ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fpsyg.2020.570365

Cognitive Diagnostic Models for Random Guessing Behaviors

Chia-Ling Hsu[1]*, *Kuan-Yu Jin*[2] *and Ming Ming Chiu*[1]

[1] Assessment Research Centre, The Education University of Hong Kong, Tai Po, Hong Kong, [2] Hong Kong Examinations and Assessment Authority, Wan Chai, Hong Kong

Many test-takers do not carefully answer every test question; instead they sometimes quickly answer without thoughtful consideration (*rapid guessing*, RG). Researchers have not modeled RG when assessing student learning with *cognitive diagnostic models* (CDMs) to personalize feedback on a set of fine-grained skills (or attributes). Therefore, this study proposes to enhance cognitive diagnosis by modeling RG via an advanced CDM with item response and response time. This study tests the parameter recovery of this new CDM with a series of simulations via Markov chain Monte Carlo methods in JAGS. Also, this study tests the degree to which the standard and proposed CDMs fit the student response data for the Programme for International Student Assessment (PISA) 2015 computer-based mathematics test. This new CDM outperformed the simpler CDM that ignored RG; the new CDM showed less bias and greater precision for both item and person estimates, and greater classification accuracy of test results. Meanwhile, the empirical study showed different levels of student RG across test items and confirmed the findings in the simulations.

Keywords: response time, rapid guessing, G-DINA model, DINA model, DINO model

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Kaiwen Man,
University of Maryland, College Park, United States
Gongjun Xu,
University of Michigan, United States

***Correspondence:**
Chia-Ling Hsu
clhsu@friends.eduhk.hk

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

Received: *07 June 2020*
Accepted: *07 September 2020*
Published: *25 September 2020*

Citation:
Hsu C-L, Jin K-Y and Chiu MM (2020) Cognitive Diagnostic Models for Random Guessing Behaviors. Front. Psychol. 11:570365. doi: 10.3389/fpsyg.2020.570365

INTRODUCTION

Cognitive diagnostic models (CDMs) assess whether test-takers have the skills needed to answer test questions (*attributes*), so that their test results can give them diagnostic feedback on their strengths and weaknesses in these attributes (Rupp et al., 2010). Specifically, a CDM analysis determines whether a person shows mastery (vs. non-mastery) of a set of fine-grained attributes (*latent class*). Teachers, clinicians and other users of test scores can use such specific information on each student or client to adapt and improve their instructions/interventions more effectively, compared to a simple, summative score.

However, some test-taking behaviors can distort current CDM results and thereby jeopardize the validity of their assessments. Recently, researchers have proposed different approaches to account for test-taking behaviors when assessing test-taker performance and item characteristics. In this study, we focus on two frequently-observed test-taking behaviors during actual tests: solution attempt and *rapid guessing* (RG; Wise and Kong, 2005). In a solution attempt, test-takers carefully try to find answers to test questions. By contrast, RG refers to test-takers quickly answering test questions without thoughtful consideration (e.g., Wise and DeMars, 2006). For instance, Meyer (2010) integrated a two-class mixture Rasch model (Rost, 1990) to classify a test-taker as either making a solution attempt or RG, but not allowing different behaviors by the same person during a test. To address this limitation, Wang and Xu (2015) proposed a model with a latent indicator to allow each test-taker to engage in either a solution attempt or RG on each item. Furthermore, the indicator can depend on either a test-taker's RG propensity or on an item-level feature (Wang et al., 2018). As RGs are typically much shorter than solution attempts, CDMs can use a test-taker's reaction time (RT) to each test question to properly model RGs and distinguish them from solution attempts (but not necessarily pre-knowledge answers, e.g., Wang et al., 2018). As no published study has proposed and tested a CDM that models RG, we do so in this study.

This study proposes a new framework of CDMs to recognize different test-taking behaviors by using RT and item responses simultaneously. This new class of CDMs: (1) models two test-taking behaviors (RG vs. solution attempt) for each item-person concurrence, (2) allows multiple switch points between RG and solution attempts among the items for each test-taker, (3) thereby yields person and item estimates with greater accuracy, and (4) generalizes to available CDMs, RT functions and other kinds of dissimilar behaviors.

