A Three-Parameter Speeded Item Response Model: Estimation and Application

Joyce Chang, Henghsiu Tsai, Ya-Hui Su, and Edward M. H. Lin

Abstract When given time constraints, it is possible that examinees leave the harder items till later and are not able to finish answering every item in time. In this paper, this situation is modeled by incorporating a speeded-effect term into a three-parameter logistic item response model. Due to the complexity of the likelihood structure, a Bayesian estimation procedure with Markov chain Monte Carlo method is presented. The methodology is applied to physics examination data of the Department Required Test for college entrance in Taiwan for illustration.

Keywords Item response model • Markov chain Monte Carlo • Test speededness

1 Introduction

Over the past few decades, there has been increasing interest in modeling response data generated from tests that are administered within an allocated time, which may be insufficient for some examinees. A test is said to be speeded if the time limit affects examinees' test performance (see, for example, Lee & Ying 2015). In order to reduce the contamination of the test speededness in modeling response

J. Chang
Department of Economics, The University of Texas at Austin, 2225 Speedway, BRB 1.116, C3100, Austin, Texas 78712, USA
e-mail: joyce.chang@utexas.edu

H. Tsai
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nangang District, Taipei 11529, Taiwan
e-mail: htsai@stat.sinica.edu.tw

Y.-H. Su (✉)
Department of Psychology, National Chung Cheng University, 168 University Road, Section 1, Min-Hsiung, Chai-Yi 62102, Taiwan
e-mail: psyyhs@ccu.edu.tw

E.M.H. Lin
Institute of Finance, National Chiao Tung University, 1001 University Road, Hsinchu 300, Taiwan
e-mail: m9281067@gmail.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_3

data, several models have been proposed in the literature. Yamamoto (1995) uses the HYBRID model to describe the behavior that an examinee may switch to a guessing strategy midway through a test due to the time constraint. Unlike the unspeeded items, which are characterized by a two-parameter logistic (2PL) model, the speeded ones are, on the other hand, characterized by a latent class based item response model. Bolt, Cohen, and Wollack (2002) use the mixture Rasch model of Rost (1990) to deal with situations where no penalty is imposed for guessing; consequently, speededness effects tend to emerge in the form of incorrect as opposed to omitted responses. Goegebeur, De Boeck, Wollack, and Cohen (2008) propose a speeded item response theory (IRT) model with gradual process change. Under this model, responses to items early in the test are governed by a 3PL model, and beyond some point the success probability gradually decreases and eventually reduces to the success probability under random guessing. Chang, Tsai, and Hsu (2014) propose the leave-the-harder-till-later speeded two-parameter logistic (LHL-2PL) model to accommodate the speeded effect. Additional literature on test speededness includes Bejar (1985), Yamamoto (1989), Yamamoto and Everson (1997), Boughton and Yamamoto (2007), Cao and Stokes (2008), and Wang and Xu (2015), among others.

In this paper, we are interested in extending the LHL-2PL model by adding a pseudo-guessing parameter. Chang, Tsai, and Hsu (2014) apply the LHL-2PL model to the physics examination data of Department Required Test (DRT) for college entrance in Taiwan, and find some evidence for the LHL mechanism in analyzing the data. Examinees have to answer 26 questions in 80 min, where the first 20 questions are multiple-choice questions that examinees should choose one correct answer out of 5 possible choices. It is then followed by 4 multiple-response questions, where out of the 5 possible, examinees need to select all the answer choices that apply, and finally 2 calculation problems. The test is administered under formula-scoring directions, where 3/4 and 1 point are deducted from the raw score for each incorrect answer made in the multiple-choice and multiple-response questions respectively. If an item is left blank, the examinee would get 0 point. Furthermore, the adjusted score would only be 0 or above for these two types of questions.

