frontiers in Psychology

ORIGINAL RESEARCH
published: 05 August 2020
doi: 10.3389/fpsyg.2020.01791

A General Three-Parameter Logistic Model With Time Effect

Zhaoyuan Zhang†, Jiwei Zhang†, Jian Tao and Ningzhong Shi**

Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China

Within the framework of item response theory, a new and flexible general three-parameter logistic model with response time (G3PLT) is proposed. The advantage of this model is that it can combine time effect, ability, and item difficulty to influence the correct-response probability. In contrast to the traditional response time models used in educational psychology, the new model incorporates the influence of the time effect on the correct-response probability directly, rather than linking them through a hierarchical method via latent and speed parameters as in van der Linden's model. In addition, the Metropolis–Hastings within Gibbs sampling algorithm is employed to estimate the model parameters. Based on Markov chain Monte Carlo output, two Bayesian model assessment methods are used to assess the goodness of fit between models. Finally, two simulation studies and a real data analysis are performed to further illustrate the advantages of the new model over the traditional three-parameter logistic model.

Keywords: Bayesian inference, deviance information criterion (DIC), item response theory (IRT), logarithm of the pseudomarginal likelihood (LPML), Markov chain Monte Carlo (MCMC), three-parameter logistic model

OPEN ACCESS

Edited by:
Jason C. Immekus, University of Louisville, United States

Reviewed by:
Rubén Maneiro, Pontifical University of Salamanca, Spain
Dandan Liao, American Institutes for Research, United States

***Correspondence:**
Jiwei Zhang
zhangjw713@nenu.edu.cn
Ningzhong Shi
shinz@nenu.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *11 October 2019*
Accepted: *29 June 2020*
Published: *05 August 2020*

Citation:
Zhang Z, Zhang J, Tao J and Shi N (2020) A General Three-Parameter Logistic Model With Time Effect. Front. Psychol. 11:1791. doi: 10.3389/fpsyg.2020.01791

1. INTRODUCTION AND MOTIVATION

Computerized assessment has become a widely accepted method of testing owing to the fact that the results produced by examinees can be quickly and accurately evaluated by virtue of the computational power that is now available. In addition, with the help of computer technology, the response times of examinees are easier to collect than in the case of traditional paper-and-pencil tests. The collected response times provide a valuable source of information on examinees and test items. For example, response times can be used to improve the accuracy of ability estimates (van der Linden, 2007; Klein Entink et al., 2009a; van der Linden and Glas, 2010; Wang et al., 2013, 2018a; Wang and Xu, 2015; Fox and Marianti, 2016; Bolsinova and Tijmstra, 2018; De Boeck and Jeon, 2019), to detect rapid guessing and cheating behavior (van der Linden and Guo, 2008; van der Linden, 2009; Wang and Xu, 2015; Pokropek, 2016; Qian et al., 2016; Skorupski and Wainer, 2017; Wang et al., 2018a,b; Lu et al., 2019; Sinharay and Johnson, 2019; Zopluoglu, 2019), to evaluate the speededness of tests (Schnipke and Scrams, 1997; van der Linden et al., 2007), and to design more efficient tests (Bridgeman and Cline, 2004; Chang, 2004; Choe et al., 2018).

1.1. Advantages of Our Model Over Traditional Response Time Models in Educational Psychology Research

Although response times in both educational and psychological research have been studied widely and in depth, there are still some deficiencies in the existing literature. Here, we compare existing response time models with our new model and analyze the advantages of our model from multiple aspects.

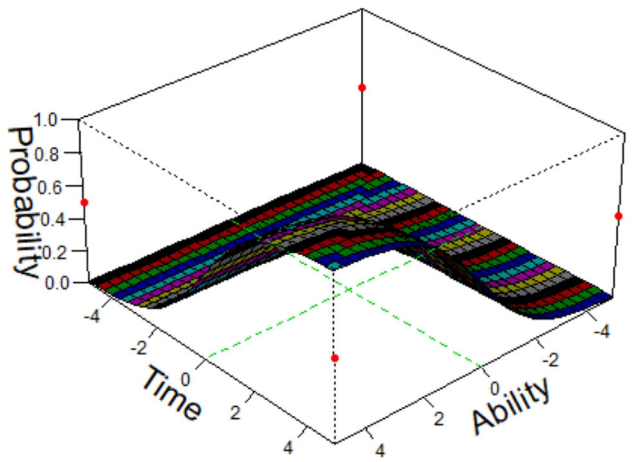

FIGURE 1 | The 3-D diagram of ability, time, and correct response probability.

Thissen (1983) proposed a joint model of response time and accuracy to describe the speed-accuracy relationship. In his model, the speed-accuracy trade-off is reflected by letting response accuracy depend on the time devoted to an item: spending more time on an item increases the probability of a correct response. Thissen's joint model can be expressed as follows:

$$\log T_{ij} = u + \eta_i + \varsigma_j - \rho(a_j\theta_i - b_j) + \varepsilon_{ij},$$

where T_{ij} is the response time of the ith examinee answering the jth item, u is a general intercept parameter, η_i and ς_j can be interpreted, respectively as the speed of examinee i and the amount of time required by item j, ρ is a regression parameter, a_j and b_j are, respectively the item discrimination and difficulty parameters, θ_i is the ability parameter for the ith examinee, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The speed–accuracy trade–off is represented by the term $a_j\theta_i - b_j$ when $\rho < 0$. When $\rho > 0$, the speed-accuracy relation is reversed. However, the way in which this model incorporates personal-level and item-level parameters means that it is unable to fully reflect the direct impact of the response time on the correct-response probability. Our new model solves this problem. The response time and the ability and item difficulty parameters are combined in an item response model that reflects the way in which the interactions among the three factors influences the correct-response probability. To provide an intuitive explanation, we use a three-dimensional diagram (**Figure 1**) to illustrate the effect of the ability and response time on the correct-response probability. A similar modeling method was proposed by Verhelst et al. (1997).

Roskam (1987, 1997) proposed a Rasch response time model integrating response time and correctness. According to this model, the probability of a correct response for the ith examinee answering the jth item can be written as

$$p(Y_{ij} = 1 \mid T_{ij}, i, j) = \frac{\theta_i T_{ij}}{\theta_i T_{ij} + \delta_j} = \frac{\exp(\xi_i + \tau_{ij} - \kappa_j)}{1 + \exp(\xi_i + \tau_{ij} - \kappa_j)},$$

where Y_{ij} denotes the response of the ith examinee answering the jth item, θ_i is the ability parameter for the ith examinee. δ_j is the item difficulty parameter for the jth item, and ξ_i, τ_{ij}, and κ_j are the logarithms of θ_i, T_{ij}, and δ_j, respectively. We can see that when T_{ij} goes to infinity, the correct-response probability $p(Y_{ij} = 1 \mid T_{ij}, i, j)$ approaches 1, no matter how difficult the item is. In fact, this type of model can only be applied to speeded tests, because a basic characteristic of such tests is that test items are quite easy, so, with unlimited time available, the answers are almost always correct. However, our new model is designed for a power test. This means that even if the examinees are given enough time, they cannot be sure to answer an item correctly, but rather they answer the item correctly with the probability of a three-parameter logistic (3PL) model.

Although there is some similarity between our model and the item response model proposed by Wang and Hanson (2005) with regard to the incorporation of response time into the traditional 3PL model, there are some major differences in concept and construction. Wang and Hanson give the probability of a correct response to item j by examinee i as

$$p(Y_{ij} = 1 \mid a_j, b_j, c_j, d_j, \theta_i, \eta_i, T_{ij}) = c_j + \frac{1 - c_j}{1 + \exp[-1.7a_j(\theta_i - b_j - \eta_i d_j / T_{ij})]},$$

where a_j, b_j, and c_j are, respectively the item discrimination, difficulty, and guessing parameters for the jth item, as in the regular 3PL model. θ_i and η_i are, respectively the ability and slowness parameters for the ith examinee, and d_j is the slowness parameter for the jth item. The item and personal slowness parameters determine the rate of increase in the probability of a correct answer as a function of response time. We will now analyze the differences between the two models.

From the perspective of model construction, the response time and the item and personal parameters are all incorporated into the same exponential function in Wang and Hanson's model, namely, $\exp[-1.7a_j(\theta_i - b_j - \eta_i d_j / T_{ij})]$, whereas in our model, the parameters and time effect appear in two different exponential functions (see the following section for a detailed description of the model): $\exp[-1.7a_j(\theta_i - b_j)] + \exp(-t^*_{ij})$. Our model considers not only the influence of the personal and item factors on the correct-response probability, but also that of the time effect. In Wang and Hanson's model, two slowness parameters associated with persons and items are introduced on the basis of the traditional 3PL model, which increases the complexity of the model. The model can be identified only by imposing stronger constraints on the model parameters. The accuracy of parameter estimation may be reduced owing to the increase in the number of model parameters. However, in our model, no such additional parameters related to items and persons are introduced, and therefore the model is more concise and easy to understand. In terms of model identifiability, our model is similar to the traditional 3PL model in that no additional restrictions need to be imposed. More importantly, parameter estimation becomes more accurate because of the addition of time information. Besides the personal ability parameter, a personal slowness parameter is included Wang and Hanson's model. In fact, their model is

a multidimensional item response theory model incorporating response time. In their model, it is assumed that these two personal parameters are independent, but this assumption may not necessarily be true in practice. For example, the lower a person's ability, the slower is their response. That is to say, there is a negative correlation between the ability parameter and the slowness parameter. More research is needed to verify this. Like other models based on the traditional 3PL model (see the next subsection), Wang and Hanson's model cannot distinguish between different abilities under different time intensities when examinees have the same response framework. However, our new model can deal with this problem very well.

In addition, our model introduces the concept of a time weight. Depending on the importance of a test (e.g., whether it is a high-stakes or a low-stakes test), the effect of the time constraint on the whole test is characterized by a time weight. This is something that cannot be dealt with by Wang and Hanson's model.

van der Linden (2007) proposed a hierarchical framework in which responses and response times are modeled separately at the measurement model level, while at a higher level, the ability and speed parameters are included in a population model to account for the correlation between them. In his approach, the latent speed parameter directly affects the response time, while the speed parameters and ability parameters are linked by the hierarchical model. It is known that in item response theory models, ability has a direct impact on the correct-response probability. Thus, we can see that the correct-response probability is related to the response time via the personal parameters (speed and ability). Van der Linden's hierarchical modeling method is unrealistic in that it includes the response time and the ability parameters in the item response model, whereas our model represents the relationships among response time, ability, and correct-response probability more simply and directly. Several other models have a similar structure to van der Linden's hierarchical model, including those of Fox et al. (2007), Klein Entink et al. (2009a,b), van der Linden and Glas (2010), Marianti et al. (2014), Wang and Xu (2015), Wang et al. (2018a), Fox and Marianti (2016), and Lu et al. (2019).