The generalized DINA model (G-DINA, de la Torre, 2011) conceptualizes and shows the utility of this framework. Specifically, the two special cases of the G-DINA model, the deterministic input, noisy "and" gate (DINA) model (Junker and Sijtsma, 2001) and its counterpart, the deterministic input, noisy "or" gate (DINO) model (Templin and Henson, 2006) are simple to compute, estimate, and interpret, so they serve as illustrations. Nevertheless, researchers can extend this approach to other CDMs, especially G-DINA-liked formulation CDMs, such as the general diagnostic model (GDM; von Davier, 2005) and the linear logistic model (Maris, 1999).

After we present the functions for describing RT and item response, we specify the new model. Next, our simulation study illustrates the new model's performance, followed by its application to real data. Lastly, we discuss the implications of this study for identifying test-taking behaviors and improving the estimation accuracy of both person and item parameters.

A NEW CDM FRAMEWORK

The new model requires distinct functions to separately specify two fundamentals for an item, RT and item response, while two main facets, person and item, affect the observed RT and item response. This section describes the adopted RT and item response functions, before specifying the new model.

The Lognormal RT Model

As cognitive test data typically resemble a lognormal distribution more closely than a normal distribution, we use a lognormal function to characterize RT (van der Linden, 2006, 2007). Let RT_{ij} be the observed RT of person i (i = 1, 2, . . ., I) to item j (j = 1, 2, . . ., J). In the lognormal function, the two parameters of *person speed* and *time intensity,* respectively, represent the two facets, person and item, as follows,

$$log\left(RT_{ij}\right) \sim N(\beta_j - \tau_i,\, 1/\kappa_j^2) \quad (1)$$

where τ_i indicates the average speed of test-taker i on a test (person speed); β_j indicates the mean time that the population needs to resolve item j (time intensity); and κ_j^2 indicates the dispersion of the logarithmized RT distribution (time discrimination parameter) of item j.

The G-DINA Model

The G-DINA model loosens some restrictions of the DINA model and its saturated form is equivalent to other general CDMs via link functions (de la Torre, 2011). Hence, the G-DINA model can (a) present different CDMs with similar formulations via various constraints and (b) substantially reduce the number of latent classes for an item – especially for models with more than five attributes. The original G-DINA model with identity link can be expressed as

$$P\left(\alpha_{ij}^{*}\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik} \quad (2)$$

For test-taker i, the reduced attribute vector α_{ij}^{*} has the required attributes for item j. The intercept for item j, δ_{j0} represents the probability of a correct response without the required attributes (*baseline probability*). The main effect δ_{jk} reflects the extent to which mastery of a single attribute α_k changes the probability of a correct response. The interaction effect $\delta_{jkk'}$ indicates the extent to which mastery of both attributes α_k and $\alpha_{k'}$ changes the probability of a correct response. The interaction effect $\delta_{j12\cdots K_j^*}$ reflects the extent to which mastery of all the required attributes $\alpha_1, \alpha_2, \cdots$, and $\alpha_{K_j^*}$ changes the probability of a correct response.

Like most CDMs, the G-DINA model requires a $J \times K$ Q-matrix (Tatsuoka, 1983), in which K knowledge attributes are required to correctly answer J items. $K_j^* = \sum_{k=1}^{K} q_{jk}$ is the number of required attributes for item j, where $q_{jk} = 1$ if the correct response to item j requires attribute k; and 0 otherwise. As the number of required attributes for item j is smaller than that of the all attribute vectors ($K_j^* < K$), the G-DINA model can reduce the number of required latent classes ($2^{K_j^*} < 2^K$) for an item. To illustrate the G-DINA-like formulations, we use two common cases: the DINA and DINO.

The DINA Model

In the non-compensatory DINA, individuals are classified into one of two latent classes for an item: (a) the attribute vectors have all of an item's required attributes (*mastery*) or (b) the attribute vectors are missing at least one of the item's required attributes (*non-mastery*). The two latent classes' corresponding probabilities for a correct response entail that (a) mastery individuals do not slip, or (b) non-mastery individuals guess correctly (Junker and Sijtsma, 2001). Thus, the DINA model can be re-formed by setting to zero, all G-DINA model parameters except δ_{j0} and $\delta_{j12\cdots K_j^*}$