Based on the discussions of Lord (1975) on formula scoring, Chang, Tsai, and Hsu (2014) argue that examinees are less likely to guess whenever they do not know the answer, and therefore, it provides some rationale for considering a speeded model in which random guessing is not allowed. However, it is also argued that examinees often know enough about the subject to eliminate some of the incorrect choices. That being the case, guessing from among the remaining options is likely to help them overcome the penalty of $1/(k-1)$, where k is the number of options, and is 5 for the first 20 multiple-choice questions (e.g., Angoff 1989). For each of the 4 multiple-response questions, there are 5 choices, and each one is graded independently, so $\mathrm{k} = 2$. That is, each choice in the multiple-response question is either true or false. In the literature, many papers also allow random guessing (or pseudo-guessing) parameters in their models, see, for example, Cao and Stokes (2008), Goegebeur, De Boeck, Wollack, and Cohen (2008), and Wang and Xu (2015). This motivates us to consider in this paper the leave-the-

harder-till-later speeded three-parameter logistic IRT (LHL-3PL) model by adding a pseudo-guessing parameter to the LHL-2PL model of Chang, Tsai, and Hsu (2014).

The rest of the paper is organized as follows. In Sect. 2, we describe the LHL-3PL model in more details. Since our model is a direct extension of Chang, Tsai, and Hsu (2014), our prior settings are the same as theirs except for the extra pseudo-guessing parameters. The prior settings for the pseudo-guessing parameters will also be mentioned in Sect. 2. A simulation study is conducted in Sect. 3 to demonstrate the validation of the Bayesian estimation procedure. Application of the LHL-3PL model to the data of Department Required Test for college entrance in Taiwan is illustrated in Sect. 4. Section 5 concludes.

2 Leave-the-Harder-till-Later Speeded Three-Parameter Logistic Item Response Model

Let Y_{pj} be the dichotomous response of examinee p on item j, where $p = 1, 2, \ldots, P$, and $J = 1, 2, \ldots, J$. Denote b_j and a_j as the location and scale parameters respectively, for item j, and θ_p as the ability parameter for examinee p. In the 2PL model (Birnbaum 1968), the probability that examinee p gets a correct response on item j is given by

$$\Pr(Y_{pj} = 1|a_j, b_j, \theta_p) = \frac{1}{1 + e^{-a_j(\theta_p - b_j)}}.$$

The parameter a_j is also known as the discrimination parameter (de Ayala 2009), or the slope parameter (Wang 2004), and the parameter b_j is called the difficulty parameter in Embretson and Reise (2000) and Wang and Xu (2015). For more descriptions and discussions of the 2PL model, see Embretson and Reise (2000), Wang (2004), and de Ayala (2009).

The three-parameter logistic (3PL) model is obtained by adding an extra parameter to the 2PL model. Under the 3PL model,

$$\Pr(Y_{pj} = 1|a_j, b_j, c_j, \theta_p) = c_j + \left(1 - c_j\right) \cdot \frac{1}{1 + e^{-a_j(\theta_p - b_j)}}.$$

The parameter c_j is referred to as the item's pseudo-guessing or pseudo-chance parameter and equals the probability of a correct response when θ approaches $-\infty$ (de Ayala 2009). It is also named the asymptotic parameter (Wang 2004) or the lower-asymptotic parameter (Embretson & Reise 2000). The 3PL model is suitable for multiple-choice cognitive items (Embretson & Reise 2000; Wang 2004).

Unlike the traditional IRT models described above, where unspeededness is implicitly assumed, Chang, Tsai, and Hsu (2014) introduce two additional parameters to the 2PL model in an attempt to capture the effect of speededness. It is assumed that the probability of a correct response in given by

$$\Pr\left(Y_{pj}=1\left|a_j,b_j,\theta_p,\tau_p,\lambda\right.\right)=\frac{e^{-\lambda\left(b_j-\tau_p\right)\cdot\mathrm{I}\left\{b_j>\tau_p\right\}}}{1+e^{-a_j\left(\theta_p-b_j\right)}},\tag{1}$$

where τ_p is the p-th examinee's threshold parameter for speededness and λ, which is always larger than zero, is the speededness rate. Indicator function $I\{\cdot\}$ is defined as

$$I\{b_j>\tau_p\}=\begin{cases}1, & b_j>\tau_p,\\ 0, & b_j\le\tau_p.\end{cases}$$