1.2. Advantages of Our Model Compared With the Traditional 3PL Model

Item response theory (IRT) models have been extensively used in educational testing and psychological measurement (Lord and Novick, 1968; van der Linden and Hambleton, 1997; Embretson and Reise, 2000; Baker and Kim, 2004). The most popular IRT model that includes guessing is the 3PL model (Birnbaum, 1968), which has been discussed in many papers and books (see e.g., Hambleton et al., 1991; van der Linden and Hambleton, 1997; Baker and Kim, 2004; von Davier, 2009; Han, 2012). However, several studies have revealed that the 3PL model has technical and theoretical limitations (Swaminathan and Gifford, 1979; Zhu et al., 2018). In this paper, we focus on another defect of the traditional 3PL model, namely, that it cannot distinguish between different abilities under different time intensities when the examinees have the same response framework. Here, we give a simulation example to illustrate the shortcomings of the traditional 3PL model and the advantages of our model (which is a general three-parameter logistic model with response time: G3PLT). We assume that 24 examinees answer three items and that the examinees can be divided into three groups of eight, with the examinees in each group having response frameworks $(1, 0, 0)$, $(0, 1, 0)$, and $(1, 1, 0)$, respectively. Here, 0 indicates that the item is answered correctly and 1 indicating that it is answered incorrectly. The item parameters of the three items are calibrated in advance and known. The discrimination, difficulty, and guessing parameters are set as in **Table 1**.

TABLE 1 | The setting of the true values of discrimination, difficulty, and guessing parameters.

Item	Discrimination	Difficulty	Guessing
1	0.8	−1	0
2	1	0	0.05
3	1.2	1	0.1

To consider the influence of different time effects on the ability of the examinees, eight time transformation values are considered: −0.2, 0.2, 0.5, 1, 2, 3, and 8. The specific settings for the time transformation values can be found in section 2. **Table 2** shows the estimated ability values from the 3PL model and from our model under different response frameworks, with the maximum likelihood method being used to estimate the ability parameter.

The following conclusions can be drawn from **Table 2**.

1. The estimated ability under the G3PLT model with the same response framework will gradually increase as the transformed time decreases from 8 to −0.2. This indicates that the examinees may have different proficiencies in responding to items. Less time is taken if the examinee has greater ability. The time effect captures exactly the information that the traditional 3PL model cannot provide. Specifically, the 3PL model cannot distinguish between abilities when there are different response times under the same response framework.
2. As an illustration, we consider the case where the transformed time is −0.2. The ability estimates under the three response frameworks $(1, 0, 0)$, $(0, 1, 0)$, and $(1, 1, 0)$ are −0.8863, 0.1408, and 1.3109, respectively. We find that the more difficult the item and the greater the number of items answered correctly, the higher are the ability estimates. Without considering the time effect, the ability estimates based on the 3PL model under the three response frameworks are −0.9339, −0.7207, and 0.6659, respectively.
3. Under the three response frameworks, the ability estimates obtained from the G3PLT model and the 3PL model are almost the same when the transformed time reaches 8. This indicates that even if the examinees are allowed enough time, they cannot be certain of answering an item correctly, but can do so only with the correct-response probability given by the 3PL model.

TABLE 2 | The comparisons of ability estimates under the frameworks of 3PL model and G3PLT model.

Examinees	Fitting model	Response framework	Transformated time t^*	Estimation of ability
1		(1, 0, 0)	−0.2	−0.8863
2		(1, 0, 0)	0	−0.8970
3		(1, 0, 0)	0.2	−0.9052
4	G3PLT	(1, 0, 0)	0.5	−0.9142
5		(1, 0, 0)	1	−0.9232
6		(1, 0, 0)	2	−0.9305
7		(1, 0, 0)	3	−0.9327
8		(1, 0, 0)	8	−0.9339
–	3PL	(1, 0, 0)	–	−0.9339
9		(0, 1, 0)	−0.2	0.1408
10		(0, 1, 0)	0	0.0614
11		(0, 1, 0)	0.2	−0.0139
12	G3PLT	(0, 1, 0)	0.5	−0.1233
13		(0, 1, 0)	1	−0.2945
14		(0, 1, 0)	2	−0.5397
15		(0, 1, 0)	3	−0.6515
16		(0, 1, 0)	8	−0.7202
–	3PL	(0, 1, 0)	–	−0.7207
17		(1, 1, 0)	−0.2	1.3109
18		(1, 1, 0)	0	1.0990
19		(1, 1, 0)	0.2	0.9791
20	G3PLT	(1, 1, 0)	0.5	0.8706
21		(1, 1, 0)	1	0.7752
22		(1, 1, 0)	2	0.7016
23		(1, 1, 0)	3	0.6785
24		(1, 1, 0)	8	0.6660
–	3PL	(1, 1, 0)	–	0.6659

We now give another example to further explain the advantages of the G3PLT model. Under the condition that the correct-response probability is the same, we consider the response times of examinees i and j when they answer the same item, and we find that these are 1 and 2 min, respectively. In general, we think that the examinee with shorter response times has a higher ability. Thus, here the ability of examinee i should be higher than that of examinee j. However, since the 3PL model does not consider response time, the difference in ability cannot be distinguished. This problem can be solved by using the G3PLT model. Because this model takes into account the information provided by response time, it can estimate the ability of examinees more objectively and accurately. As shown in **Figure 2**, for the same item, L1 represents the item characteristic curve corresponding to the case where examinees need a long response time ($t_1^* = 4.41$), and L2 represents the item characteristic curve corresponding to the case where examinees need a short response time ($t_1^* = 1.94$). When $p = 0.86$ is given as the correct-response probability, the estimated ability under L1 is 0, while the estimated ability under L2 is 0.88. Therefore, according to the evaluation results from the G3PLT model, the examinees with shorter times should have higher abilities, whereas the 3PL model is unable to distinguish between the two cases. In addition, it can be seen from the figure that when the ability is fixed at 0, the probabilities of a correct response under the two characteristic curves L1 and L2 are 0.86 and 0.52, respectively. This indicates that under the same ability condition, the correct-response probability of the examinees with short response times is lower than that of the examinees with long response times.

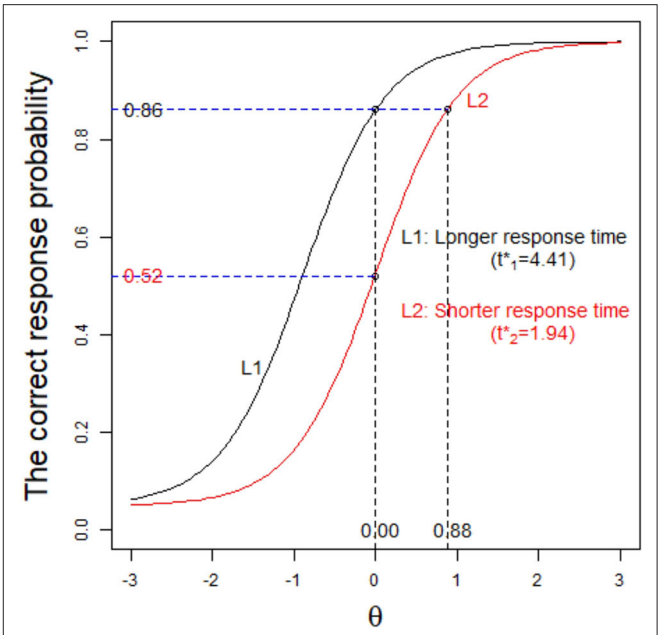

FIGURE 2 | The item characteristic curve based on different time intensities.

The remainder of this paper is organized as follows. Section 2 presents a detailed introduction to the proposed G3PLT model. Section 3 provides a computational strategy based on a Metropolis–Hastings within Gibbs sampling algorithm to meet computational challenges for the proposed model. Two Bayesian model comparison criteria are also discussed in section 3. In section 4, simulation studies are conducted to examine the performance of parameter recovery using the Bayesian algorithm and to assess model fit using the deviance information criterion (DIC) and the logarithm of the pseudomarginal likelihood (LPML). A real data analysis based on the Program for International Student Assessment (PISA) is presented in section 5. We conclude with a brief discussion and suggestions for further research in section 6.

2. THE MODEL AND ITS IDENTIFICATION

2.1. The General Three-Parameter Logistic Model With Response Time (G3PLT)

Let the examinees be indexed by $i = 1, 2, \ldots, N$ and the items by $j = 1, 2, \ldots, J$. Let θ denote the parameters representing the effects of the abilities of the examinees, and let a_j, b_j, and c_j denote

the item effects, which are generally interpreted, respectively as discrimination power, difficulty, and success probability in the case of random guessing. If Y_{ij} denotes the response of the ith examinee answering the jth item, then the corresponding correct-response probability can be expressed as

$$\begin{aligned} p_{ij} &= p(Y_{ij} = 1 \mid a_j, b_j, c_j, \theta_i, t_{ij}^*) \\ &= c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)] + \exp(-t_{ij}^*)}, \end{aligned} \tag{2.1}$$

where D is a constant equal to 1.7. The influence of the time effect on the probability is described by the term $\exp(-t_{ij}^*)$.

2.2. Time Transformation Function

It is obvious that when the response time of each item is very short, the correct-response probability of an item is reduced. In addition, we know that it is impossible for an examinee to answer an item 100% correctly even if they are given enough time to think about the item, and this can be attributed to limitations of the examinee's ability. When examinees are given enough time to answer each item, our model will reduce to the traditional 3PL model, and each item is answered correctly with the corresponding 3PL model correct-response probability. To make the model fully represent the requirement that the correct-response probability varies with time and to eliminate the effects of different average response times for each item in different tests, we consider the following time transformation:

$$t_{ij}^* = f(t_{ij}) = \frac{\log t_{ij} - \mu_t}{\sigma_t} + W, \tag{2.2}$$

where μ_t is the logarithm of the average time spent by all examinees in answering all items, and σ_t is the corresponding standard deviation. W denotes the time weight, which is equal to zero or a positive integer. From the simulation study and real data analysis, we find that the G3PLT model reduces to the traditional 3PL model when the time weight increases to 8, and therefore we restrict the weight to values in the range 0–8. An increase in the time weight indicates that the time factor of the test has a small influence on the correct-response probability of the examinee.

Proposition 1. *Suppose that the correct-response probability* $p(Y_{ij} = 1 \mid a_j, b_j, c_j, \theta_i, t_{ij}^*)$ *is given by Equation (2.1). Then, we have the following results:*

1. *As the transformed time* $t_{ij}^* \to +\infty$*, the G3PLT model reduces to the 3PL model. That is,*

$$p_{ij} \to c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)]}. \tag{2.3}$$

In other words, it is impossible for the examinee to answer the item 100% correctly even if they are given enough time to think about the item, which can be attributed to the limitations of the examinee's ability.

