, averaging over all items and replicates. Similarly, RMSE for  $\theta_p$ 's or  $\tau_p$ 's is obtained from averaging over all examinees and replicates. Furthermore, to visualize the bias for  $\{b_j\}$  more closely, the estimated location parameter averaged over replicates for each item is plotted against its true value in Figures 3 and 4. The estimates under the LHL-IRT model fitting are represented by circles, and the estimates under the IRT model fitting are represented by solid dots.

We first look at the results for the Rasch case. For data generated from LHL-Rasch, the identifiability constraints imposed on the ability distribution of the two fitted models are identical

TABLE 2.

RMSE of estimates from LHL-2PL fitting and 2PL fitting under data generated from the LHL-2PL model (10 replicates).

Parameter	$P = 250$		$P = 500$		$P = 1,000$	
	LHL	2PL	LHL	2PL	LHL	2PL
$\mathbf{b}$	0.2811	0.2770	0.1994	0.2380	0.1472	0.2063
$\mathbf{a}$	0.3297	0.4151	0.2258	0.3865	0.1606	0.4238
$\boldsymbol{\theta}$	0.3691	0.3596	0.3664	0.3622	0.3653	0.3579
$\boldsymbol{\tau}$	0.5304	—	0.5472	—	0.4906	—

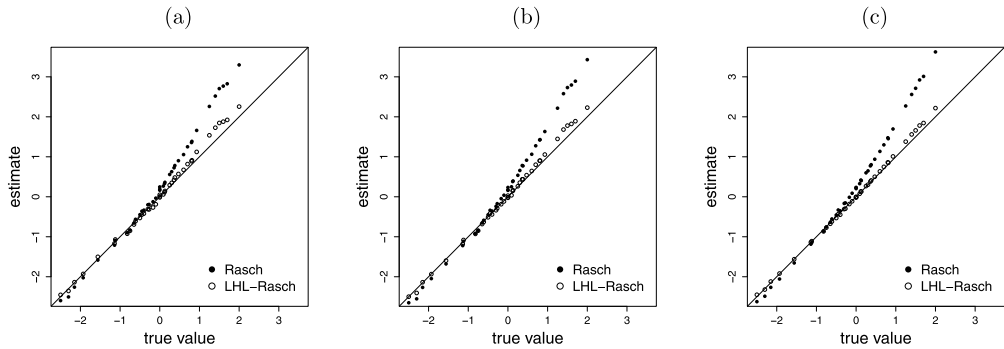


FIGURE 3.

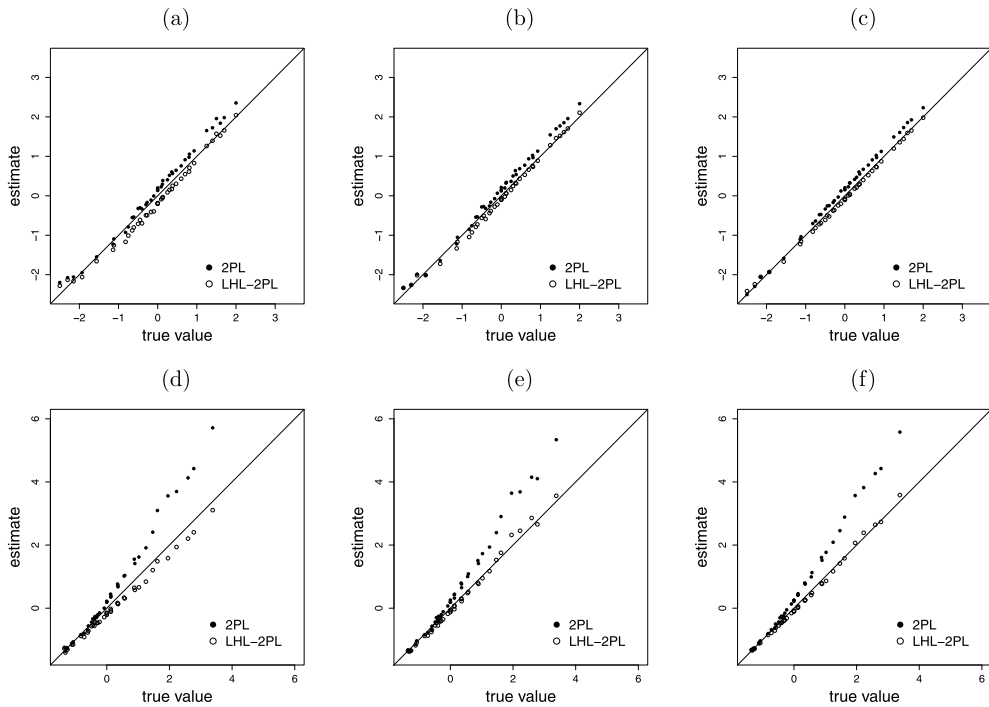
Estimate of  $b_j$  versus true value of  $b_j$ , averaged over 10 replications, under Rasch fittings (*solid dots*) and LHL-Rasch fittings (*circles*): (a)  $P = 250$ ; (b)  $P = 500$ ; (c)  $P = 1,000$ .

FIGURE 4.