The rationality behind the model is as follows. When encountering an item, the examinee would decide if he would get into solving process right away by the level of difficulty of the item. If its difficulty exceeds one's threshold, τ_p, i.e., $b_j>\tau_p$, the item is considered time-consuming and would be retained till a later test period. It is further assumed that the first-skipped item would be answered with the probability of $e^{-\lambda(b_j-\tau_p)}$. In other words, the model can be partitioned into two parts: (1) whether to solve or not, and (2) whether the answer is correct. The two stages are given by

$$Z_{pj}|(b_j,\tau_p,\lambda)\sim\text{Bernoulli}\left(e^{-\lambda(b_j-\tau_p)\cdot I\{b_j>\tau_p\}}\right),$$

$$Y_{pj}|(a_j,b_j,\theta_p,Z_{pj})\sim\text{Bernoulli}\left(\frac{1}{1+e^{-a_j\left(\theta_p-b_j\right)}}\cdot Z_{pj}\right),$$

where Z_{pj} denotes whether the item is being answered or not.

As discussed in Sect. 1, for the DRT data, the first 20 questions and the 21st to the 24th questions are multiple-choice questions and multiple-response questions respectively, and are therefore, naturally suitable for a 3PL model, where a pseudo-guessing parameter is included. Specifically, we consider the LHL-3PL model (to be defined below). For the last 2 calculation problems, we simply set the corresponding pseudo-guessing parameters to be zero. Under the LHL-3PL model,

$$\Pr\left(Y_{pj}=1\left|a_j,b_j,c_j,\theta_p,\tau_p,\lambda\right.\right)=c_j+\left(1-c_j\right)\cdot\frac{e^{-\lambda\left(b_j-\tau_p\right)\cdot\mathrm{I}\left\{b_j>\tau_p\right\}}}{1+e^{-a_j\left(\theta_p-b_j\right)}},\tag{2}$$

where $0<c_j<1$. We want to compare our proposed LHL-3PL model with the LHL-2PL of Chang, Tsai, and Hsu (2014) to explore the role of random guessing in the DRT data, so we adopt the assumptions, including the normality of the joint distribution of θ_p and τ_p, prior settings and the MCMC-based estimation procedure of Chang, Tsai, and Hsu (2014). For the pseudo-guessing parameter c_j, we transform it into the real number scale γ_j, and assume

$$\gamma_j=\log\left(\frac{c_j}{1-c_j}\right)\sim\mathrm{N}\left(\mu_\gamma,\sigma_\gamma^2\right),\tag{3}$$

Table 1 RMSE of estimates from LHL-3PL fitting under data generated from the LHL-3PL model (10 replicates)

Parameter\ P	250	500	1,000
$\boldsymbol{b}$	0.9521	0.9392	0.7881
$\boldsymbol{a}$	1.4735	0.8152	0.7369
$\boldsymbol{c}$	0.0897	0.0978	0.0978
$\boldsymbol{\theta}$	0.5645	0.5387	0.5306
$\boldsymbol{\tau}$	2.8719	2.8198	2.7675

and

$$\mu_\gamma \sim \mathrm{N}\left(\mu, \sigma^2\right), \qquad \sigma_\gamma^2 \sim \mathrm{Inv-Gamma}\left(\alpha, \beta\right), \tag{4}$$

where $\mu = 0, \sigma^2 = 1, \alpha = \beta = 3$.

Bayesian estimation method has been widely used in IRT modeling, see, for example, Swaminathan and Gifford (1982, 1985, 1986), Mislevy (1986), Bolt, Cohen, and Wollack (2002), van der Linden (2007), Cao and Stokes (2008), Fox (2010), Meyer (2010), and Chang, Tsai, and Hsu (2014).

3 Simulation Study

In this section, we conduct a simulation study to evaluate the performance of the MCMC method in estimating the parameters. All computations were performed using some Fortran code with IMSL subroutines.