2. *As the transformed time* $t_{ij}^* \to -\infty$ *(the original time* $t_{ij} \to 0$*), the correct-response probability of the G3PLT model tends to zero. That is,*

$$p_{ij} = c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j) + \exp(-t_{ij}^*)]} \downarrow 0. \tag{2.4}$$

When there is not enough time to answer items (e.g., at the end of the examination), any item answered by the examinee must be one that requires only a very short time to finish. As the response time continues to shorten, the correct-response probability is reduced.

3. *The G3PLT model can be reduced to a G2PLT model by constraining the lower asymptote parameter* c_j *to be zero, and a G1PLT model can be obtained by further constraining* a_j *to be the same across all items.*

2.3. Asymptotic Properties of the Model

Let p_j be the correct-response rate for the jth item. When the transformed time $t_{ij}^* \to +\infty$, the model in Equation (2.1) can be written as

$$\begin{aligned} &\lim_{t_{ij}^* \to +\infty} \left\{ c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)] + \exp(-t_{ij}^*)} \right\} \\ &\quad = c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)]} = p_j. \end{aligned} \tag{2.5}$$

The ability can be obtained as

$$\theta_i = b_j - \frac{1}{Da_j} \log\left(\frac{1 - p_j}{p_j - c_j} \right). \tag{2.6}$$

Next, we will use a specific example to explain the meaning of Equations (2.5) and 2.6. Assuming that $p_j = 0.5$, $a_j = 1.5$, $b_j = 1$, and $c_j = 0.1$, we obtain $\theta_i = 0.8$ from Equation (2.6). This result indicates that even if examinee i has sufficient response time to finish item j, the examinee's ability should be at least 0.8 (the intersection of the vertical asymptote and the x-axis in **Figure 3**) if the correct response probability reaches 0.5; otherwise, no matter how long a response time is allowed, the examinee's correct-response probability cannot reach 0.5. This is like a primary school pupil attempting to solve a college math problem, because the pupil's ability is so low that no matter how much time he is given, he cannot get a correct answer to item j other than by guessing. Moreover, when the ability $\theta_i \to +\infty$, the model in Equation (2.1) can be written as

$$\begin{aligned} &\lim_{\theta_i \to +\infty} \left\{ c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)] + \exp(-t_{ij}^*)} \right\} \\ &\quad = c_j + \frac{1 - c_j}{1 + \exp(-t_{ij}^*)} = p_j. \end{aligned} \tag{2.7}$$

The transformed time t_{ij}^* can be obtained as

$$t_{ij}^* = -\log\left(\frac{1 - p_j}{p_j - c_j} \right). \tag{2.8}$$

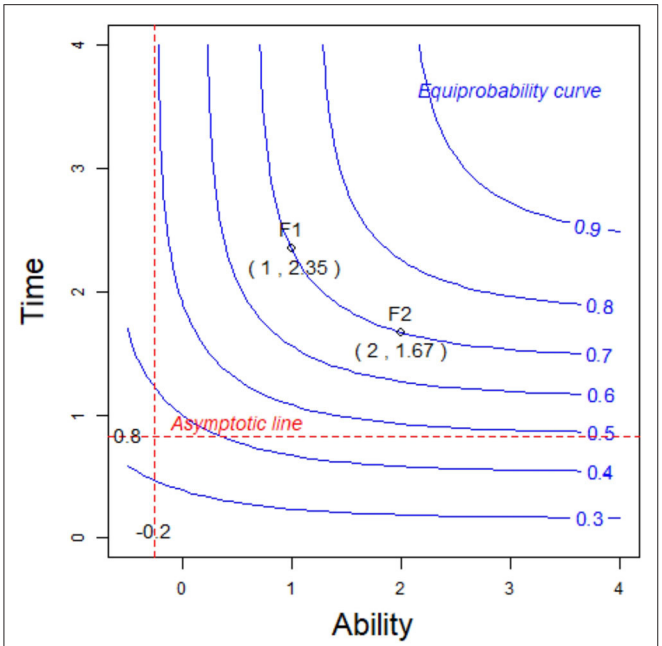

FIGURE 3 | The equiprobability curve based on the ability and time.

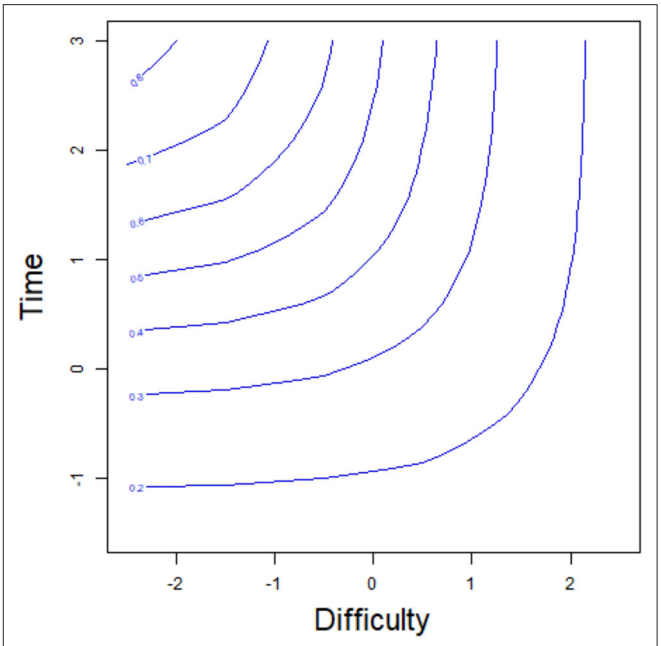

FIGURE 4 | The equiprobability curve based on the item difficulty and time.

We again assume that $p_j = 0.5$, $a_j = 1.5$, $b_j = 1$, and $c_j = 0.1$. From (2.8), the transformed time t_{ij}^* is about -0.2. This result indicates that even if the examinee i has a strong ability, the transformed time required to answer item j should not be less than -0.2 (the intersection of the horizontal asymptote and the y-axis in **Figure 3**) if the correct-response probability reaches 0.5; otherwise, no matter how strong the ability of the examinee, it is impossible to reach a correct-response probability of 0.5. This is like a college student solving a primary school math problem. Although the college student's ability is very strong, she cannot finish the item in a very short time. In addition, the correct-response probability of the examinees is the same for two points on the equiprobability curve. For example, for the two examinees F1 and F2 with the same correct-response probability 0.7 in **Figure 3**, the examinee F1 with low ability (1) takes a long time (2.35), while the response time (1.67) of the examinee F2 with high ability (2) is short to obtain the same correct-response probability. Similarly, the equiprobability curve based on item difficulty and time is shown in **Figure 4**. The correct-response probability is the same for two points on the equiprobability curve. The item with high difficulty takes a long time, while the response time of the item with low difficulty is short, giving the same correct-response probability.

2.4. Model Identification

To ensure identification of the G3PLT model, either the scale of latent traits or the scale of item parameters has to be restricted (Birnbaum, 1968; Lord, 1980; van der Linden and Hambleton, 1997). In this paper, we set the mean and variance of the latent traits to zero and one, respectively (Bock and Aitkin, 1981). The mean of the latent trait is fixed to remove the trade-off between θ_i and b_j in location, and the variance of the latent trait is fixed to remove the trade-off among θ_i, b_j, and a_j in scale.

3. BAYESIAN INFERENCE

3.1. Prior and Posterior Distributions

In a Bayesian framework, the posterior distribution of the model parameters is obtained based on the observed data likelihood (sample information) and prior distributions (prior information). In general, these two kinds of information have an important influence on the posterior distribution. However, in large-scale educational assessment, the number of examinees is often very large. Therefore, the likelihood information plays a dominant role, and the selection of different priors (informative or non-informative) has no significant influence on the posterior inference (van der Linden, 2007; Wang et al., 2018a). Based on previous results (Wang et al., 2018a), we adopt the informative prior distribution to analyze the following simulation studies and real data. The specific settings are as follows. For the latent ability, we assume a standardized normal prior, i.e., $\theta_i \sim N(0,1)$ for $i = 1,\ldots,N$. The prior distribution for the discrimination parameter a_j is a lognormal distribution, i.e., $a_j \sim \log N(0,1)$ for $j = 1,\ldots,J$. The prior distribution for the difficulty parameter b_j is a standardized normal distribution, i.e., $b_j \sim N(0,1)$ for $j = 1,\ldots,J$. For the guessing parameter, we assume a Beta distribution, i.e., $c_j \sim \text{Beta}(2,10)$ for $j = 1,\ldots,J$. Then, the joint posterior distribution of the parameters given the data is as

follows:

$$p(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \mid \boldsymbol{Y}, \boldsymbol{T}) \propto \left[\prod_{i=1}^{N}\prod_{j=1}^{J} p(Y_{ij} \mid \theta_i, a_j, b_j, c_j, T_{ij})\right]\prod_{i=1}^{N} p(\theta_i) \times \prod_{j=1}^{J} p(a_j)p(b)p(c_j). \tag{3.1}$$

3.2. Bayesian Estimation

Bayesian methods have been widely applied to estimate parameters in complex IRT models (see e.g., Albert, 1992; Patz and Junker, 1999a,b; Béguin and Glas, 2001; Rupp et al., 2004). In this study, the Metropolis-Hastings within Gibbs algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994; Chib and Greenberg, 1995; Chen et al., 2000) is used to draw samples from the full conditional posterior distributions because the parameters of interest do not have conjugate priors within the framework of the IRT model.

Detailed MCMC Sampling Process

Step 1: Sample the ability parameter θ_i for the ith examinee. We independently draw θ_i^* from the normal proposal distribution, i.e., $\theta_i^* \sim N(\theta_i^{(r-1)}, v_\theta^2)$. The prior of θ_i is assumed to follow a normal distribution with mean μ_θ and variance σ_θ^2, i.e., $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$. Therefore, the acceptance probability is given by

$$\alpha(\theta_i^{(r-1)}, \theta_i^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_i \mid \theta_i^*, \boldsymbol{a}^{(r-1)}, \boldsymbol{b}^{(r-1)}, \boldsymbol{c}^{(r-1)}, \boldsymbol{T}_i)p_{\text{prior}}(\theta_i^* \mid \mu_\theta, \sigma_\theta^2)}{p(\boldsymbol{Y}_i \mid \theta_i^{(r-1)}, \boldsymbol{a}^{(r-1)}, \boldsymbol{b}^{(r-1)}, \boldsymbol{c}^{(r-1)}, \boldsymbol{T}_i)p_{\text{prior}}(\theta_i^{(r-1)} \mid \mu_\theta, \sigma_\theta^2)}\right\}. \tag{3.2}$$

Otherwise, the value of the preceding iteration is retained, i.e., $\theta_i = \theta_i^{(r-1)}$. Here, $\boldsymbol{Y}_i = (Y_{i1}, Y_{ti2}, \ldots, Y_{iJ})$, $\boldsymbol{T}_i = (Y_{i1}, Y_{ti2}, \ldots, Y_{iJ})$, $\boldsymbol{a} = (a_1, a_2, \ldots, a_J)$, $\boldsymbol{b} = (b_1, b_2, \ldots, b_J)$, and $\boldsymbol{c} = (c_1, c_2, \ldots, c_J)$. In Equation (3.3), $p(\boldsymbol{Y}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{T}_i) = \prod_{j=1}^{J}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$, where p_{ij} is given in Equation (2.1).