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik} \tag{3}$$

In Eq. 3, $\delta_{j0} = g_j$ is the probability of a correct response to item j for a non-mastery test-taker i, where g_j is the guessing parameter for item j; $\delta_{j0} + \delta_{j12\cdots K_j^*} = 1 - s_j$ is the probability of a correct response to item j for a mastery test-taker i, where s_j is the slipping parameter for item j. In the DINA, a mastery test-taker with all the required attributes (K_j^*) for item j generally answers it correctly and other test-takers generally answer it incorrectly. Like the DINA, Eq. 3 shows that except for the attribute vector $\alpha_j^* = 1_{K_j^*}$ (in which $1_{K_j^*}$ is a vector of ones with length K_j^*), other latent classes ($2^{K_j^*} - 1$) have the same probability of correctly answering item j. As shown in Eq. 3, this probability increases only after mastering all the required attributes. Under the DINA model assumption (Junker and Sijtsma, 2001), the G-DINA has two parameters per item (see Eq. 3).

The DINO Model

Unlike the non-compensatory DINA, the compensatory DINO only entails at least one of the required attributes to answer an item, so the parameters in G-DINA are set to

$$\delta_{jk} = (-1)\, \delta_{jk'k''} = \cdots = (-1)^{K_j^*+1}\, \delta_{j1,2,\cdots,K_j^*} \tag{4}$$

where $k = 1, \cdots, K_j^*$, $k' = 1, 2, \cdots K_j^* - 1$, and $k'' > k', \cdots, K_j^*$. The orders of the interactions vary the alternating sign, and the quantities of the main effects and interactions have the same value.

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \delta_{jk}\alpha_{ik}. \tag{5}$$

For a test-taker i with at least one of the required attributes, the probability of answering item j without slipping (s_j') is $\delta_{j0} + \delta_{jk} = 1 - s_j'$. Likewise, for a test-taker i with none of the required attributes, the probability of correctly answering item j is the guessing parameter, $\delta_{j0} = g_j'$. Unlike the DINA, all latent classes except for the attribute vector $\alpha_j^* = 0_{K_j^*}$ (a vector of zeros and of length K_j^*) have the same probability of correctly answering item j. Like the DINA, the DINO only needs two parameters for an item (Eq. 5, Templin and Henson, 2006).

To use both information of RT and item response, two functions must be specified. Hence, RT-GDINA, RT-DINA and RT-DINO jointly model RT and item response with the lognormal distribution (Eq. 1) and G-DINA, DINA and DINO, respectively.

New Class of CDMs

We introduce a new class of G-DINA to account for varying test-taking behaviors. RT (RT_{ij}) and item response (Y_{ij}) are modeled individually. As test-takers can switch between RG and solution behaviors, like Wang and Xu (2015), a latent indicator (ξ) is employed, where if test-taker i tries to solve item j, $\xi_{ij} = 1$ (0 otherwise; RG is specified in this study). Incorporating this latent indicator into the lognormal RT model extends Eqs 1–6

$$\begin{cases} log\left(RT_{ij}\right) \sim N\left(\beta_j - \tau_i, 1/\kappa_j^2\right), & \text{if } \xi_{ij} = 1; \\ log\left(RT_{ij}\right) \sim N\left(\beta_0, 1/\kappa_0^2\right), & \text{if } \xi_{ij} = 0. \end{cases} \tag{6}$$

indicates that the logarithmized RT is normally distributed as Eq. 1 if test-taker i solves item j from solution attempt ($\xi_{ij} = 1$), and it is normally distributed with mean time intensity β_0 and time discrimination κ_0^2 if test-taker i responds to item j with a RG ($\xi_{ij} = 0$). For a RG on item j by test-taker i, RT is constant.

Likewise, adding ξ_{ij} to the G-DINA yields

$$P\left(\alpha_{ij}^*\right) = \xi_{ij}\left(\delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}\right) + \left(1 - \xi_{ij}\right)\delta_j^* \tag{7}$$

The G-DINA model is the underlying model for a solution attempt on item j by test-taker i ($\xi_{ij} = 1$). We assume that a RG on item j by test-taker i ($\xi_{ij} = 0$) yields δ_j^*. For simplicity, like Wise and DeMars (2006), we assume that test-taker i has the same probability of correctly answering item j both by RG and by guessing with none of the required attributes ($\delta_j^* = \delta_{j0}$), that is, guessing randomly for all options. Hence, Eq. 7 can be re-written as

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \xi_{ij}\left(\sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}\right) \tag{8}$$

The latent indicator ξ_{ij} in Eqs 6–8 is a binary result of test-taker i on item j's behavior (solution attempt vs. RG). It can be modeled by a Bernoulli distribution with π_j, the marginal probability of the solution attempt. Using the DINA and DINO to identify RGs and solution attempts, Eqs 3 and 5 are re-written, respectively, as Eqs 8 and 9.