We first describe the true data generating process. We consider $J = 40$, $P = 250$, 500, and 1,000. Let $\boldsymbol{a} = (a_1, \cdots, a_J)$, $\boldsymbol{b} = (b_1, \cdots, b_J)$, $\boldsymbol{c} = (c_1, \cdots, c_J)$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_P)$, and $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_P)$. The true values of $\boldsymbol{a}$ and $\boldsymbol{b}$ are the same as those considered in Sect. 4 of Chang, Tsai, and Hsu (2014). For the true values of $\boldsymbol{c}$, we set $c_j = (40.5 - j)/40$, for $j = 1, \ldots, 40$. The true value of λ equals 1. For $p = 1, \ldots, P$, (θ_p, τ_p) are independently and identically sampled from a bivariate normal distribution with the marginal distribution of θ_p and τ_p being $N(0, 1)$ and $N(0.2, 0.5)$, respectively, and the correlation being 0.8.

We produce 40,000 MCMC draws with the first 10,000 draws as burn-in. For each parameter, the posterior mean was calculated as our Bayes estimates, based on 30,000 MCMC draws after burn-in. We repeat the exercise 10 times, and the root mean squared error (RMSE) of the posterior means are summarized in Table 1. From Table 1, it is clear that, in general, the RMSE decreases with the value P, except for the parameter $\boldsymbol{c}$. However, the RMSE's of the parameter $\boldsymbol{c}$ are the smallest, and those of the parameter $\boldsymbol{\tau}$ are the largest. From $P = 250$ to $P = 1{,}000$, the RMSE's of the parameter $\boldsymbol{a}$ become half.

4 Application

In this section, the proposed LHL-3PL model and the MCMC procedure described in the previous section are applied to the data of the physics examination of the 2010 Department Required Test for college entrance in Taiwan provided by College Entrance Examination Center (CEEC). The data from 1,000 randomly sampled examinees contains the original responses and nonresponses information, but we treat both nonresponses and incorrect answers the same way and code them as $Y_{pj} = 0$ as suggested by Chang, Tsai, and Hsu (2014). As for the calculation part, the response Y_{pj} is coded as 1 whenever the original score is more than 7.5 out of 10 points, and zero otherwise.

The four models, including the 2PL, LHL-2P, 3PL, and the LHL-3PL models, are fitted to the data using Bayesian analysis. For the 3PL and the LHL-3PL models, we set $c_{25} = c_{26} = 0$ because guessing is in theory not possible. Further comparison is made via Bayesian model selection criterion, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde 2002), described below.

We use the posterior means as the point estimates for parameters of interest. Let $\boldsymbol{\xi} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{\tau}, \lambda)$, and $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}, \hat{\lambda})$ be the posterior mean of $\boldsymbol{\xi}$ under the fitted LHL-3PL model given data $\mathbf{y} = (\mathbf{y_1}, \cdots, \mathbf{y_P})$, where $\mathbf{y_p} = (\mathrm{y}_{p1}, \cdots, \mathrm{y}_{pJ})$. The DIC for the fitted LHL-3PL model is defined as

$$\mathrm{DIC} = D(\hat{\boldsymbol{\xi}}) + 2p_D, \tag{5}$$

where

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

In (5), the first term $D(\hat{\boldsymbol{\xi}})$ measures the goodness-of-fit, and the second term p_D, which represents the effective number of parameters used in the model, is the difference between posterior mean deviance and deviance evaluated at the posterior means of the parameters. The DIC for the other three fitted models are defined similarly. 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 the second row of Table 2. The LHL-3PL has a smallest DIC, indicating the best fitting performance of the LHL-3PL as compared to the other models after compensating for model complexity.