Step 2: Sample the difficulty parameter b_j for the jth item. We independently draw b_j^* from the normal proposal distribution, i.e., $b_j^* \sim N(b_j^{(r-1)}, v_j^2)$. The prior of b_j is assumed to follow a normal distribution with mean μ_b and variance σ_b^2, i.e., $b_j \sim N(\mu_b, \sigma_b^2)$. The acceptance probability is given by

$$\alpha(b_j^{(r-1)}, b_j^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}^{(r)}, a_j^{(r-1)}, b_j^*, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(b_j^* \mid \mu_b, \sigma_b^2)}{p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}^{(r)}, a_j^{(r-1)}, b_j^{(r-1)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(b_j^{(r-1)} \mid \mu_b, \sigma_b^2)}\right\}. \tag{3.3}$$

Otherwise, the value of the preceding iteration is retained, i.e., $b_j = b_j^{(r-1)}$. Here, $\boldsymbol{Y}_j = (Y_{1j}, Y_{2j}, \ldots, Y_{Nj})$, $\boldsymbol{T}_j = (T_{1j}, T_{2j}, \ldots, T_{Nj})$, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$. In Equation (3.3), $p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}, a_j, b_j, c_j, \boldsymbol{T}_j) = \prod_{i=1}^{n}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$.

Step 3: Sample the discrimination parameter a_j for the jth item. We independently draw a_j^* from the log-normal proposal distribution, i.e., $a_j^* \sim \log N(\log a_j^{(r-1)}, v_a^2)$. In addition, $p_{\text{prior}}(a_j)$ is a lognormal prior distribution, i.e., $a_j \sim \log N(\mu_a, \sigma_a^2)$. The acceptance probability is given by

$$\alpha(a_j^{(r-1)}, a_j^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}^{(r)}, a_j^*, b_j^{(r)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(a_j^* \mid \mu_a, \sigma_a^2)a_j^*}{p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}^{(r)}, a_j^{(r-1)}, b_j^{(r)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(a_j^{(r-1)} \mid \mu_a, \sigma_a^2)a_j^{(r-1)}}\right\}. \tag{3.4}$$

Otherwise, the value of the preceding iteration is retained, i.e., $a_j = a_j^{(r)}$. In Equation (3.4), $(\boldsymbol{Y}_j \mid \boldsymbol{\theta}, a_j, b_j, c_j, \boldsymbol{T}_j) = \prod_{i=1}^{n}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$.

Step 4: Sample the guessing parameter c_j for the jth item. We independently draw c_j^* from the uniform proposal distribution, i.e., $c_j^* \sim U(c_j^{(r-1)} - 0.01, c_j^{(r-1)} + 0.01)$. The prior of c_j is assumed to follow a Beta distribution, i.e., $c_j \sim \text{Beta}(\upsilon_1, \upsilon_2)$. Therefore, the acceptance probability is given by

$$\alpha(c_j^{(r-1)}, c_j^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^*, \boldsymbol{T}_j)p_{\text{prior}}(c_j^* \mid \upsilon_1, \upsilon_2)}{p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(c_j^{(r-1)} \mid \upsilon_1, \upsilon_2)}\right\}. \tag{3.5}$$

Otherwise, the value of the preceding iteration is retained, i.e., $c_j = c_j^{(r)}$. In Equation (3.5), $p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}, a_j, b_j, c_j, \boldsymbol{T}_j) = \prod_{i=1}^{n}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$.

3.3. Bayesian Model Assessment

Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) for model comparison when the number of parameters is not clearly defined. The DIC is an integrated measure of model fit and complexity. It is defined as the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity. We write $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{ij},\ i = 1, \ldots, N,\ j = 1, \ldots, J)$, where $\boldsymbol{\Omega}_{ij} = (\theta_i, a_j, b_j, c_j)'$. Let $\{\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(R)}\}$, where $\boldsymbol{\Omega}^{(r)} = (\boldsymbol{\Omega}_{ij}^{(r)},\ i = 1, \ldots, N,\ j = 1, \ldots, J)$, $\boldsymbol{\Omega}_{ij}^{(r)} = (\theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)})'$ for $i = 1, \ldots, N, j = 1, \ldots, J$, and $r = 1, \ldots, R$, denote an Markov chain Monte Carlo (MCMC) sample from the posterior distribution in Equation (3.1). The joint likelihood function of the responses can be written as

$$L(\boldsymbol{Y} \mid \boldsymbol{\Omega}, \boldsymbol{T}) = \prod_{i=1}^{N}\prod_{j=1}^{J} f(y_{ij} \mid \theta_i, a_j, b_j, c_j, t_{ij}), \tag{3.6}$$

where $f(y_{ij} \mid \theta_i, a_j, b_j, c_j, t_{ij})$ is the response probability of the G3PLT model. The logarithm of the joint likelihood function in Equation (3.6) evaluated at $\boldsymbol{\Omega}^{(r)}$ is given by

$$\log L(\boldsymbol{Y} \mid \boldsymbol{\Omega}^{(r)}, \boldsymbol{T}) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij}). \tag{3.7}$$

The joint log-likelihoods for the responses, $\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij})$, $i = 1, \ldots, N$ and $j = 1, \ldots, J$, are readily available from MCMC sampling outputs, and therefore $\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij})$ in Equation (3.7) is easy to compute. The effective number of parameters in the models is defined by

$$p_D = \overline{\text{Dev}(\boldsymbol{\Omega})} - \text{Dev}(\widehat{\boldsymbol{\Omega}}), \tag{3.8}$$

where $\overline{\text{Dev}(\boldsymbol{\Omega})}$ is a Monte Carlo estimate of the posterior expectation of the deviance function $\text{Dev}(\boldsymbol{\Omega}) = -2\log L(\boldsymbol{Y} \mid \boldsymbol{\Omega}, \boldsymbol{T})$, and the term $\text{Dev}(\widehat{\boldsymbol{\Omega}})$ is computed by plugging the mean of the simulated values of $\boldsymbol{\Omega}$ into $\text{Dev}(\cdot)$, where $\widehat{\boldsymbol{\Omega}} = \sum_{r=1}^{R} \boldsymbol{\Omega}^{(r)}/R$. More specifically,

$$\begin{aligned} \overline{\text{Dev}(\boldsymbol{\Omega})} &= -\frac{2}{R}\sum_{r=1}^{R} \log L(\boldsymbol{Y} \mid \boldsymbol{\Omega}^{(r)}), \\ \text{Dev}(\widehat{\boldsymbol{\Omega}}) &= -2\log L(\boldsymbol{Y} \mid \widehat{\boldsymbol{\Omega}}). \end{aligned} \tag{3.9}$$

The DIC can now be formulated as follows:

$$\text{DIC} = \widehat{\text{Dev}(\boldsymbol{\Omega})} + 2p_D = \widehat{\text{Dev}(\boldsymbol{\Omega})} + 2\left[\overline{\text{Dev}(\boldsymbol{\Omega})} - \widehat{\text{Dev}(\boldsymbol{\Omega})}\right], \tag{3.10}$$

A model with a smaller DIC value fits the data better.

Another method is to use the logarithm of the pseudomarginal likelihood (LPML) (Geisser and Eddy, 1979; Ibrahim et al., 2001) to compare different models. This is also based on the log-likelihood functions evaluated at the posterior samples of model parameters. The detailed calculation process is as follows.

We let $U_{ij,\max} = \max_{1 \le r \le R}[-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij})]$, and a Monte Carlo estimate of the conditional predictive ordinate (CPO) (Gelfand et al., 1992; Chen et al., 2000) is then given by

$$\begin{aligned} \log(\widehat{\text{CPO}_{ij}}) &= -U_{ij,\max} \\ &- \log\left\{\frac{1}{R}\sum_{r=1}^{R} \exp[-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij}) - U_{ij,\max}]\right\}. \end{aligned} \tag{3.11}$$

Note that the maximum value adjustment used in $\log(\widehat{\text{CPO}_{ij}})$ plays an important role in numerical stabilization in the computation of $\exp[-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij}) - U_{ij,\max}]$ in Equation (3.11). A summary statistic of the $\widehat{\text{CPO}_{ij}}$ is the sum of their logarithms, which is called the LPML and is given by

$$\text{LPML} = \sum_{i=1}^{N}\sum_{j=1}^{J} \log(\widehat{\text{CPO}_{ij}}). \tag{3.12}$$

A model with a larger LPML has a better fit to the data.

3.4. Accuracy Evaluation of Parameter Estimation

To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. In the following simulation study, 200 replications are used. Five indices are used to assess the accuracy of the parameter estimates. Let ϑ be the parameter of interest. Assume that $M = 200$ data sets are generated. Also, let $\widehat{\vartheta}^{(m)}$ and $\text{SD}^{(m)}(\vartheta)$ denote the posterior mean and the posterior standard deviation of ϑ obtained from the mth simulated data set for $m = 1, \ldots, M$.

The bias for the parameter ϑ is defined as

$$\text{Bias}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\vartheta}^{(m)} - \vartheta), \tag{3.13}$$

and the mean squared error (MSE) for ϑ is defined as

$$\text{MSE}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\vartheta}^{(m)} - \vartheta)^2. \tag{3.14}$$

The simulation SE is the square root of the sample variance of the posterior estimates over different simulated data sets. It is defined as

$$\text{Simulation SE}(\vartheta) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\vartheta}^{(m)} - \frac{1}{M}\sum_{\ell=1}^{M}\widehat{\vartheta}^{(\ell)}\right)^2}, \tag{3.15}$$

and the average of posterior standard deviation is defined as

$$\text{SD}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}\text{SD}^{(m)}(\vartheta). \tag{3.16}$$

The coverage probability based on the 95% highest probability density (HPD) intervals is defined as

$$\begin{aligned} &\text{CP}(\vartheta) \\ &= \frac{\#\text{ of 95\% (HPD) intervals containing } \vartheta \text{ in } M \text{ simulated data sets}}{M}. \end{aligned} \tag{3.17}$$

4. SIMULATION STUDY

4.1. Simulation 1

We conduct a simulation study to evaluate the recovery performance of the combined MCMC sampling algorithm based on different simulation conditions.