Estimates of  $\tilde{b}_j$ 's versus true value of  $\tilde{b}_j$ 's, averaged over 10 replications, under 2PL fittings (*solid dots*) and LHL-2PL fittings (*circles*), where  $\tilde{b}_j = b_j$  in (a)–(c) and  $\tilde{b}_j = b_j \cdot a_j$  in (d)–(f): (a) (d)  $P = 250$ ; (b) (e)  $P = 500$ ; (c) (f)  $P = 1,000$ .

to the true value used in data generation. As a result, the estimates of  $b_j$ 's from the two models and their true values are thought to be of comparable scale. The bias, due to ignoring the LHL speededness scenario, of  $\{b_j\}$  associated with the Rasch model fitting is very pronounced, as shown in Figure 3, as well as the large RMSE given in Table 1. Moreover, for those location parameters with large true values, the larger the true value is, the larger the bias for the Rasch fitting will be. This is to be expected because for an item with a larger location parameter, the number of examinees with their thresholds smaller than the location parameter will be greater, and, therefore, the impact of speededness is greater. In addition, the RMSE and bias of location parameters reduce as the sample size increases for the LHL-Rasch fitting. It is also shown in Table 1 that the performance for estimating  $\theta_p$ 's is not much different between the two model fittings, which is not surprising due to the constraint on  $\{\theta_p\}$  for model identifiability.

Under the LHL-2PL setting, it is shown in Table 2 that for moderate and large data sets with  $P = 500$  and  $P = 1,000$ , the RMSE's for both parameter groups  $\{b_j\}$  and  $\{a_j\}$  under the 2PL fittings are larger than those under the LHL-2PL fittings, coming from bias due to ignoring the LHL speededness phenomenon. Compared to the LHL fitting, the 2PL fitting gives a positive bias for items with larger  $b_j$  on the original scale of location parameters, as shown in Figure 4(a)–(c). In this 2PL case, the estimates of  $a_j$  and  $b_j$  would influence each other, and, therefore, we may also examine the estimation bias on the logit scale (i.e.,  $b_j^* = a_j b_j$ ) to account for the effects of estimates of  $b_j$  and  $a_j$  simultaneously. As shown in Figure 4(d)–(f), the bias for estimating  $b_j^*$  due to ignoring the LHL speededness mechanism is even more pronounced under this logit scale than that from the original scale of location parameters.

To sum up, the Bayesian procedure yields sensible estimates for the LHL-IRT models. Moreover, based on our limited experiences, the advantage of fitting the LHL-2PL model against the conventional 2PL model is more pronounced for examinees of a large sample size (e.g.,  $P = 500$  or 1,000).

## 5. Application

We applied our methodology to the data of Department Required Test for college entrance in Taiwan. The data are responses on the physics examination for 1,000 examinees randomly sampled from a total of 35,357 examinees who took the test in 2010, provided by College Entrance Examination Center (CEEC). Examinees have to answer 26 questions, including 20 multiple-choice questions, 4 multiple-response questions and 2 calculation problems, within 80 minutes. In addition, it is a formula-score test, namely, some points are deducted for each incorrect answer. Based on such scoring scheme, examinees are less likely to guess whenever they do not know the answer (Lord, 1975). This provides some rationale for considering a speeded model in which random guessing is not allowed.

The data released by CEEC contain the original response and nonresponse information for each examinee and each item, for both the multiple-choice and multiple-response questions. We treat both nonresponses and incorrect answers the same way and code them as  $Y_{pj} = 0$ . Each calculation problem contains several sequential questions and is graded by professionals. An examinee will obtain 10 points if both the final answer and the solution process are correct, or otherwise only a partial credit will be granted. Only the total points (0–10) for each calculation problem are available from the data. Thus, considering the scoring scheme for the calculation problems, the response  $Y_{pj}$  is coded as 1 whenever the original score is more than 7.5 out of 10 points, and zero otherwise.

In order to obtain some evidence for the LHL mechanism in analyzing the physics examination data, we consider the following procedures before fitting LHL models. Under the LHL scenario, responses to the harder items are influenced by both test speededness and ability. It

is as if there were a “speeded factor” that also affects the test performance. Hence, it would be expected that additional local dependence exists among harder items beyond the traditional IRT fitting. To examine such possible local dependence, the traditional IRT fitting is compared to the multidimensional item response theory (MIRT) model fittings in terms of the likelihood and Akaike information criterion (AIC). More specifically, the confirmatory MIRT (see, for example, De Boeck & Wilson, 2004) is adopted, and only more difficult items are related to the second factor. The “difficulty” is determined by the location parameter estimate  $\hat{b}_j$  under the 2PL fitting. The MIRT provides a better goodness of fit compared to 2PL even after compensating the model complexity when the overall local dependence among the responses of enough difficult items can be captured by the second factor. For example, the log likelihood and AIC for the 2PL fitting are  $-12,907.103$  and  $25,918.207$ , respectively, whereas they are  $-12,886.105$  and  $25,894.211$  for the MIRT with the 8 most difficult items associated to the second factor. The comparison results suggest the existence of local dependence among harder items beyond the 2PL fitting, and provide some rationale to fit the LHL models.

Four models studied in the simulations are fitted to the data using Bayesian analysis. The posterior means for parameters, obtained from an MCMC chain of about 40,000 iterations in total and 30,000 iterations after burn-in, are used as the parameter estimates. We make further comparisons among these four models using the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), a Bayesian model selection criterion briefly described in the following. Let  $\xi_3 = (\mathbf{a}, \mathbf{b}, \boldsymbol{\theta}, \boldsymbol{\tau}, \lambda)$  and  $\hat{\xi}_3$  be the posterior mean of  $\xi_3$  given data  $\mathbf{y}$ . The DIC for a fitted model is defined as

$$DIC = D(\hat{\xi}_3) + 2p_D, \quad (6)$$

where

$$\begin{aligned} D(\hat{\xi}_3) &= -2 \log f(\mathbf{y} | \hat{\xi}_3), \\ p_D &= E_{\xi_3 | \mathbf{y}}[-2 \log f(\mathbf{y} | \xi_3)] - D(\hat{\xi}_3), \end{aligned}$$