Apart from DIC, the Bayesian model-data fit checking techniques, such as posterior predictive model checking (PPMC), has also been used in the literature. See, for example, Li, Bolt, and Fu (2006), Sinharay, Johnson, and Stern (2006), and Huang and Hung (2010). The procedure runs as follows:

Table 2 DIC for physics examination data of the Department Required Test for college entrance in Taiwan

Model	2PL	LHL-2PL	3PL	LHL-3PL
DIC	24,671.99	24,717.57	24,506.24	24,416.17

Step 1. Compute the realized discrepancy measure from the observed data set $\mathbf{y}$.
Step 2. Generate a draw of parameter $\boldsymbol{\xi}$ from the posterior distribution.
Step 3. Draw a data set $\tilde{\mathbf{y}}$ from the model, using the parameter $\boldsymbol{\xi}$ drawn in Step 2.
Step 4. Compute the value of the predictive discrepancy measure from the above draws of parameters and data set $\tilde{\mathbf{y}}$.
Step 5. Repeat Steps 2–4 1,000 times to compute the posterior predictive p-value (PPP-value).

The PPP-value is defined to be the percent of times that the predictive discrepancy measure is larger than its realized counterpart. An extreme PPP-value (PPP-value larger than 0.975 or smaller than 0.025) suggests that the model fits the data poor (Li, Bolt, & Fu 2006, p. 11). Following from Li, Bolt, and Fu (2006) and Sinharay, Johnson, and Stern (2006), we use the sample odds ratio (e.g. Agresti 2002p. 45) as the discrepancy measure in our study. The sample odds ratio is defined to be $OR = (n_{11}n_{00})/(n_{10}n_{01})$, where n_{jk} denotes the number of individuals scoring j on the first item and k on the second item, $j, k = 0, 1$. The sample odds ratio tests item response association between a pair of items. Here, we have $J = 26$ items, resulting in $J(J-1)/2 = 325$ pairs, and therefore, 325 PPP-values. The number of extreme PPP-values of the four fitted models are all zeros, indicating the goodness of fits of these four models.

Let $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_J)$, where, for $j = 1, \ldots, J$, $\pi_j = \sum_{p=1}^{P} y_{pj}/P$. Thus, for $j = 1, \ldots, 24$, π_j represents the percent of examinees who respond correctly to question j, and for $j = 25$ and 26, it represents the percent of examinees whose original score is more than 7.5.

Now, we compare the estimates of these four models. Since the estimates of 2PL and LHL-2PL are similar, and those of 3PL and LHL-3PL are similar, we only compare those of LHL-2PL and LHL-3PL in the following. Figure 1a shows the plots of $\hat{c}_j$ and π_j, over $j = 1, \ldots, 26$. Recall that $c_{25} = c_{26} = 0$. From Fig. 1a, we see that fewer examinees score more than 7.5 or above in the calculation problems than getting a correct answer on each of the multiple-choice questions or the multiple-response questions. Figure 1b reveals that there are some discrepancies between the estimated discrimination parameters $\hat{\boldsymbol{a}}$ under the LHL-3PL and the LHL-2PL model, whereas the estimated difficulty parameters $\hat{\boldsymbol{b}}$ are very close (Fig. 1c). The sample correlations between the estimates under the two models are 0.177 and 0.969 for $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ respectively (Table 3).

The sample correlation matrix of $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{c}}$ and $\boldsymbol{\pi}$ under LHL-2PL and LHL-3PL given in Table 4 shows that $\boldsymbol{\pi}$ is highly correlated with $\hat{\boldsymbol{b}}$, and is negatively correlated (although the correlation is moderate) with $\hat{\boldsymbol{a}}$ under LHL-3PL while almost uncorrelated under LHL-2PL. For $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$, there is a moderate correlation