Simulation Design

The following manipulated conditions are considered: (a) test length $J = 20$ or 60 and (b) number of examinees $N = 500$, 1,000, or 2,000. Fully crossing different levels of these two factors yields six conditions (two test lengths × three sample sizes). Next, the true values of the parameters are given. True item discrimination parameters a_j are generated from a truncated normal distribution, i.e., $a_j \sim N(1, 0.2)\text{I}(a_j > 0)$, $j = 1, 2, \ldots, N$, where the indicator function $\text{I}(A)$ takes a value of 1 if A is true and a value of 0 if A is false. The item difficulty parameters b_j are generated from a standardized normal distribution. The item guessing parameters c_j are generated from a Beta distribution,

i.e., $c_j \sim \text{Beta}(2, 10)$. In addition, the ability parameters of the examinees, θ_i, are also generated from a standardized normal distribution. In each simulation condition, 200 replications (replicas) are considered. Next, we generate the response time data for each examinee based on the following facts:

1. The difficulty of each item has a direct impact on the response time. That is to say, the time spent on simple items is shorter, and the time spent on difficult items is longer.
2. In addition, the ability of each examinee also has a direct impact on the response time. That is to say, examinees with higher ability spend less time on an item.
3. Depending on the importance of the test (high-stakes test or low-stakes test), the effect of the time constraint on the whole test should be characterized by the time weighting.

In Wang and Xu (2015, p. 459), the average logarithms of the response times for each item based on the solution behavior follow a normal distribution. That is, $\log t_j \sim N(0.5, 0.25)$, $j = 1, 2, \ldots, J$, where the average time t_j spent on item j is about 1.64872 ($= e^{0.5}$) min. We take the standardized transformation $t_j^* = f(t_j) = (\log t_j - 0.5)/0.5$, so that $t_j^* \sim N(0, 1)$, where $-\infty < t_j^* < +\infty$.

Next, we consider the premise that the easier an item, the shorter is the response time. The true values of the difficulty parameter and the transformed time t_j^* for each item are arranged in order from small to large, i.e., $b_1 < b_2 < \cdots < b_{J-1} < b_J$ and $t_1^* < t_2^* < \cdots < t_{J-1}^* < t_J^*$. The corresponding item–time pairs can be written as $(b_1, t_1^*) < (b_2, t_2^*) < \cdots < (b_{J-1}, t_{J-1}^*) < (b_J, t_J^*)$. The response time of each examinee is generated from a normal distribution, i.e., $t_{ij}^* \sim N(t_j^*, 0.5)$, where $j = 1, \ldots, J$. Moreover, for a given item j, the premise that examinees with higher ability spend less time on the item needs to be satisfied. Therefore, we arrange $\theta_{1j} > \theta_{2j} > \cdots > \theta_{N-1,j} > \theta_{N,j}$, and $t_{1j}^* < t_{2j}^* < \cdots < t_{N-1,j}^* < t_{N,j}^*$. The corresponding ability–time pairs can be obtained by arranging the true values of the ability parameter and the transformed time t_{ij}^*, i.e., (θ_{ij}, t_{ij}^*). The time weights range from 0 to 8. The higher the value of the time weight, the weaker is the influence of the time factor of the test on the correct-response probability of the examinee. In this simulation study, we assume that the time factor of the test has an important influence on the correct-response probability of the examinee. Therefore, we set the time weight to 1 in this simulation. Based on the true values of the parameters and the response time data, the response data can be simulated using the G3PLT model given by Equation (2.1).

Convergence Diagnostics

To evaluate the convergence of the parameter estimations, we only consider convergence in the case of minimum sample sizes. That is, the test length is fixed at 20, and the number of examinees is 500. Two methods are used to check the convergence of our algorithm. One is the "eyeball" method to monitor convergence by visually inspecting the history plots of the generated sequences (Zhang et al., 2007; Hung and Wang, 2012), and the other is the Gelman–Rubin method (Gelman and Rubin, 1992; Brooks and Gelman, 1998) for checking the convergence of the parameters.

The convergence of the Bayesian algorithm is checked by monitoring the trace plots of the parameters for consecutive sequences of 10,000 iterations. The trace plots show that all parameter estimates stabilize after 5,000 iterations and then converge quickly. Thus, we set the first 5,000 iterations as the burn-in period. As an illustration, four chains started at overdispersed starting values are run for each replication. The trace plots of three randomly selected items are shown in **Figure 5**. In addition, we find that the potential scale reduction factor (PSRF) (Brooks and Gelman, 1998) values for all parameters are less than 1.2, which ensures that all chains converge as expected.

Recovery of Item Parameters

The average bias, MSE, SD, SE, and CP for discrimination, difficulty, and guessing parameters based on six different simulation conditions are shown in **Table 3**. The following conclusions can be drawn.

1. Given the total test length, when the number of individuals increases from 500 to 2,000, the average MSE, SD, and SE for the discrimination, difficulty, and guessing parameters show a decreasing trend. For example, for a total test length of 20 items, when the number of examinees increases from 500 to 2,000, the average MSE of all discrimination parameters decreases from 0.0088 to 0.0072, the average SE of all discrimination parameters decreases from 0.0022 to 0.0014, and the average SD of all discrimination parameters decreases from 0.0085 to 0.0066. The average MSE of all difficulty parameters decreases from 0.0436 to 0.0213, the average SE of all difficulty parameters decreases from 0.0272 to 0.0122, and the average SD of all difficulty parameters decreases from 0.0362 to 0.0143. The average MSE of all guessing parameters decreases from 0.0019 to 0.0013, the average SE of all guessing parameters decreases from 0.0007 to 0.0006, and the average SD of all guessing parameters decreases from 0.0013 to 0.0008.
2. The average SDs of the item parameters are larger than their average SEs. This indicates that the fluctuations of the posterior means of item parameters between different replications are small compared with their fluctuations within each replication.
3. Under the six simulated conditions, the average CPs of the discrimination, difficulty, and guessing parameters are about 0.950.
4. When the number of examinees is held fixed but the number of items increases from 20 to 40, the average MSE, SD, and SE show that the recovery results for the discrimination, difficulty and guessing parameters do not change much, which indicates that the Bayesian algorithm is stable and there is no reduction in accuracy due to an increase in the number of items.

In summary, the Bayesian algorithm provides accurate estimates of the item parameters for various numbers of examinees and items. Therefore, it can be used as a guide to practice.

Recovery of Ability Parameters

Next, we evaluate the recovery of latent ability from the plots of the true values and the estimates in **Figure 6**. For a fixed number

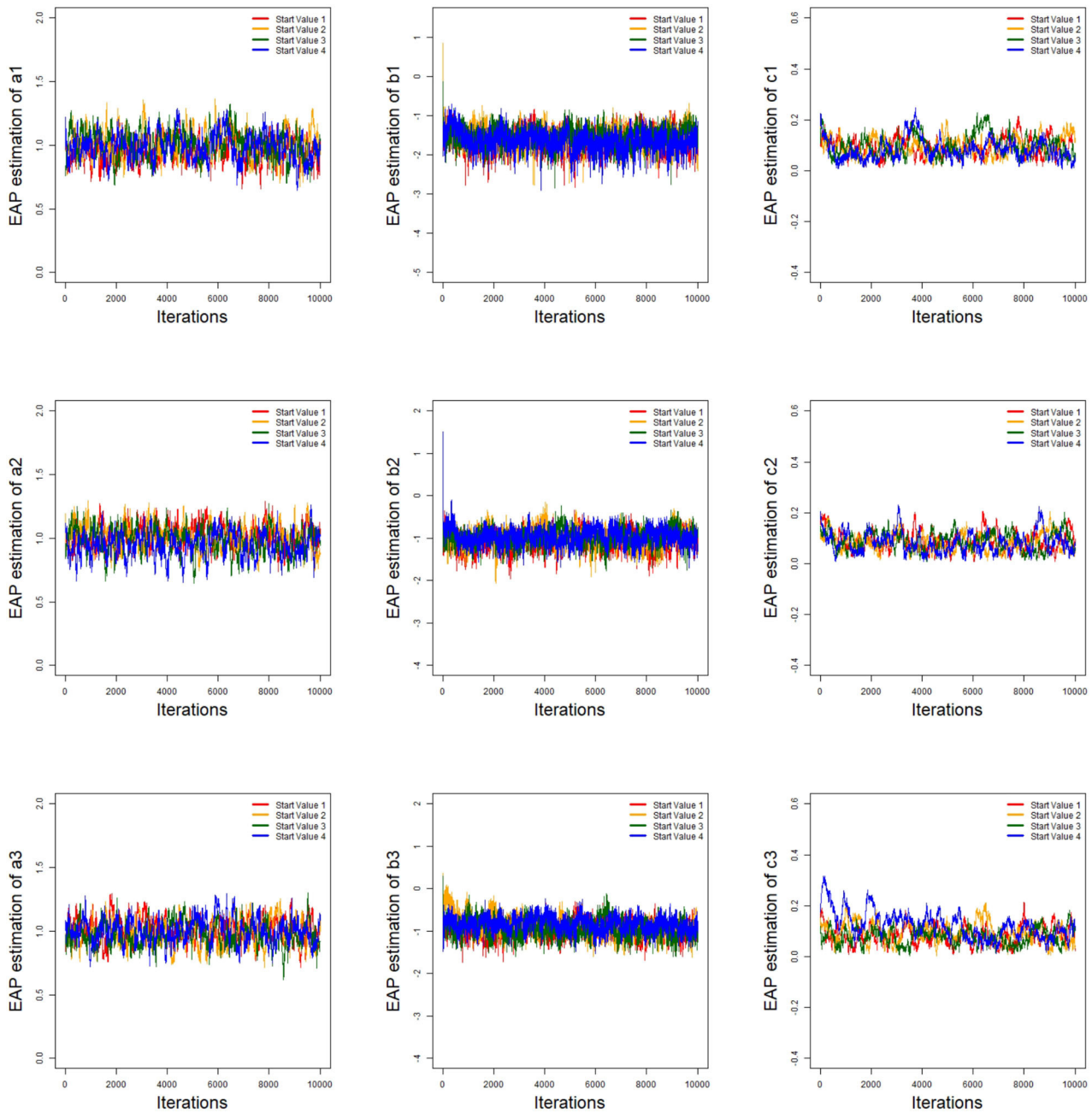

FIGURE 5 | The trace plots of three randomly selected items for the simulation study 1.

of examinees (500 or 1,000), when the number of items increases from 20 to 60, the ability estimates become more accurate, with the true values and the estimates basically lying on the diagonal line. Note that the estimated abilities are the average of 200 replication estimates. Because of the increase in the number of items, the probability of the situation in which all items are answered correctly by the high-ability examinees and incorrectly by the low-ability examinees, leading to a large deviation of the ability estimators, is reduced. Therefore, the estimated values and the true values of the ability at the end of the curve are closer to the diagonal line when the number of items is 60. In summary, the Bayesian sampling algorithm also provides accurate estimates of the ability parameters in term of the plots of the true values and the estimates.