in which the expectation in  $p_D$  is taken with respect to the posterior distribution of  $\xi_3$ . In (6), the first term  $D(\hat{\xi}_3)$  measures the goodness-of-fit, and the second term  $p_D$  represents the effective number of parameters used in the model. A smaller DIC is preferred, which selects a model with a better goodness-of-fit and simultaneously maintains the model complexity to be as simple as possible. The resulting DIC values for the four fitted models are listed in Table 3. The LHL-2PL is the model with the smallest DIC, indicating the best fitting performance among all the models after compensating for model complexity. The LHL-2PL and LHL-Rasch models provide better fit than their respective non-speeded counterparts and, therefore, the leave-the-harder-till-later mechanism is thought to describe the data better. On the other hand, the amount of DIC reduction from LHL-2PL to LHL-Rasch is greater than that from LHL-2PL to 2PL, indicating that the scale parameters also provide a flexible feature to improve the fitting for this data set.

The estimates  $\hat{\theta}_p$  and  $\hat{\tau}_p$  for all examinees ( $P = 1,000$ ) from the LHL-2PL fitting are plotted in Figure 5(a). In addition, the estimate  $\hat{\theta}_p$  versus the length of 95 % posterior interval for  $\tau_p$  are shown in Figure 5(b). For examinees with higher ability,  $\hat{\theta}_p$  and  $\hat{\tau}_p$  are positively correlated, as shown in Figure 5(a), while the correlation is about zero for other examinees. This result coincides with the intuition that, for the better performance group, if the ability of an examinee is higher, the examinee-specific threshold, which is compared to location parameters for items considered to be difficult and skipped first, will be greater. In contrast, for an examinee in the worse performance group, the probability of correctly answering an item according to the 2PL model is low and, therefore, in LHL-2PL the speeded term makes little influence on  $\pi_{pj}$ , similar to what we note in Section 2.1 and Figure 1. In other words, for examinees with lower ability, the data provide less information for the thresholds, and, thus, the posterior intervals are expected to

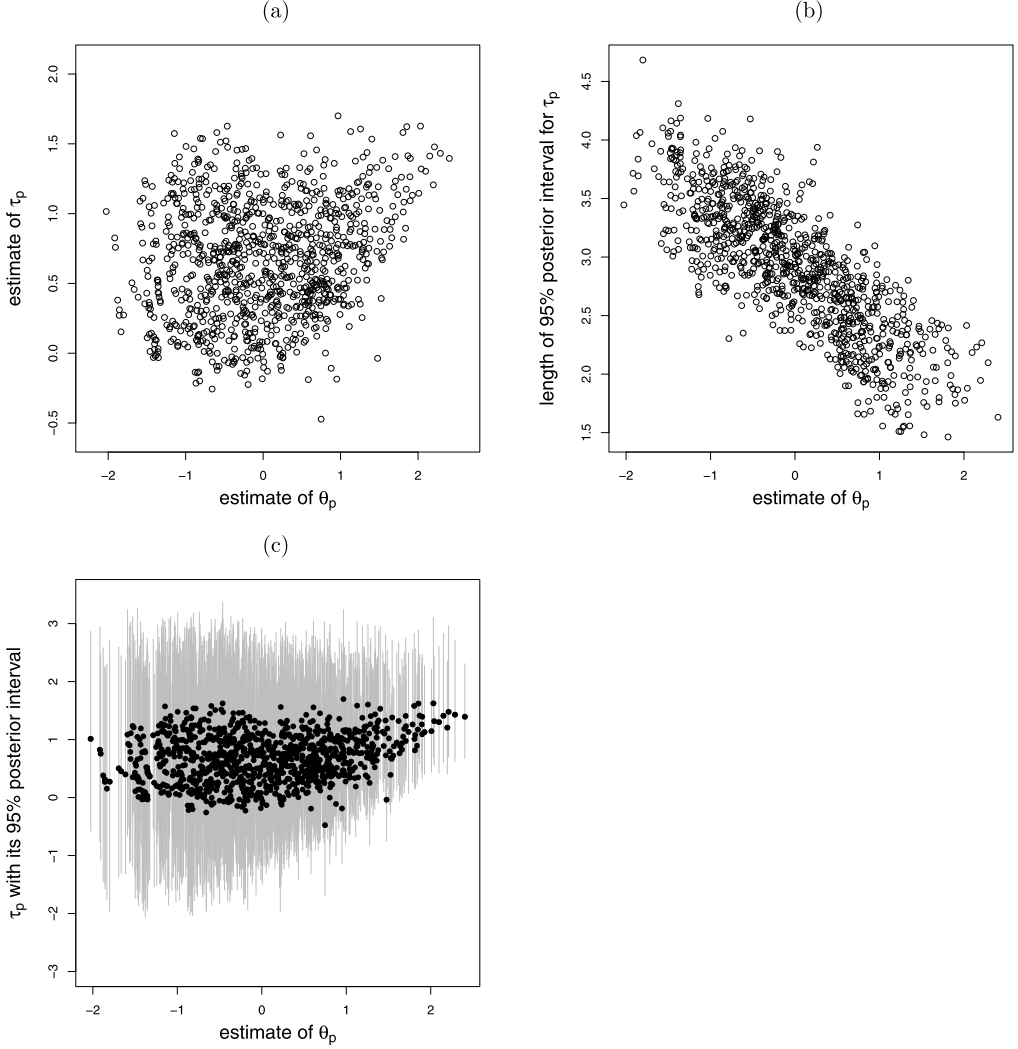


FIGURE 5.