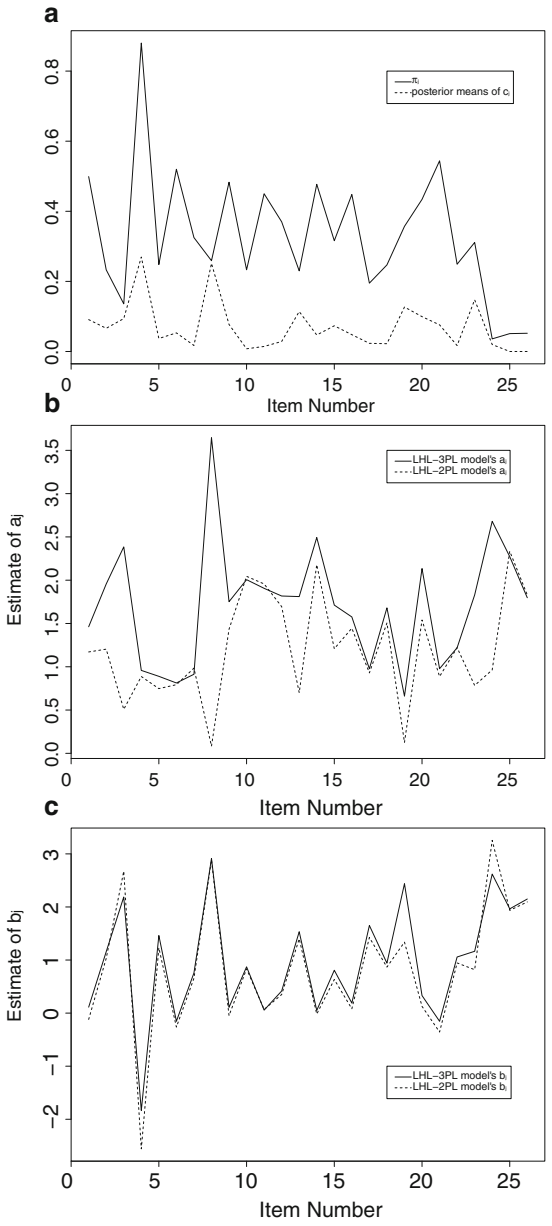

Fig. 1 (**a**) Plots of $\boldsymbol{\pi}$ and $\hat{\boldsymbol{c}}$, for $j = 1, \ldots, 26$; (**b**) plots of $\hat{\boldsymbol{a}}$ under LHL-3PL and LHL-2PL; (**c**) plots of $\hat{\boldsymbol{b}}$ under LHL-3PL and LHL-2PL

under LHL-3PL, whereas there is a low and negative correlation under LHL-2PL. For $\boldsymbol{\pi}$ and $\hat{\boldsymbol{c}}$, they are moderately correlated.