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \xi_{ij}\delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik} \tag{9}$$

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \xi_{ij}\delta_{jk}\alpha_{ik} \tag{10}$$

If test-taker i tries to solve item j ($\xi_{ij} = 1$), Eqs 6, 8–10 reduce, respectively, to Eqs 1–3, and 5. Thus, the lognormal, G-DINA, DINA, and DINO are special cases of our proposed new CDM framework. Likewise, jointly modeling RT and item response with a latent indicator via Eqs 6 and 8–10 are, respectively, represented as RT-GDINA-RG, RT-DINA-RG and RT-DINO-RG.

To illustrate this approach, we combine the lognormal RT distribution and the easy-to-understand DINA and DINO models, with the latent indicator ξ (RT-DINA-RG, RT-DINO-RG) and without it (RT-DINA, RT-DINO). We estimate their parameters via the Bayesian method with the Markov chain Monte Carlo (MCMC) algorithm in the freeware JAGS (Plummer, 2017). For the JAGS code and the priors for the estimated parameters of the RT-DINA-RG and RT-DINO-RG models (see **Appendix**).

SIMULATION STUDY 1: PARAMETER RECOVERY OF RT-DINA-RG

Design

In simulation study 1, we evaluated the parameter recovery of the RT-DINA-RG for a test of 30 dichotomous items measuring five non-compensatory attributes. See the artificial Q-matrix in **Table 1**. The guessing (g_j) and slipping (s_j) parameters were randomly generated, respectively, from the uniform distributions of $U(0.05,\ 0.3)$ and $U(0.05,\ 0.2)$, which reflect a high quality test. This data-generating procedure for the 30 simulated items yielded item discrimination indices (IDI) that ranged from 0.51 to 0.88, indicating a test with high measurement quality (Lee et al., 2012).

TABLE 1 | Specified Q-matrix and item parameters in simulation 1.

Item	q_1	q_2	q_3	q_4	q_5	π_j
1	1	0	0	0	0	0.9
2	0	1	0	0	0	0.9
3	0	0	1	0	0	0.9
4	0	0	0	1	0	0.9
5	0	0	0	0	1	0.9
6	1	0	0	0	0	0.8
7	0	1	0	0	0	0.8
8	0	0	1	0	0	0.8
9	0	0	0	1	0	0.8
10	0	0	0	0	1	0.8
11	1	1	0	0	0	0.9
12	1	0	1	0	0	0.9
13	1	0	0	1	0	0.9
14	1	0	0	0	1	0.9
15	0	1	1	0	0	0.9
16	0	1	0	1	0	0.8
17	0	1	0	0	1	0.8
18	0	0	1	1	0	0.8
19	0	0	1	0	1	0.8
20	0	0	0	1	1	0.8
21	1	1	1	0	0	0.9
22	1	1	0	1	0	0.9
23	1	1	0	0	1	0.9
24	1	0	1	1	0	0.9
25	1	0	1	0	1	0.9
26	1	0	0	1	1	0.8
27	0	1	1	1	0	0.8
28	0	1	1	0	1	0.8
29	0	1	0	1	1	0.8
30	0	0	1	1	1	0.8

We manipulated two conditions. In the RG condition, the marginal probability of RG ($1 - \pi_j$) was set for items at two levels: 0.1 and 0.2 (Wang et al., 2018). To describe the dynamic latent indicator of person i on item j in the RG condition, the ξ-parameter was generated from a Bernoulli distribution with probability either of 0.8 or 0.9. In the RT-DINA-RG, mean item time intensity (β_0) and item discrimination (κ_0) were (a) set, respectively, at 2 and 1.6 for rapid guessers ($\xi_{ij} = 0$) and (b) generated, respectively, from $U(2,\ 4)$ and $U(0.15,\ 2)$ for normal test-takers ($\xi_{ij} =\ 1$). In non-RG condition, RG never occurs, and the RT-DINA served as the data-generating model, yielding parameters similar to the RT-DINA-RG. Mean item time intensity and item discrimination can be generated to accommodate various test situations (e.g., Man et al., 2018), but they do not affect the use of the proposed model. Therefore, we leave this interesting topic for further study.