Fitting LHL-2PL to the physics examination data: (a)  $\hat{\tau}_p$  versus  $\hat{\theta}_p$ ; (b) length of 95 % posterior interval of  $\tau_p$  versus  $\hat{\theta}_p$ ; (c)  $\hat{\tau}_p$  (solid dots) and 95 % posterior interval of  $\tau_p$  (gray bars) versus  $\hat{\theta}_p$ .

be wider, as shown in Figure 5(b). The estimates  $\hat{\theta}_p$ ,  $\hat{\tau}_p$  (the solid dots) and the corresponding 95 % posterior interval for  $\tau_p$  (gray vertical bars) are plotted in Figure 5(c), showing that even though  $\hat{\tau}_p$ 's in the low ability group fluctuate up and down, the fluctuations are relatively small compared to the variation of their posterior distribution. Thus, this gives some explanation as to why the correlation between their  $\hat{\theta}_p$ 's and  $\hat{\tau}_p$ 's is about zero.

The posterior distribution of  $\lambda$  is plotted in Figure 6(a). The posterior mean and 95 % posterior interval for  $\lambda$  are 0.67 and [0.48, 0.95], respectively. The speeded term  $\exp\{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}\}$  with respect to the value of  $b_j - \tau_p$ , based on the posterior mean, 2.5 % and 97.5 % quantiles for  $\lambda$ , is demonstrated in Figure 6(b). All  $P \cdot J$  values of  $\hat{b}_j - \hat{\tau}_p$  are collected, and their distribution is summarized by a nonparametric curve, using the function *density* in software R, in Figure 6(c). About 47 % of these values are greater than 0, indicating the proportion of (examinee, item) pairs suffering from test speededness. The comparison of Figure 6(b) and Figure 6(c)

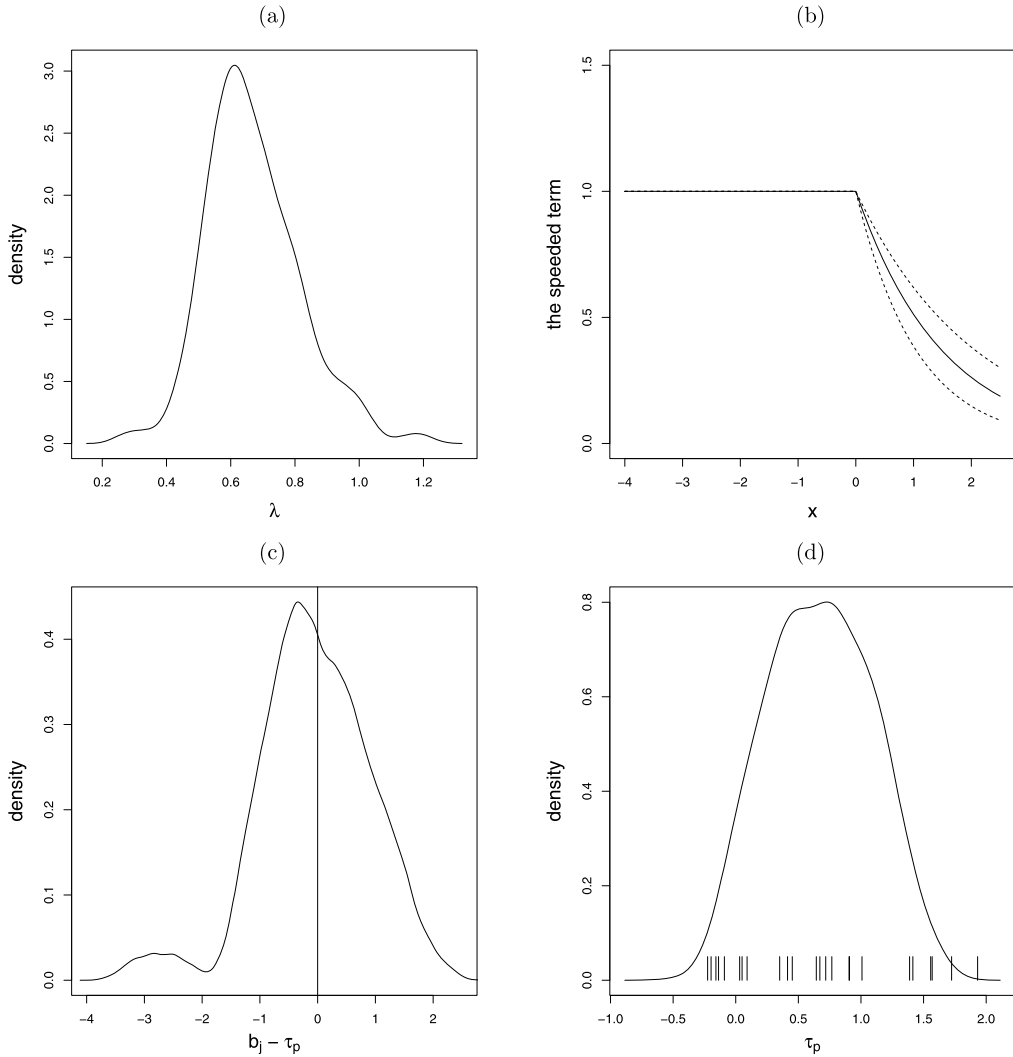


FIGURE 6.