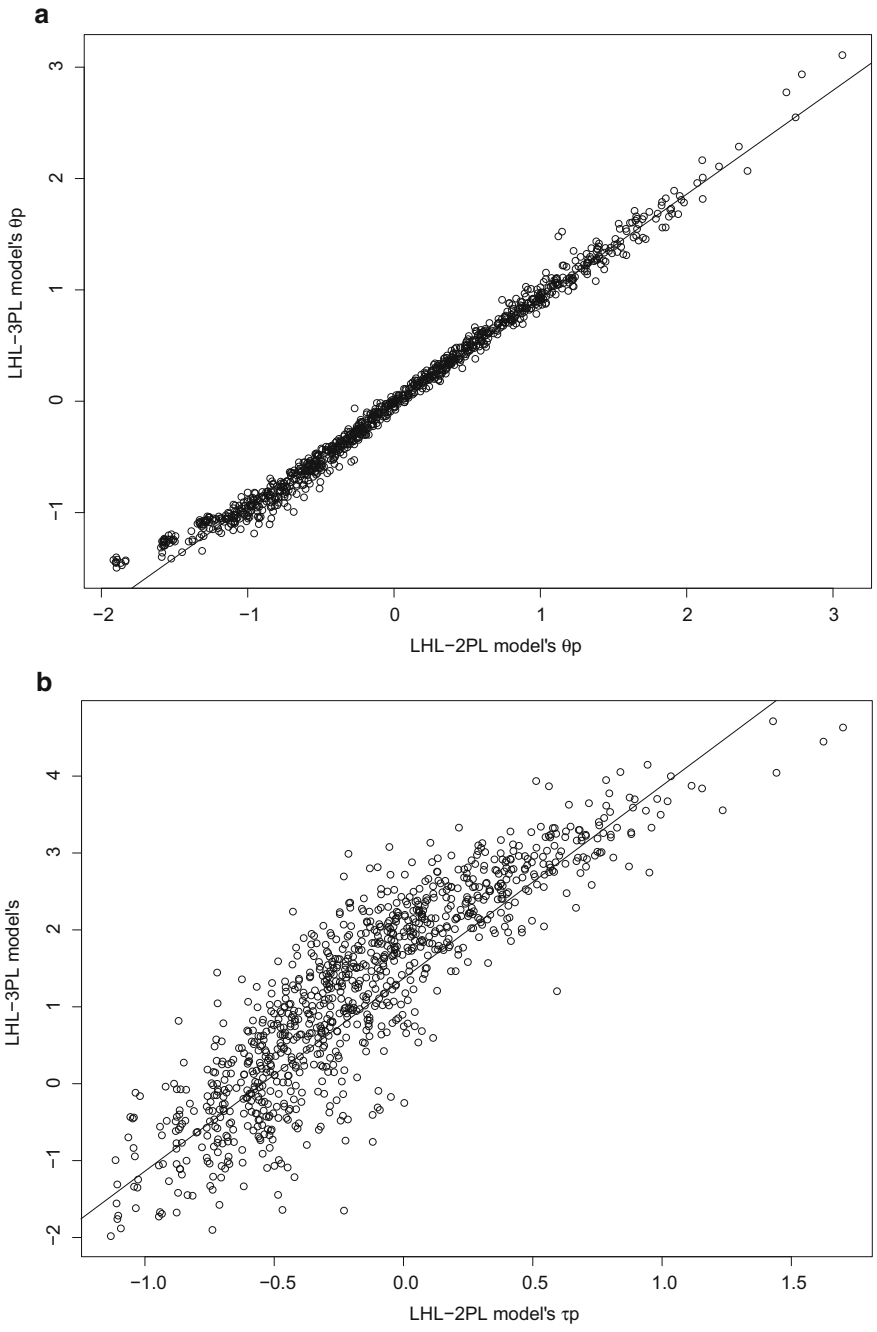

Fig. 2 (**a**) Scatter plot of $\hat{\boldsymbol{\theta}}$ under LHL-3PL against LHL-2PL; (**b**) Scatter plot of $\hat{\boldsymbol{\tau}}$ under LHL-3PL against LHL-2PL

Table 3 Sample correlations between the estimates under LHL-3PL and LHL-2PL

	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\tau}}$	$\hat{\boldsymbol{a}}$	$\hat{\boldsymbol{b}}$
Correlation	0.994	0.882	0.177	0.969

Table 4 Sample correlations of the estimates for LHL-3PL, with their counterparts for LHL-2PL enclosed by parentheses

	$\hat{\boldsymbol{a}}$	$\hat{\boldsymbol{b}}$	$\hat{\boldsymbol{c}}$
$\boldsymbol{\pi}$	−0.387(−0.067)	−0.877(−0.919)	0.492
$\hat{\boldsymbol{a}}$		0.417(−0.204)	0.150
$\hat{\boldsymbol{b}}$			−0.132

Figure 2a shows that the estimated $\hat{\boldsymbol{\theta}}$ under both models yields very similar results. Figure 2b, however, shows that there is a larger difference between the estimated examinee-specific threshold parameters. Indeed, the variations of $\hat{\boldsymbol{\tau}}$ under LHL-3PL are much larger than those of LHL-2PL. This may be due to the inclusion of the extra pseudo-guessing parameters in the LHL-3PL model. The sample correlations between the estimates under LHL-3PL and LHL-2PL are 0.994 and 0.882 for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ respectively (Table 3).