4.2. Simulation 2

In this simulation study, we use the DIC and LPML model assessment criteria to evaluate model fitting. Two issues need further study. The first is whether the two criteria can accurately identify the true model that generates data from numerous fitting

TABLE 3 | Evaluating the accuracy of parameters based on six different simulated conditions in simulation study 1.

	No. of items=20											
	No. of examinees 500				No. of examinees 1,000				No. of examinees 2,000			
Item parameter	Bias	MSE	SE	SD	CP	Bias	MSE	SE	SD	CP	Bias	MSE	SE	SD	CP
Discrimination a	−0.0162	0.0088	0.0022	0.0085	0.9513	−0.0081	0.0083	0.0019	0.0076	0.9503	−0.0038	0.0072	0.0014	0.0066	0.9480
Difficulty b	−0.0134	0.0436	0.0272	0.0362	0.9385	−0.0103	0.0290	0.0166	0.0213	0.9410	−0.0068	0.0213	0.0122	0.0143	0.9287
Guessing c	−0.0031	0.0019	0.0007	0.0013	0.9315	−0.0026	0.0016	0.0006	0.0010	0.9378	−0.0014	0.0013	0.0006	0.0008	0.9283
	No. of items=60											
	No. of examinees 500				No. of examinees 1,000				No. of examinees 2,000			
Item parameter	Bias	MSE	SE	SD	CP	Bias	MSE	SE	SD	CP	Bias	MSE	SE	SD	CP
Discrimination a	0.0159	0.0082	0.0023	0.0082	0.9543	0.0132	0.0081	0.0019	0.0071	0.9393	0.0005	0.0074	0.0014	0.0059	0.9343
Difficulty b	−0.0345	0.0447	0.0245	0.0339	0.9574	−0.0112	0.0233	0.0153	0.0205	0.9497	−0.0086	0.0163	0.0098	0.0121	0.9296
Guessing c	0.0071	0.0016	0.0005	0.0011	0.9389	0.0061	0.0013	0.0005	0.0008	0.9328	0.0025	0.0011	0.0005	0.0006	0.9484

The Bias, MSE, SD, SE, and CP denote the average; Bias, MSE, SE, SD, and CP for the parameters. a *represents all discrimination parameters,* b *represents all difficulty parameters and* c *represents all guessing parameters.*

models. The second concerns the influence of different time weights in the G3PLT model on model fitting.

Simulation Design

In this simulation, the number of examinees is $N = 1,000$ and the test length is fixed at 20. Six item response models will be considered: the traditional 3PL model and the G3PLT model with time weights $W = 0$, 2, 4, 6, and 8. Thus, we evaluate the model fitting in the following five cases:

- Case 1. True model: G3PLT model with time weight 0 vs. Fitted model: 3PL model, G3PLT model with time weight 0.
- Case 2. True model: G3PLT model with time weight 2 vs. Fitted model: 3PL model, G3PLT model with time weight 2.
- Case 3. True model: G3PLT model with time weight 4 vs. Fitted model: 3PL model, G3PLT model with time weight 4.
- Case 4. True model: G3PLT model with time weight 6 vs. Fitted model: 3PL model, G3PLT model with time weight 6.
- Case 5. True model: G3PLT model with time weight 8 vs. Fitted model: 3PL model, G3PLT model with time weight 8.

The true values and prior distributions for the parameters are the same as in Simulation 1. To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. The results of Bayesian model assessment based on the 200 replications are shown in **Table 4**. Note that the following results for DIC and LPML are based on the average of 200 replications.

From **Table 4**, we find that when the G3PLT model with time weight 0 (G3PLT0) is the true model, the G3PLT0 model is chosen as the better-fitting model according to the results for DIC and LPML, which is what we expect to see. The medians of DIC and LPML are respectively 25 324.43 and −13231.77. The differences between the G3PLT0 model and 3PL model in the medians of DIC and LPML are −33.72 and 199.23, respectively. Similarly, when the G3PLT model with time weight 2 (G3PLT2) is the true model, the G3PLT2 model is also chosen as the better-fitting model according to the results for DIC and LPML. The medians of DIC and LPML are respectively 22 777.38 and −12221.93. The differences between the G3PLT2 model and 3PL model in the medians of DIC and LPML are −74.07 and 21.75, respectively. However, when the time weight increases from 4 to 8, the medians of DIC for the 3PL model and G3PLT model are basically the same. This shows that the 3PL model is basically the same as the G3PLT model with time weights 4, 6, and 8, which is attributed to the fact that the G3PLT model reduces to the traditional 3PL model when the time weight increases from 4 to 8. Based on the results for LPML, we find that the power of LPML to distinguish between the true G3PLT4 (6, 8) model and the 3PL model is stronger than that of DIC, because the LPMLs of the two models differ greatly. For example, the difference between the G3PLT8 model and 3PL model in the median of LPML is 46.45.

In summary, the two Bayesian model assessment criteria can accurately identify the true model that generates data. In addition, the process of transformation of the G3PLT model into the traditional 3PL model is also reflected by the differences in DIC and LPML. Therefore, the two Bayesian model assessment criteria are effective and robust and can guide practice.

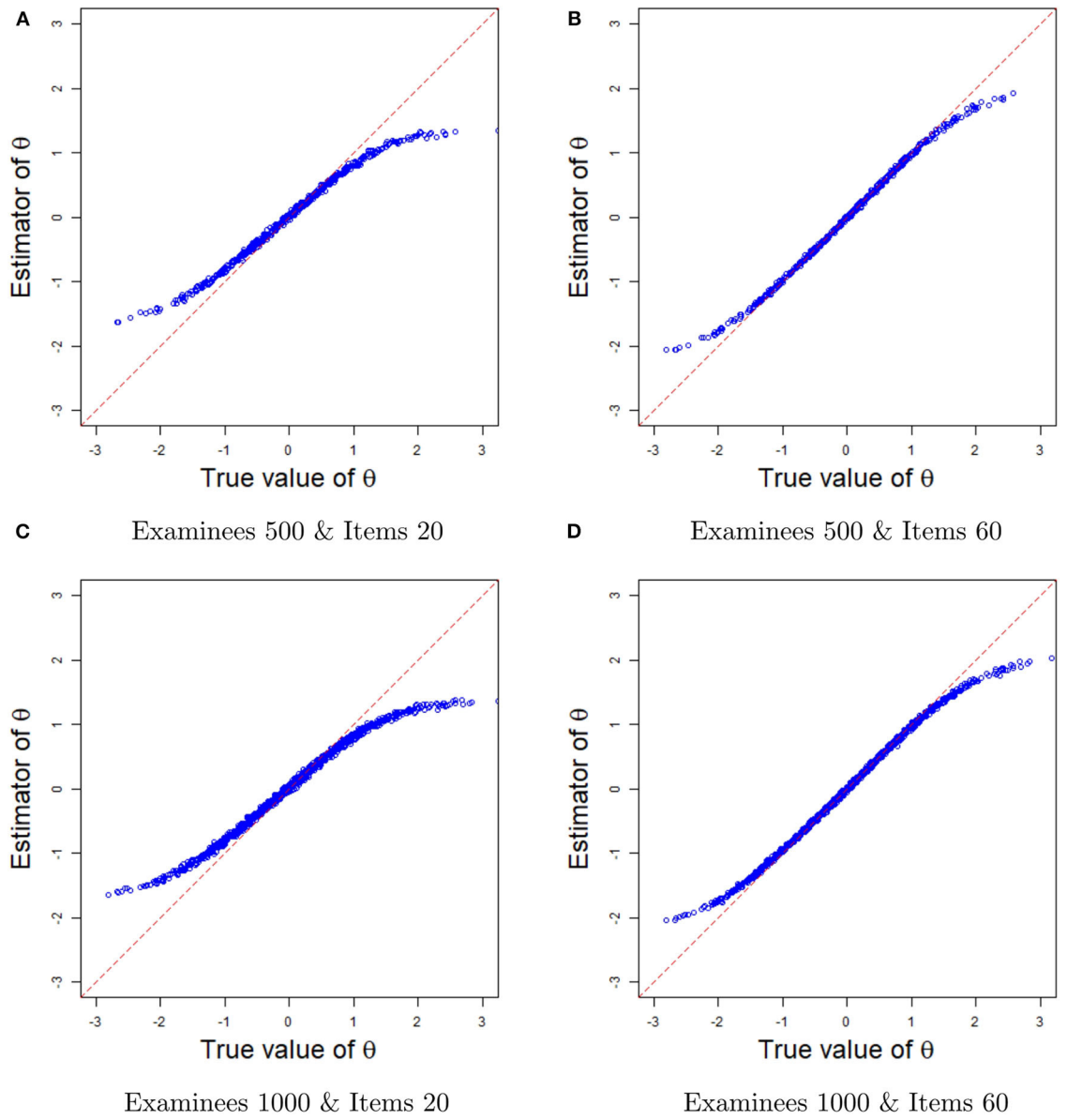

FIGURE 6 | The comparisons between ability estimates and true values in different sample sizes. **(A)** The comparisons between ability estimates and true values based on 500 examinees and 20 items. **(B)** The comparisons between ability estimates and true values based on 500 examinees and 60 items. **(C)** The comparisons between ability estimates and true values based on 1,000 examinees and 20 items. **(D)** The comparisons between ability estimates and true values based on 1,000 examinees and 60 items.

5. REAL DATA

5.1. Data Description

In this example, the 2015 computer-based Program for International Student Assessment (PISA) science data are used. From among the many countries that have participated in the computer-based assessment of the sciences, we choose the students from the USA as the object of analysis. Students with Not Reached (original code 6) or Not Response (original code 9) are removed in this study, where Not Reached and Not Response (omitted) are treated as missing data. The final 548 students are used to answer 16 items, and the corresponding response times are recorded. All 16 items are scored using a dichotomous scale. The 16 items are respectively CR083Q01S, CR083Q02S, CR083Q03S, CR083Q04S, DR442Q02C, DR442Q03C, DR442Q05C, DR442Q06C, CR442Q07S, CR245Q01S, CR245Q02S, CR101Q01S, CR101Q02S, CR101Q03S, CR101Q04S, and CR101Q05S. The frequency histogram of logarithmic response times and the correct rate for each item are shown in **Figure 7**.

5.2. Bayesian Model Assessment

To evaluate the impact of different time weights on the PISA data and to analyze the differences between the G3PLT model and

TABLE 4 | The results of Bayesian model assessment in Simulation 2.