Fitting LHL-2PL to the physics examination data: (a) posterior distribution for  $\lambda$ ; (b) the speeded term  $\exp\{-\lambda x \cdot I\{x > 0\}\}$  with posterior mean for  $\lambda$  (solid line) and 95 % posterior interval for  $\lambda$  (broken line); (c) the distribution for the collection of  $\hat{b}_j - \hat{\tau}_p$  for all  $p$  and  $j$  (the nonparametric curve), and no speededness effect for the left part of the vertical line ( $\hat{b}_j - \hat{\tau}_p = 0$ ); (d) the comparison of  $\hat{b}_j$ 's (small lines in the  $x$ -axis) with the distribution of  $\hat{\tau}_p$ 's (the nonparametric curve).

gives us some rough idea on the degree of test speededness for all the (examinee, item) pairs. We may further investigate the degree of speededness based on the plot of the spread of the threshold estimates (the nonparametric curve) and the location parameter estimates (the small line in the  $x$ -axis) in Figure 6(d). Since the thresholds for most of the examinees are relatively small compared to the largest location parameter, there is a certain degree of test speededness here. For instance, for an examinee whose threshold is 0.6, which is about the mean of the thresholds, 7 items would have a less than 0.6 probability of being in the problem solving process. Due to such a large speededness effect, it is expected that there is a great difference between the 2PL fitting and the LHL-2PL fitting. Figure 7 shows the scatter plot of the location parameter estimates under the 2PL fitting versus those under the LHL-2PL fitting. Fitting 2PL indeed yields larger

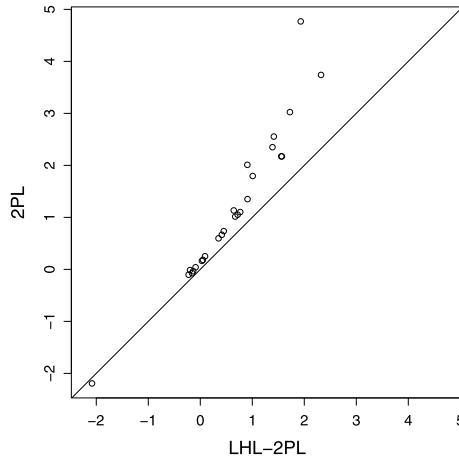


FIGURE 7.

The physics examination data: estimate of  $b_j$  from the 2PL fitting versus the LHL-2PL fitting.

TABLE 3.

The goodness of fit, model complexity and DIC of the LHL-2PL, LHL-Rasch, 2PL, and Rasch models, fitted to the physics examination data.

Model	$D(\hat{\xi}_3)$	$p_D$	DIC
LHL-2PL	22,670.51	1,056.36	24,783.22
LHL-Rasch	23,658.33	845.03	25,348.40
2PL	23,137.45	882.69	24,902.83
Rasch	23,831.76	844.49	25,520.74

location parameter estimates than fitting LHL-2PL. In addition, the larger the location parameter estimates is, the larger the difference will be. This coincides with the results of our simulations and intuition.

The 2010 physics examination is considered as more difficult than past physics examinations and requiring too much computation. Based on our analysis, the LHL-IRT models provide better fit compared to their IRT counterparts via the criterion DIC. This indicates that there is indeed a certain degree of “speededness effect” in this examination. However, our scoring scheme may bring some additional variability other than speededness effect into the data. More specifically, the “speededness effect” in our model might not only capture the effect of not having enough time to solve the items, but also account for the possible effect of being unwilling to guess when the examinee is uncertain about the answer under formula scoring. In this particular case, the “speededness effect” would be interpreted as the test performance affected by both the time limit and the unwillingness to guess under formula scoring. The above observation suggests a broader interpretation of the speededness effect rather than pure test speededness under certain scenarios. In conclusion, LHL-IRT is valuable in understanding test behavior and worth considering to avoid biased estimates due to possible speededness effects.

## 6. Discussion

In this study, a speeded IRT model and its estimation procedure are proposed. The underlying mechanism that in a high-stakes test with a time limit, an examinee may answer those easier

items first and leave the harder ones till a later test period is modeled by incorporating a speeded-effect term into the traditional IRT model. Our simulation results showed that parameters were recovered well through the proposed Bayesian estimation procedure. Moreover, fitting the current model reduced the RMSE and bias of  $b_j$ 's as compared to simply fitting a traditional IRT model when the underlying mechanism is LHL in a speeded test.

Generally speaking, the proposed parametric model is not simply for the LHL scenario. Actually, it accommodates mechanisms for which the probability of attempting one item for a particular examinee  $p$  varies according to the magnitude of the location parameter, once the location parameter is larger than an examinee-specific threshold. Furthermore, the larger the magnitude is, the smaller the probability will be. As an example, in a low-stakes test, unmotivated examinees may only answer easier items and randomly select some of harder ones to answer, and consequently the fitted "speededness effect" could be explained as the degree of lacking motivation. As another example, similar to the case in Section 5, if an examinee tries a difficult item and is unwilling to guess under uncertainty due to formula scoring, the LHL-IRT model may also be suitable for such a case in which the fitted "speededness" effect will be explained as willingness to guess or answer under uncertainty. In conclusion, the LHL-IRT model can be more widely used for situations other than simply the LHL scenario. One has to carefully interpret the fitted speededness effect depending upon the situation.

Under the framework of using parametric models to illustrate certain mechanisms during the process of a speeded test, there are mainly two classes of models. One assumes that items are answered by item ordering, and hence only the later items will be affected by the time limit, taking IRT-GPC, HYBRID, IRT-CG, and the MRM with ordinal constraints approach, for example. In the other class of models, it is hypothesized that examinees cannot fully reflect their ability on items with larger location parameters due to the time limit. LHL-IRT and IRT-DG belong to this class. However, in an achievement test with paper-pencil format, items are usually listed all at once and are not necessarily ordered by their location parameters. Therefore, fitting the two classes of model might reach different results. For a test without relatively difficult items appearing before some easier items, the models with item-ordering assumption are suitable. In contrast, for a test with several parts of questions which are ordered by the location parameters within each part, it is quite possible that examinees try their best on all the easier items first and then attempt those harder ones, in order to obtain higher score. Under this circumstance, it is advised to fit the LHL-IRT or IRT-DG to accommodate the mechanism better.