We simulated 1,000 test-takers across conditions, and each test-taker had generated five latent attributes with positive correlations, following Henson and Douglas (2005) procedure. Specifically, we randomly generated 1,000 vectors with five values, $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \alpha_{i4}, \alpha_{i5})'$, $i = 1, 2, \ldots, 1{,}000$, from a multivariate normal distribution with no interaction, $MVN(\mathbf{0.5}, \Sigma)$ with Σ diagonal elements of 1.0 and others of 0.5. A cut-off value of 0.253 ($z_{0.6}$) indicated mastery of the attribute (if $\alpha_{ik} > 0.253$, $\alpha_{ik} = 1$; otherwise, $\alpha_{ik} = 0$), yielding ∼60% mean mastery of each attribute, which generally ranged from easy to moderate. The person speed parameter (τ_i) was generated from $N(0,\ 0.3^2)$. Each condition was replicated 100 times from an R script.

Both the RT-DINA and RT-DINA-RG were fit to these data to test three hypotheses: (1) with some RG, the RT-DINA-RG efficiently recovers item and person estimates; (2) ignoring RG via the RT-DINA yields biased item parameter estimates, less accurate classification of attribute mastery, and less reliable person speed estimates; and (3) with no RG, the RT-DINA-RG performs as well as the RT-DINA. To evaluate the recovery of item parameters, the bias and root mean squared error (RMSE) were computed as dependent variables:

$$\text{Bias}\left(\hat{\nu}\right) = \sum_{r=1}^{100} \left(\hat{\nu}_r - \nu\right)/100 \tag{11}$$

$$\text{RMSE}\left(\hat{\nu}\right) = \sqrt{\sum_{r=1}^{100} \left(\hat{\nu}_r - \nu\right)^2/100} \tag{12}$$

where ν and $\hat{\nu}_r$ indicate respectively, true and estimated values in the r-th replication of an item parameter. We examined test-takers' true and estimated latent classes to evaluate the classification accuracy of each attribute. The reliability of the person speed parameter was computed as:

$$\text{Reliability}\left(\hat{\tau}\right) = \text{Correlation}\left(\hat{\tau}, \tau\right)^2 \tag{13}$$