	Fitted model			3PL	G3PLT0	G3PLT2	G3PLT4	G3PLT6	G3PLT8
True Model	G3PLT0	DIC	Q_1	25297.63	**25270.07**	–	–	–	–
			Median	25358.15	**25324.43**	–	–	–	–
			Q_3	25412.52	**25379.70**	–	–	–	–
			IQR	114.88	**109.63**	–	–	–	–
		LPML	Q_1	−13456.37	**−13251.64**	–	–	–	–
			Median	−13431.01	**−13231.77**	–	–	–	–
			Q_3	−13406.19	**−13218.86**	–	–	–	–
			IQR	50.17	**32.77**	–	–	–	–
	G3PLT2	DIC	Q_1	22742.44	–	**22677.65**	–	–	–
			Median	22851.46	–	**22777.38**	–	–	–
			Q_3	22953.34	–	**22890.79**	–	–	–
			IQR	210.89	–	**213.14**	–	–	–
		LPML	Q_1	−12274.46	–	**−12246.10**	–	–	–
			Median	−12243.68	–	**−12221.93**	–	–	–
			Q_3	−12221.43	–	**−12200.33**	–	–	–
			IQR	53.02	–	**45.76**	–	–	–
	G3PLT4	DIC	Q_1	20529.71	–	–	**20522.24**	–	–
			Median	20614.41	–	–	**20613.60**	–	–
			Q_3	20711.15	–	–	**20708.31**	–	–
			IQR	181.44	–	–	**186.06**	–	–
		LPML	Q_1	−11322.69	–	–	**−11263.87**	–	–
			Median	−11300.75	–	–	**−11239.84**	–	–
			Q_3	−11273.01	–	–	**−11219.60**	–	–
			IQR	49.67	–	–	**44.26**	–	–
	G3PLT6	DIC	Q_1	20210.35	–	–	–	**20206.43**	–
			Median	20295.34	–	–	–	**20294.27**	–
			Q_3	20386.09	–	–	–	**20384.67**	–
			IQR	175.73	–	–	–	**178.23**	–
		LPML	Q_1	−11102.84	–	–	–	**−11144.73**	–
			Median	−11079.08	–	–	–	**−11121.81**	–
			Q_3	−11052.10	–	–	–	**−11098.77**	–
			IQR	50.74	–	–	–	**45.96**	–
	G3PLT8	DIC	Q_1	20014.40	–	–	–	–	**20013.64**
			Median	20111.34	–	–	–	–	**20112.86**
			Q_3	20191.08	–	–	–	–	**20189.52**
			IQR	176.68	–	–	–	–	**175.87**
		LPML	Q_1	−11083.24	–	–	–	–	**−11032.39**
			Median	−11053.93	–	–	–	–	**−11007.48**
			Q_3	−11026.44	–	–	–	–	**−10981.35**
			IQR	56.79	–	–	–	–	**51.03**

Note that the 3PL denotes three parameter logistic model, and the G3PLTw denotes the general three parameter logistic model with time weight w, where w = 0, 2, 4, 6, 8.

the traditional 3PL model in fitting the data, both models are used to fit the data. G3PLT models with different time weights $W = 0, 1, 2, 3, 4, 5, 6, 7,$ and 8 are considered. In the estimation procedure, the setting of the prior distributions is the same as in Simulation 1. In all of the Bayesian computations, we use 10,000 MCMC samples after a burn-in of 5,000 iterations for each model to compute all posterior estimates.

Table 5 shows the results for DIC and LPML under the 3PL model and the G3PLT model with different time weights. According to DIC and LPML, we find that the G3PLT model with time weight 6 is the best-fitting model, with DIC and LPML values of 8389.316 and −4196.672, respectively. The G3PLT model with time weight 0 is the worst-fitting model, with DIC and LPML values of 9708.940 and −4792.301, respectively. That the G3PLT model with time weight 0 is the worst fitting model can be attributed to the fact that the influence of the time effect on the correct-response probability is relatively weak for the PISA data. This is consistent with the the evaluation purpose of the PISA test, which is a nonselective and low-stakes test. Examinees lack motivation to answer each item carefully, and therefore the

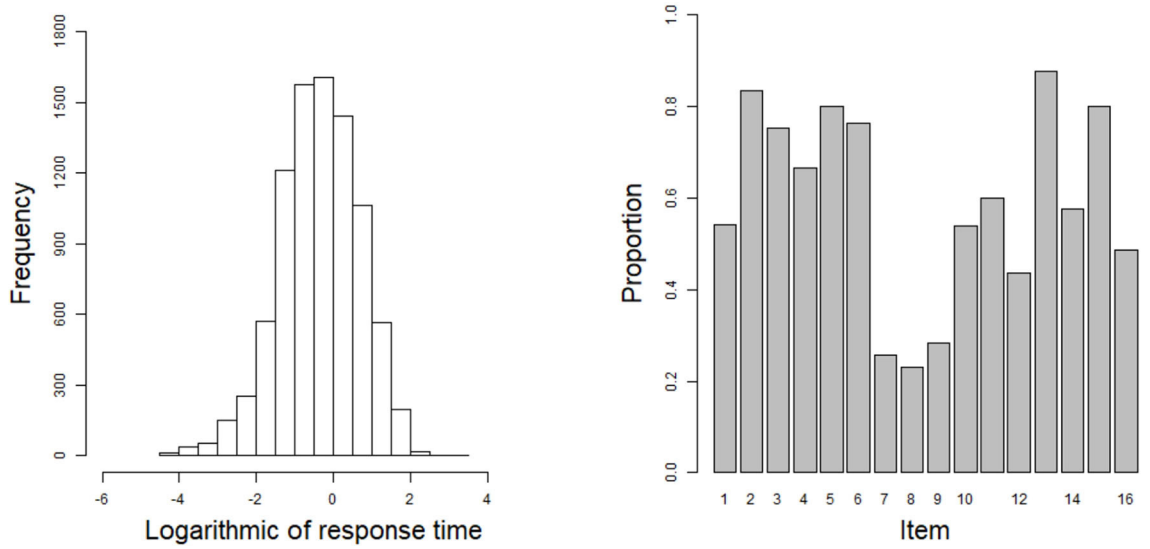

FIGURE 7 | The frequency histogram of logarithmic response times and the correct rate for each item in the real data.

time effect cannot be reflected. However, when the time weight of the G3PLT model increases from 5 to 8, the DIC and LPML values are basically the same as those in the case of the 3PL model. The model fitting results once again verify that our G3PLT model reduces to the traditional 3PL model when the time weight increases to a certain value. Next, we will analyze the PISA data based on the G3PLT model with time weight 6.

5.3. Analysis of Item Parameters

The estimated results for the item parameters are shown in **Table 6**. We can see that the expected a posteriori (EAP) estimates of the nine item discrimination parameters are greater than one. This indicates that these items can distinguish well between different abilities. In addition, the EAP estimates of the 11 difficulty parameters are less than zero, which indicates that 10 items are slightly easier than the other six. The three most difficult items are items 8 (DR442Q06C), 7 (DR442Q05C), and 9 (CR442Q07S). The EAP estimates of the difficulty parameters for these three items are, respectively 1.085, 0.900, and 0.839. The corresponding correct rates for the three items in **Figure 7** are 0.231, 0.257, and 0.285. The most difficult three items have the lowest correct rates, which is consistent with our intuition. The six EAP estimates of the guessing parameters are larger than 0.1. The three items that the examinees are most likely to answer correctly by guessing are items 11 (CR245Q02S), 12 (CR101Q01S), and 10 (CR245Q01S). The EAP estimates of the guessing parameters for these three items are respectively 0.132, 0.128, and 0.117. Among the 16 items, item 7 is the best design item owing to the fact that it has high discrimination and difficulty estimates, and the guessing parameter has the lowest estimate in all of the items. Next, we use the posterior standard deviation (SD) to evaluate the degree of deviation from the EAP estimate. The average SD of all discrimination parameters is about 0.005, the average SD of all difficulty parameters is about 0.010, and the average SD of all guessing parameters is about 0.001. We can see that the average SD values of the three parameters are very small, indicating that the estimated values fluctuate near the posterior mean.

5.4. Analysis of Personal Parameters

Next, we analyze the differences between the estimated abilities of examinees in the 3PL model and in the G3PLT model under the same response framework, together with the reasons for these differences. We consider four examinees with same response framework for the 16 items, (1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1). They are examinee 60, examinee 313, examinee 498, and examinee 210, and the corresponding response times for these examinees to answer the 16 items are 25.80, 29.36, 35.48, and 41.44 min. Under the framework of the 3PL model, the estimated abilities of the four examinees are the same, 1.45. However, taking into account the time factors for the four examinees, the estimated abilities are different according to the G3PLT model with time weight 6. The estimated abilities are 1.46, 1.42, 1.41, and 1.38, respectively. We find that under the same response framework, as the response times of the examinees increase from 25.80 to 41.44 min, the estimated abilities of the examinees show a decreasing trend. This indicates that examinees with short response times are more proficient in answering these items than examinees with long response times. Therefore, the ability of examinees with short response times to answer 15 items correctly should be higher than that of examinees with long times. This once again shows that our G3PLT model is reasonable. By incorporating the time effect into the IRT model, the interpretation of the latent construct essentially shifts: before we were measuring whether students could answer items correctly, now we are measuring whether students can answer items correctly and quickly.

TABLE 5 | The results of Bayesian model assessment in real data analysis.

	3PL	G3PLT0	G3PLT1	G3PLT2	G3PLT3
DIC	8392.374	9708.940	9217.431	8825.986	8561.295
LPML	−4197.832	−4792.301	−4565.769	−4395.082	−4275.351
	G3PLT4	**G3PLT5**	**G3PLT6**	**G3PLT7**	**G3PLT8**
DIC	8441.556	8398.835	8389.316	8391.581	8390.254
LPML	−4221.003	−4200.857	−4196.672	−4197.678	−4197.906

TABLE 6 | The results of item parameter estimation in real data analysis.