There are several strong assumptions in LHL-IRT, such as the same speededness rate  $\lambda$  for all the examinees and not allowing for guessing. In other words, the current study merely serves as a first attempt to realize the LHL mechanism by formulating the parametric model in the simplest way. Some extensions to relax these assumptions may be considered in future research. In addition, we do not distinguish between incorrect responses and nonresponse (Holman & Glas, 2005; Lord, 1983; O'Muircheartaigh & Moustaki, 1999) in LHL-IRT. However, in a test on which examinees are less likely to guess, test speededness may result in higher nonresponse rates for some items, and hence it brings more information to take nonresponse into account. Glas and Pimentel (2008) apply a nonresponse model to the missing responses on the last few items for test speededness modeling. It remains an issue to incorporate the information of nonresponse into other mechanisms of test speededness.

### Acknowledgements

We thank the associate editor and reviewers for valuable comments and suggestions which make this work more complete and the College Entrance Examination Center (CEEC) for providing the data.



Nan-Jung Hsu and Yu-Wei Chang's research has been supported by the National Science Council under Grant Nos. 101-2118-M-007-003, 100-2922-I-007-181, and 101-2922-I-007-202.

### Appendix: Full Conditionals and Sampling Scheme in Metropolis–Hastings Algorithm

Let  $\boldsymbol{\eta} = (\boldsymbol{\xi}_1, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}_2)$  be a vector consisting of all parameters in the model, all the random effects in (5), and the parameters in the first-stage prior. For any component  $x$  in  $\boldsymbol{\eta}$ , denote  $\boldsymbol{\eta}_{-x}$  as all components of  $\boldsymbol{\eta}$ , except  $x$ . At the  $(t + 1)$ -th iteration in the MCMC constructed by the Metropolis–Hastings algorithm, considering the conditional distribution of  $x$ , superscript  $(t)$  for all parameters in  $\boldsymbol{\eta}_{-x}$  stands for their current values. Current values indicate values at  $(t + 1)$ -th step for parameters whose  $(t + 1)$ -th iteration have been sampled; values at  $t$ th step for parameters whose  $(t + 1)$ -th iteration have not yet been sampled.

Let  $c_\theta$ ,  $c_\tau$ ,  $c_b$ ,  $c_a$ ,  $c_\lambda$ ,  $c_{\sigma_\tau^2}$ , and  $c_\rho$  be scales (variances) for the jumping distributions of the Metropolis–Hastings algorithm. These scales should be tuned in order to get suitable acceptance rates.

At  $(t + 1)$ -th iteration of the MCMC procedures, the sampling scheme is as follows:

- For  $\theta_p$ ,  $\tau_p$ , and  $b_j$ : a random walk jumping proposal is adopted for the Metropolis–Hastings algorithm. Their full conditionals are

$$\begin{aligned} f(\theta_p | \mathbf{y}, \boldsymbol{\eta}_{-\theta_p}) &\propto e^{\frac{-\theta_p^2}{2(1-\rho^2)}} \cdot e^{\frac{\rho\theta_p(\tau_p - \mu_\tau)}{(1-\rho^2)\sigma_\tau}} \cdot \prod_{j=1}^J \frac{[1 + e^{-a_j(\theta_p - b_j)} - e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}]^{1-y_{pj}}}{1 + e^{-a_j(\theta_p - b_j)}}, \\ f(\tau_p | \mathbf{y}, \boldsymbol{\eta}_{-\tau_p}) &\propto \prod_{j=1}^J [e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}]^{y_{pj}} \\ &\quad \cdot \prod_{j=1}^J [1 + e^{-a_j(\theta_p - b_j)} - e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}]^{1-y_{pj}} \\ &\quad \cdot e^{\frac{-1}{2(1-\rho^2)} \left[ \frac{(\tau_p - \mu_\tau)^2}{\sigma_\tau^2} - \frac{2\rho\theta_p(\tau_p - \mu_\tau)}{\sigma_\tau} \right]}, \\ f(b_j | \mathbf{y}, \boldsymbol{\eta}_{-b_j}) &\propto \prod_{p=1}^P \left[ \frac{1}{1 + e^{-a_j(\theta_p - b_j)}} \right]^{y_{pj}} \cdot \prod_{p=1}^P e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\} \cdot y_{pj}} \\ &\quad \cdot \prod_{p=1}^P \left[ \frac{1 + e^{-a_j(\theta_p - b_j)} - e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}}{1 + e^{-a_j(\theta_p - b_j)}} \right]^{1-y_{pj}} \cdot e^{\frac{-(b_j - \mu_b)^2}{2\sigma_b^2}}. \end{aligned}$$

Take  $\theta_p$  for example. Sample the candidate  $\theta_p^*$  from  $N(\theta_p^{(t)}, c_\theta)$ , and the chain of  $\theta_p^{(\cdot)}$  ‘jumps’ to  $\theta_p^*$  with probability

$$\min \left\{ \frac{f(\theta_p^* | \mathbf{y}, \boldsymbol{\eta}_{-\theta_p^*})}{f(\theta_p^{(t)} | \mathbf{y}, \boldsymbol{\eta}_{-\theta_p^{(t)}})}, 1 \right\},$$

otherwise,  $\theta_p^{(t+1)} = \theta_p^{(t)}$ . Sampling schemes for  $\tau_p$  and  $b_j$  are similar to that of  $\theta_p$  and are omitted here.