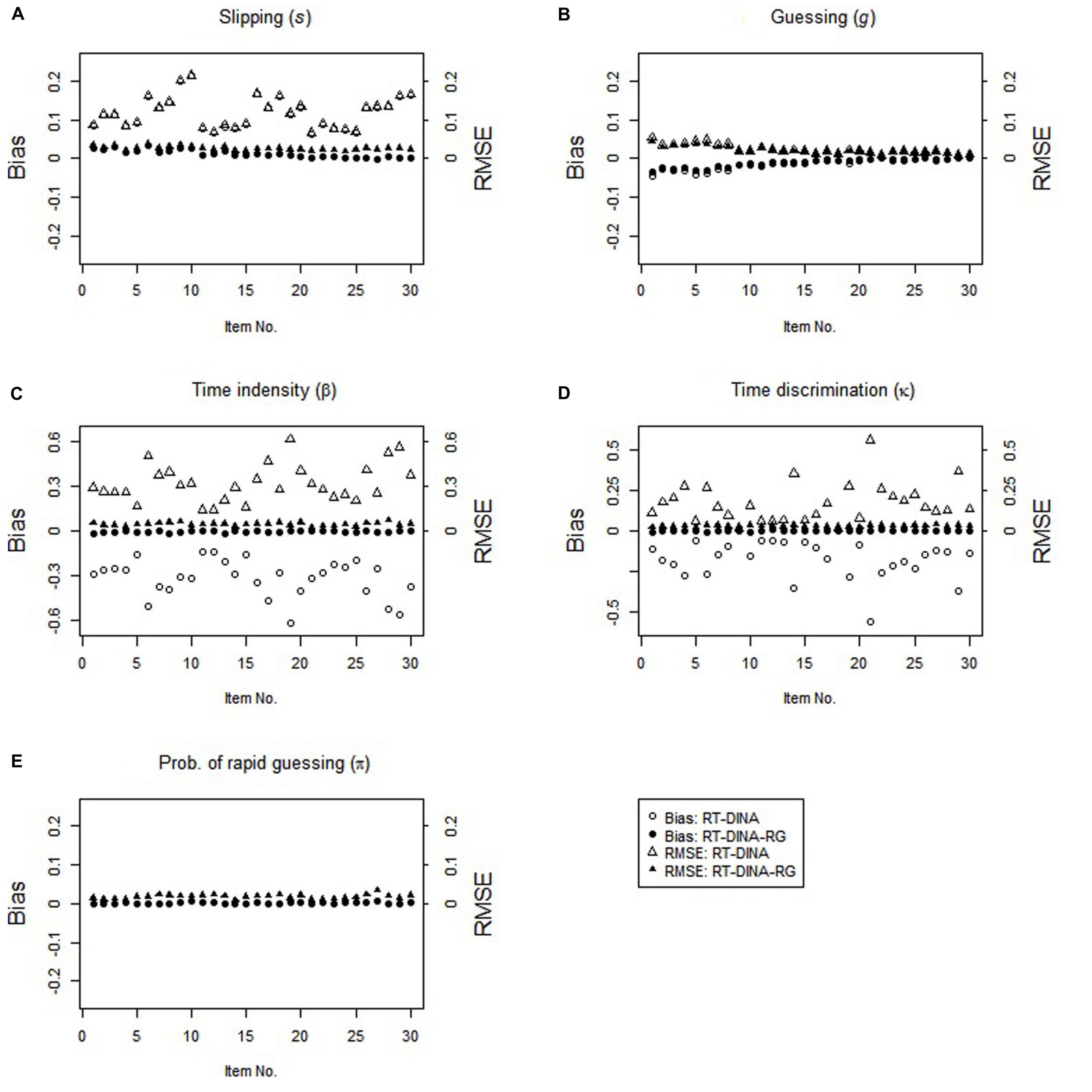

FIGURE 1 | Parameter recovery for the RT-DINA and RT-DINA-RG under the RG condition in simulation 1.

Results

In the RG condition, the RT-DINA-RG generally yielded unbiased parameter estimates, whereas the RT-DINA overestimated the slipping parameters and underestimated the item intensity, guessing, and time discrimination parameters (see **Figure 1**). Greater RG increased the severities of slipping overestimation and time intensity underestimation. For test-takers without the required attributes, RG did not influence the success rate, so ignoring RG did not substantially influence estimation of the guessing parameters. Across five attributes and 100 replications, mean classification accuracy was higher for the RT-DINA-RG than the RT-DINA (0.936 > 0.924), suggesting that ignoring RG reduces the accuracy of attribute classification. Also, the RT-DINA-RG outperformed the RT-DINA on reliability of the person speed parameter (M: 0.66 > 0.57).

In the non-RG condition, both RT-DINA and RT-DINA-RG recovered the parameters well (see **Figure 2**). The bias and RMSE for the π-parameter in the RT-DINA-RG were nearly zero. Also, both models yielded practically identical classification accuracy (M = 96.6%) and reliability of person speed parameter