Figure 1a and b reveals that item 8 has a π that is very close to its c-parameter estimate and it has very different a-parameter estimates in the LHL-2PL and LHL-3PL. We therefore compute the estimated probability that item 8 is answered correctly in these four models. We first consider the LHL-3PL model. This is done as follows. Recall that we produce 40,000 MCMC draws with the first 10,000 draws as burn-in. For $p = 1, \ldots, P$, for each draw after burn-in, we compute the probability that $\{Y_{p8} = 1\}$ using Eq. (2), then we take the average over all the last $30,000$ draws to get an estimate of the probability that $\{Y_{p8} = 1\}$. Then, we take the average over $p = 1, \ldots, P$, to get the estimate of the probability that item 8 is answered correctly. We repeat the computation for the other 3 models. The estimated values are 0.26642, 0.25783, 0.26285, and 0.25860 in the 2PL, 3PL, LHL-2PL, and LHL-3PL models, respectively. Since $\pi_8 = 0.259$, we see that the estimate in the LHL-3PL model is closest to π_8. However, the interpretations under the LHL-2LP and LHL-3PL models are quite different. In the LHL-3PL model, item 8 has the highest b-parameter estimate, meaning that it is the most difficult one, and most examinees answer it correctly just by guessing. This may or may not be true, and deserves a further study by putting some stronger priors on the $\boldsymbol{c}$-parameter instead of using a two-layer hierarchical prior in this study to reduce the impact of the prior settings.

5 Concluding Remarks

In this study, we extend the LHL-2PL model to the LHL-3PL model by adding a pseudo-guessing parameter. Then, we apply the LHL-3PL model to the physics examination data of the Department Required Test for college entrance in Taiwan. The test consists of three types of questions, including multiple-choice, multiple-

response, and calculation problems. The percent of examinees who responded correctly are the lowest for the two calculation problems. The estimated pseudo-guessing parameters for the multiple-choice and multiple-response questions range from 0.0077 to 0.2694, indicating some evidence of random guessing. This may due to the fact that examinees often know enough about the subject to eliminate some of the incorrect choices. Therefore, guessing from among the remaining options is likely to help them beat the odds of random guessing. We found that the estimated ability parameters almost unaffected by adding a pseudo-guessing parameter to the model. The changes in the estimated difficulty parameters are also slim. Changes are mainly in some of the estimated discrimination parameters and many of the estimated examinee-specific threshold parameters for the speededness effect. In sum, we find some evidence for the LHL mechanism as well as for random guessing.

In the LHL-3PL model, we consider the case that all the examinees share the same speededness rate λ. It is interesting to relax the assumption in a further study. Another interesting future work is to put some stronger priors on the $\boldsymbol{c}$-parameter.

Acknowledgements The research was supported by Academia Sinica and the Ministry of Science and Technology of the Republic of China under grant number MOST 102-2118-M-001 -007 -MY2. The authors would like to thank the co-editor, Professor Wen-Chung Wang, and Dr. Yu-Wei Chang for their helpful comments and suggestions, and the College Entrance Examination Center (CEEC) for providing the data.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: Wiley.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement, 26*, 323–336.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the test of English as a foreign language* (Research Rep. RR-85-11). Princeton: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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.

Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). New York: Springer.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73*, 209–230.

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika, 79*, 255–274.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: L. Erlbaum Associates.

Fox, J.-P. (2010). *Bayesian item response modeling-theory and applications*. New York: Springer.

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.

Huang, H.-Y., & Hung, S.-P. (2010). Implementation and application of Bayesian three-level IRT random intercept latent regression model. *Chinese Journal of Psychology, 52*, 309–326. (in Chinese)

Lee, Y.-H., & Ying, Z. (2015). A mixture cure-rate model for responses and response times in time-limit tests. *Psychometrika, 80*, 748–775.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3–21.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*, 7–11.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*, 521–538.

Mislevy, R. L. (1986). Bayes modal estimation in item response theory. *Psychometrika, 51*, 177–195.

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298–321.

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. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175–192.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the twoparameter logistic model. *Psychometrika, 50*, 349–364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589–601.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308.

Wang, W.-C. (2004). Rasch measurement theory and application in education and psychology. *Journal of Education and Psychology, 27*, 637–694 (in Chinese).

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*, 456–477.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (ETS Research Rep. No. RR-89-41). Princeton: Educational Testing Service.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Technical Rep. No. TR-10). Princeton: Educational Testing Service.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost (Ed.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Munster: Waxmann.