- For positive quantities  $a_j$ ,  $\lambda$  and  $\sigma_\tau^2$ : a lognormal jumping proposal is adopted for the Metropolis–Hastings algorithm. Their full conditionals are

$$\begin{aligned}
 f(a_j | \mathbf{y}, \boldsymbol{\eta}_{-a_j}) &\propto \prod_{p=1}^P \frac{[1 + e^{-a_j(\theta_p - b_j)} - e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}]^{y_{pj}}}{1 + e^{-a_j(\theta_p - b_j)}} \cdot e^{\frac{-a_j}{\beta_a}}, \\
 f(\lambda | \mathbf{y}, \boldsymbol{\eta}_{-\lambda}) &\propto e^{-\lambda \sum_{p=1}^P \sum_{j=1}^J (b_j - \tau_p) \cdot I\{b_j > \tau_p\} \cdot y_{pj}} \\
 &\quad \cdot \prod_{p=1}^P \prod_{j=1}^J [1 + e^{-a_j(\theta_p - b_j)} - e^{-\lambda(b_j - \tau_p) \cdot I\{b_j > \tau_p\}}]^{1 - y_{pj}} \cdot e^{\frac{-\lambda}{\beta_\lambda}}, \\
 f(\sigma_\tau^2 | \mathbf{y}, \boldsymbol{\eta}_{-\sigma_\tau^2}) &\propto \frac{1}{\sigma_\tau^{2(\alpha_0 + \frac{P}{2} + 1)}} \cdot e^{\frac{-1}{\sigma_\tau^2} [\beta_0 + \frac{\sum_{p=1}^P (\tau_p - \mu_\tau)^2}{2(1 - \rho^2)}] + \frac{1}{\sigma_\tau} [\frac{\sum_{p=1}^P \rho \theta_p (\tau_p - \mu_\tau)}{1 - \rho^2}]}
 \end{aligned}$$

Take  $\lambda$  for example. Sample the candidate  $\lambda^*$  from Lognormal( $\log \lambda^{(t)}, c_\lambda$ ), and the chain of  $\lambda^{(\cdot)}$  ‘jumps’ to  $\lambda^*$  with probability

$$\min \left\{ \frac{f(\lambda^* | \mathbf{y}, \boldsymbol{\eta}_{-\lambda^*}) \cdot \lambda^*}{f(\lambda^{(t)} | \mathbf{y}, \boldsymbol{\eta}_{-\lambda^{(t)}}) \cdot \lambda^{(t)}}, 1 \right\},$$

otherwise,  $\lambda^{(t+1)} = \lambda^{(t)}$ . Sampling schemes for  $a_j$  and  $\sigma_\tau^2$  are similar to that of  $\lambda$  and are omitted here.

- For  $\rho$ : we first transform  $\rho$  into the real number scale,

$$\zeta = \log \left( \frac{1 + \rho}{1 - \rho} \right),$$

and the Metropolis–Hastings algorithm is then implemented on the new scale. The full conditional for  $\rho$  is

$$\begin{aligned}
 f(\rho | \mathbf{y}, \boldsymbol{\eta}_{-\rho}) &\propto (1 - \rho^2)^{\frac{-P^2}{2}} \cdot e^{\frac{-1}{2(1 - \rho^2)} \cdot \sum_{p=1}^P (\theta_p^2 + \frac{(\tau_p - \mu_\tau)^2}{\sigma_\tau^2})} \\
 &\quad \cdot e^{\frac{-1}{2(1 - \rho^2)} \cdot (\frac{-2\rho}{\sigma_\tau}) \cdot \sum_{p=1}^P \theta_p (\tau_p - \mu_\tau)} \cdot I\{-1 \leq \rho \leq 1\}.
 \end{aligned}$$

As to the sampling scheme, first, consider a random walk jumping proposal  $J_t(\cdot | \cdot)$  on the scale of  $\zeta$ :

$$J_t(\zeta^* | \zeta^{(t)}) \sim N(\zeta^{(t)}, c_\rho). \quad (7)$$

By change of variable, the jumping proposal  $J_t(\rho^* | \rho^{(t)})$  can be obtained from (7). Sample the candidate  $\zeta^*$  from  $N(\log \frac{1 + \rho^{(t)}}{1 - \rho^{(t)}}, c_\rho)$ , and the chain of  $\rho^{(\cdot)}$  “jumps” to  $\rho^* = \frac{2 \exp(\zeta^*)}{1 + \exp(\zeta^*)}$  with probability

$$\min \left\{ \frac{f(\rho^* | \mathbf{y}, \boldsymbol{\eta}_{-\rho^*}) \cdot (1 + \rho^*)(1 - \rho^*)}{f(\rho^{(t)} | \mathbf{y}, \boldsymbol{\eta}_{-\rho^{(t)}}) \cdot (1 + \rho^{(t)})(1 - \rho^{(t)})}, 1 \right\},$$

otherwise,  $\rho^{(t+1)} = \rho^{(t)}$ .