Parameter	EAP	SD	HPDI	Parameter	EAP	SD	HPDI
a_1	0.980	0.003	[0.873, 1.120]	a_9	1.199	0.003	[1.116, 1.312]
a_2	0.927	0.003	[0.824, 1.025]	a_{10}	0.821	0.004	[0.688, 0.946]
a_3	0.986	0.004	[0.857, 1.114]	a_{11}	1.059	0.006	[0.890, 1.200]
a_4	1.034	0.003	[0.928, 1.139]	a_{12}	1.004	0.007	[0.874, 1.195]
a_5	0.893	0.007	[0.723, 1.047]	a_{13}	1.037	0.006	[0.899, 1.198]
a_6	1.084	0.005	[0.965, 1.211]	a_{14}	1.011	0.005	[0.883, 1.137]
a_7	1.216	0.005	[1.062, 1.336]	a_{15}	0.986	0.006	[0.848, 1.190]
a_8	1.087	0.004	[0.974, 1.203]	a_{16}	0.803	0.002	[0.715, 0.917]
b_1	−0.065	0.009	[−0.240, 0.111]	b_9	0.839	0.007	[0.670, 0.995]
b_2	−1.405	0.014	[−1.617, −1.170]	b_{10}	0.065	0.020	[−0.186, 0.391]
b_3	−0.921	0.010	[−1.085, −0.693]	b_{11}	−0.147	0.016	[−0.374, 0.114]
b_4	−0.519	0.009	[−0.700, −0.321]	b_{12}	0.530	0.014	[0.324, 0.795]
b_5	−1.187	0.021	[−1.430, −0.849]	b_{13}	−1.608	0.015	[−1.846, −1.369]
b_6	−0.920	0.011	[−1.124, −0.730]	b_{14}	−0.083	0.012	[−0.280, 0.149]
b_7	0.900	0.007	[0.726, 1.069]	b_{15}	−1.145	0.016	[−1.429, −0.933]
b_8	1.085	0.007	[0.876, 1.236]	b_{16}	0.272	0.016	[0.062, 0.547]
c_1	0.065	0.000	[0.018, 0.120]	c_9	0.042	0.000	[0.016, 0.069]
c_2	0.098	0.001	[0.026, 0.189]	c_{10}	0.117	0.001	[0.029, 0.192]
c_3	0.079	0.001	[0.017, 0.156]	c_{11}	0.132	0.001	[0.056, 0.216]
c_4	0.079	0.001	[0.015, 0.143]	c_{12}	0.128	0.001	[0.071, 0.190]
c_5	0.107	0.002	[0.028, 0.199]	c_{13}	0.093	0.001	[0.028, 0.176]
c_6	0.092	0.001	[0.019, 0.158]	c_{14}	0.115	0.001	[0.034, 0.177]
c_7	0.026	0.000	[0.006, 0.045]	c_{15}	0.097	0.002	[0.022, 0.185]
c_8	0.032	0.000	[0.009, 0.056]	c_{16}	0.103	0.001	[0.035, 0.165]

6. CONCLUDING REMARKS

In this paper, we propose a new and flexible general three-parameter logistic model with response time (G3PLT), which is different from previous response time models, such as the hierarchical model framework proposed by van der Linden (2007), in which the response and the response time are considered in different measurement models, while a high-level model represents the correlation between latent ability and speed through a population distribution. However, our model integrates latent ability, time, and item difficulty into a item response model to comprehensively consider the impact on the probability of correct response. This approach to modeling is simpler and more intuitive. In addition, time weights are introduced in our model to investigate the influence of time intensity limited by different tests on the correct-response probability. When the time weight reaches 8, our model reduces to the traditional 3PL model, which indicates that the time factor has little influence on the correct-response probability. The examinees then answer each item correctly with the response probability given by the 3PL model.

However, the computational burden of the Bayesian algorithm becomes excessive when large numbers of examinees or items are considered or a large MCMC sample size is used. Therefore, it is desirable to develop a standalone R package associated with C++ or Fortran software for more extensive large-scale assessment programs.

Other issues should be investigated in the future. First of these is whether the G3PLT model can be combined with a

multilevel structure model to analyze the influence of covariates on the latent ability at different levels, for example, to explore the influence of the time effect, gender, and socioeconomic status on latent ability. Second, although we have found that for different examinees with the same response framework, the ability estimates from the 3PL model is the same, those from the G3PLT model differ greatly. Examinees who take less time should be more proficient in answering items, and their ability should be higher than that of examinees who take longer. "Proficiency" is a latent skill that is not the same as latent ability. Whether we can connect proficiency and latent ability through a multidimensional 3PLT model to analyze their relationship is also an important topic for our future research. Third, our new model can also be used to detect various abnormal response behaviors, such as rapid guessing and cheating, with the aim of eliminating deviations in ability estimates caused by such behaviors.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.oecd.org/pisa/data/.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant no. 11571069) and Foundation for basic and applied basic research of Guangdong Province (Grant no. 2019A1515110448).

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibb sampling. *J. Educ. Stat.* 17, 251–269. doi: 10.3102/10769986017003251

Baker, F. B., and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques, 2nd Edn.* New York, NY: Marcel Dekker.

Béguin, A. A., and Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* 66, 541–561. doi: 10.1007/BF02296195

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA:Addison-Wesley), 397–479.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Bolsinova, M., and Tijmstra, J. (2018). Improving precision of ability estimation: getting more from response times. *Br. J. Math. Stat. Psychol.* 71, 13–38. doi: 10.1111/bmsp.12104

Bridgeman, B., and Cline, F. (2004). Effects of differentially time-consuming tests on computer adaptive test scores. *J. Educ. Measure.* 41, 137–148. doi: 10.1111/j.1745-3984.2004.tb01111.x

Brooks, S., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Chang, H. (2004). "Computerized testing, E-rater, and generic algorithm: Psychometrics to support emerging technologies," in *Invited Symposium, 28th International Congress of Psychology* (Beijing).

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York, NY: Springer.

Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *Am. Stat.* 49, 327–335. doi: 10.1080/00031305.1995.10476177

Choe, E. M., Kern, J. L., and Chang, H-H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *J. Educ. Behav. Stat.* 43, 135–158. doi: 10.3102/10769986177 23642

De Boeck, P., and Jeon, M. (2019). An overview of models for response times and processes in cognitive tests? *Front. Psychol.* 10:102. doi: 10.3389/fpsyg.2019.00102

Embretson, S. E, and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Earlbaum Associates.

Fox, J.-P., Klein Entink, R. H., and van der Linden, W. J. (2007). Modeling of responses and response times with the package CIRT. *J. Stat. Softw.* 20, 1–14. doi: 10.18637/jss.v020.i07

Fox, J.-P., and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivar. Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection. *J. Am. Stat. Assoc.* 74, 153–160. doi: 10.1080/01621459.1979.10481632

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling based methods (with discussion)," in *Bayesian Statistics*, Vol. 4, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford, UK: Oxford University Press), 147–167.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Han, K. T. (2012). *Fixing the c Parameter in the Three-Parameter Logistic Model.* Practical Assessment, Research and Evaluation, 17. Available online at: http://pareonline.net/getvn.asp?v=17andn=1

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109. doi: 10.1093/biomet/57.1.97

Hung, L.-F., and Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *J. Educ. Behav. Stat.* 37, 231–255. doi: 10.3102/1076998611402503

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York, NY: Springer.

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., and Fox, J.-P. (2009b). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychol. Methods* 14, 54–75. doi: 10.1037/a0014877

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: AddisonWesley.

Lu, J., Wang, C., Zhang, J., W. and Tao, J. (2019). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *Br. J. Math. Stat. Psychol.* doi: 10.1111/bmsp.12175

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., and Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *J. Educ. Behav. Stat.* 39, 426–451. doi: 10.3102/1076998614559412

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Patz, R. J., and Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.* 24, 146–178. doi: 10.3102/10769986024002146

Patz, R. J., and Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *J. Educ. Behav. Stat.* 24, 342–366. doi: 10.3102/10769986024004342

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *J. Educ. Behav. Stat.* 41, 300–325. doi: 10.3102/1076998616636618

Qian, H., Staniewska, D., Reckase, M., and Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educ. Measure.* 35, 38–47. doi: 10.1111/emip.12102

Roskam, E.E. (1987). "Toward a psychometric theory of intelligence," in *Progress in Mathematical Psychology*, eds E. E. Roskam and R. Suck (Amsterdam: North-Holland), 151–174.

Roskam, E. E. (1997). "Models for speed and time-limit tests," in *Handbook of Modern Item Response Theory*, eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 187–208.

Rupp, A. A., Dey, D. K., and Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling. *Struct. Equat. Model.* 11, 424–451. doi: 10.1207/s15328007sem1103_7

Schnipke, D. L., and Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *J. Educ. Measure.* 34, 213–232. doi: 10.1111/j.1745-3984.1997.tb00516.x

Sinharay, S., and Johnson, M. S. (2019). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *Br. J. Math. Stat. Psychol.* doi: 10.1111/bmsp.12187. [Epub ahead of print].

Skorupski, W. P., and Wainer, H. (2017). "The case for Bayesian methods when investigating test fraud," in *Handbook of Detecting Cheating on Tests*, eds G. J. Cizek and J. A. Wollack (London: Routledge), 214–231.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and vander Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Swaminathan, H., and Gifford, J. A. (1979). *Estimation of Parameters in the Three-Parameter Latent Trait Model* (Report No. 90). Amherst: Laboratory of Psychometric and Evaluation Research; School of Education; University of Massachusetts.

Thissen, D. (1983). "Timed teting: An approach using item response theory," in *New Horizons in Testing*, ed D. J. Weiss (New York, NY: Academic Press), 179–203.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.* 22, 1701–1762. doi: 10.1214/aos/1176325750

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 297–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *J. Educ. Behav. Stat.* 34, 378–394. doi: 10.3102/1076998609332107

van der Linden, W. J., Breithaupt, K., Chuah, S. C., and Zhang, Y. (2007). Detecting differential speededness in multistage testing. *J. Educ. Measure.* 44, 117–130. doi: 10.1111/j.1745-3984.2007.00030.x

van der Linden, W. J., and Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika* 75, 120–139. doi: 10.1007/s11336-009-9129-9

van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

van der Linden, W. J., and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

Verhelst, N. D., Verstralen, H. H. F. M., and Jansen, M. G. (1997). "A logistic model for time-limit tests," in *Handbook of Modern Item Response Theory*, eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 169–185.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement* 7, 110–114. doi: 10.1080/15366360903117079

Wang, C., Fan, Z., Chang, H.-H., and Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J. Educ. Behav. Stat.* 38, 381–417. doi: 10.3102/1076998612461831

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Xu, G., and Shang, Z. (2018a). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Wang, C., Xu, G., Shang, Z., and Kuncel, N. (2018b). Detecting aberrant behavior and item preknowledge: a comparison of mixture modeling method and residual method. *J. Educ. Behav. Stat.* 43, 469–501. doi: 10.3102/1076998618767123

Wang, T., and Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Appl. Psychol. Measure.* 29, 323–339. doi: 10.1177/0146621605275984

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., and Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* 31, 374–383. doi: 10.1177/0165025407077764

Zhu, Z. M., Wang, C., and Tao, J. (2018). A two-parameter logistic extension model: an efficient variant of the three-parameter logistic model. *Appl. Psychol. Measure.* 21, 1–15. doi: 10.1177/0146621618800273

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). *Educ. Psychol. Measure.* 79, 931–961. doi: 10.1177/0013164419839439

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.