The full conditionals for other components in  $\boldsymbol{\eta}$  are standard as follows:

- For  $\mu_\tau$ , sample  $\mu_\tau^{(t+1)}$  from

$$\begin{aligned}
 &N \left( \frac{\sigma_0^{2(t)} (\sum_{p=1}^P \tau_p^{(t)}) + \mu_0^{(t)} (1 - (\rho^{(t)})^2) \sigma_\tau^{2(t)} - \sigma_\tau^{(t)} \sigma_0^{2(t)} \rho^{(t)} (\sum_{p=1}^P \theta_p^{(t)})}{P \cdot \sigma_0^{2(t)} + (1 - (\rho^{(t)})^2) \sigma_\tau^{2(t)}}, \right. \\
 &\quad \left. \frac{(1 - (\rho^{(t)})^2) \sigma_\tau^{2(t)} \sigma_0^{2(t)}}{P \cdot \sigma_0^{2(t)} + (1 - (\rho^{(t)})^2) \sigma_\tau^{2(t)}} \right).
 \end{aligned}$$

- For  $\mu_b$ , sample  $\mu_b^{(t+1)}$  from

$$N\left(\frac{\sigma_2^2(\sum_{j=1}^J b_j^{(t)}) + \sigma_b^{2(t)} \mu_2}{\sigma_b^{2(t)} + J \cdot \sigma_2^2}, \frac{\sigma_b^{2(t)} \sigma_2^2}{\sigma_b^{2(t)} + J \cdot \sigma_2^2}\right).$$

- For  $\sigma_b^2$ , sample  $\sigma_b^{2(t+1)}$  from

$$\text{Inv-Gamma}\left(\alpha_1 + \frac{J}{2}, \beta_1 + \frac{1}{2} \sum_{j=1}^J (b_j^{(t)} - \mu_b^{(t)})^2\right).$$

- For  $\beta_a$ , sample  $\beta_a^{(t+1)}$  from

$$\text{Inv-Gamma}\left(\alpha_2 + J, \beta_2 + \sum_{j=1}^J a_j^{(t)}\right).$$

- For  $\beta_\lambda$ , sample  $\beta_\lambda^{(t+1)}$  from

$$\text{Inv-Gamma}(\alpha_3 + 1, \beta_3 + \lambda^{(t)}).$$

- For  $\mu_\zeta$ , sample  $\mu_\zeta^{(t+1)}$  from

$$N\left(\frac{\log(\frac{1+\rho^{(t)}}{1-\rho^{(t)}}) \sigma_3^2 + \sigma_\zeta^{2(t)} \mu_3}{\sigma_\zeta^{2(t)} + \sigma_3^2}, \frac{\sigma_\zeta^{2(t)} \sigma_3^2}{\sigma_\zeta^{2(t)} + \sigma_3^2}\right).$$

- For  $\sigma_\zeta^2$ , sample  $\sigma_\zeta^{2(t+1)}$  from

$$\text{Inv-Gamma}\left(\alpha_5 + \frac{1}{2}, \beta_5 + \frac{1}{2} \left(\log \frac{1+\rho^{(t)}}{1-\rho^{(t)}} - \mu_\zeta^{(t)}\right)^2\right).$$

- For  $\mu_0$ , sample  $\mu_0^{(t+1)}$  from

$$N\left(\frac{\sigma_0^{2(t)} \mu_1 + \sigma_1^2 \mu_\tau^{(t)}}{\sigma_0^{2(t)} + \sigma_1^2}, \frac{\sigma_0^{2(t)} \sigma_1^2}{\sigma_0^{2(t)} + \sigma_1^2}\right).$$

- For  $\sigma_0^2$ , sample  $\sigma_0^{2(t+1)}$  from

$$\text{Inv-Gamma}\left(\alpha_6 + \frac{1}{2}, \beta_6 + \frac{1}{2} (\mu_\tau^{(t)} - \mu_0^{(t)})^2\right).$$

- For  $\beta_0$ , sample  $\beta_0^{(t+1)}$  from

$$\text{Gamma}\left(\alpha_0 + \alpha_4, \frac{\sigma_\tau^{2(t)} \beta_4}{\sigma_\tau^{2(t)} + \beta_4}\right).$$

Notice that the full conditional  $f(x|\boldsymbol{\eta}_{-x})$  is usually a product of many terms. A good tip to avoid numerical problems during implementation is to work on the log scale in both the numerator and the denominator first for computing the acceptance ratio in the Metropolis–Hastings procedure.

- Bejar, I.I. (1985). *Test speededness under number-right scoring: an analysis of the test of English as a foreign language* (Research Report RR-85-11). Princeton: Educational Testing Service.
- Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2002). Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39, 331–348.
- Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computerized-adaptive test scores. *Journal of Educational Measurement*, 41, 137–148.
- Cao, J., & Stokes, S.L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, 73, 209–230.
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models*. New York: Springer.
- Evans, F.R., & Reilly, R.R. (1972). A study of test speededness as a source of bias. *Journal of Educational Measurement*, 9, 123–131.
- Fox, J.-P. (2010). *Bayesian item response modeling-theory and applications*. New York: Springer.
- Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457–511.
- Glas, C.A.W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, 68, 907–922.
- Goegebeur, Y., De Boeck, P., Wollack, J.A., & Cohen, A.S. (2008). A speeded item response model with gradual process change. *Psychometrika*, 73, 65–87.
- Holman, R., & Glas, C.A.W. (2005). Modeling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical & Statistical Psychology*, 58, 1–18.
- Kingston, N.M., & Dorans, N.J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8, 147–154.
- Lord, F.M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, 12, 7–11.
- Lord, F.M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, 48, 477–482.
- O’Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society. Series A*, 162, 177–194.
- Oshima, T.C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31, 200–219.
- Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271–282.
- 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*, 64, 583–616.
- Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51, 589–601.
- van der Linden, W.J. (2011). Setting time limits on tests. *Applied Psychological Measurement*, 35, 183–199.
- van der Linden, W.J., Breithaupt, K., Chuah, S.C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, 44, 117–130.
- Wollack, J.A., Cohen, A.S., & Wells, C.S. (2003). A method of maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, 40, 307–330.
- 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).
- Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). New York: Waxmann.
- Yen, W.M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.

Manuscript Received: 5 JUN 2012

Final Version Received: 3 DEC 2012