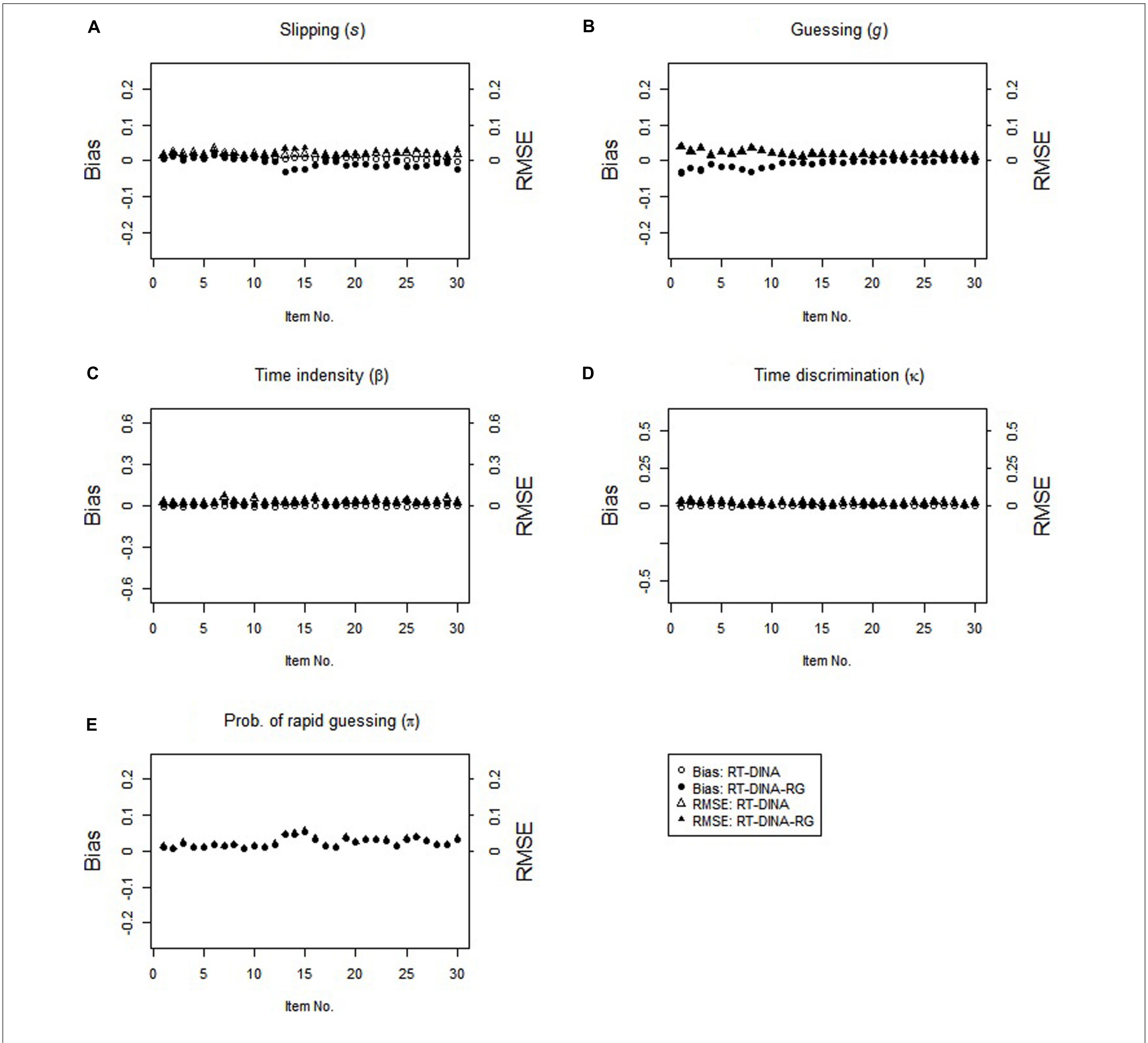

FIGURE 2 | Parameter recovery for the RT-DINA and RT-DINA-RG under the non-RG condition in simulation 1.

(M = 0.76) across 100 replications. Hence, overfitting the RT-DINA-RG to data without RG showed no significant harm. In brief, the simulation results supported our three hypotheses.

SIMULATION STUDY 2: PARAMETER RECOVERY OF RT-DINO-RG

Design

Study 2 simulated compensatory attributes and analyzed parameter recovery by the RT-DINO and RT-DINO-RG. The item responses and RTs were generated for (a) the RG condition with the RT-DINO-RG and (b) the non-RG condition with the RT-DINO. The parameters, data generation and evaluation criteria were the same as those in simulation study 1. Paralleling study 1, we test three hypotheses: (1) with some RG, the RT-DINO-RG efficiently recovers item and person estimates; (2) ignoring RG via the RT-DINO yields biased item parameter estimates and less accurate classification of attribute mastery; and (3) with no RG, the RT-DINO-RG performs as well as the RT-DINO.

Results

The study 2 results resemble the study 1 results (see **Figure 3**). In the RG condition, the RT-DINO-RG recovered the parameters well, whereas the RT-DINO overestimated the slipping parameters and underestimated the item intensity, guessing, and time discrimination parameters. Greater RG

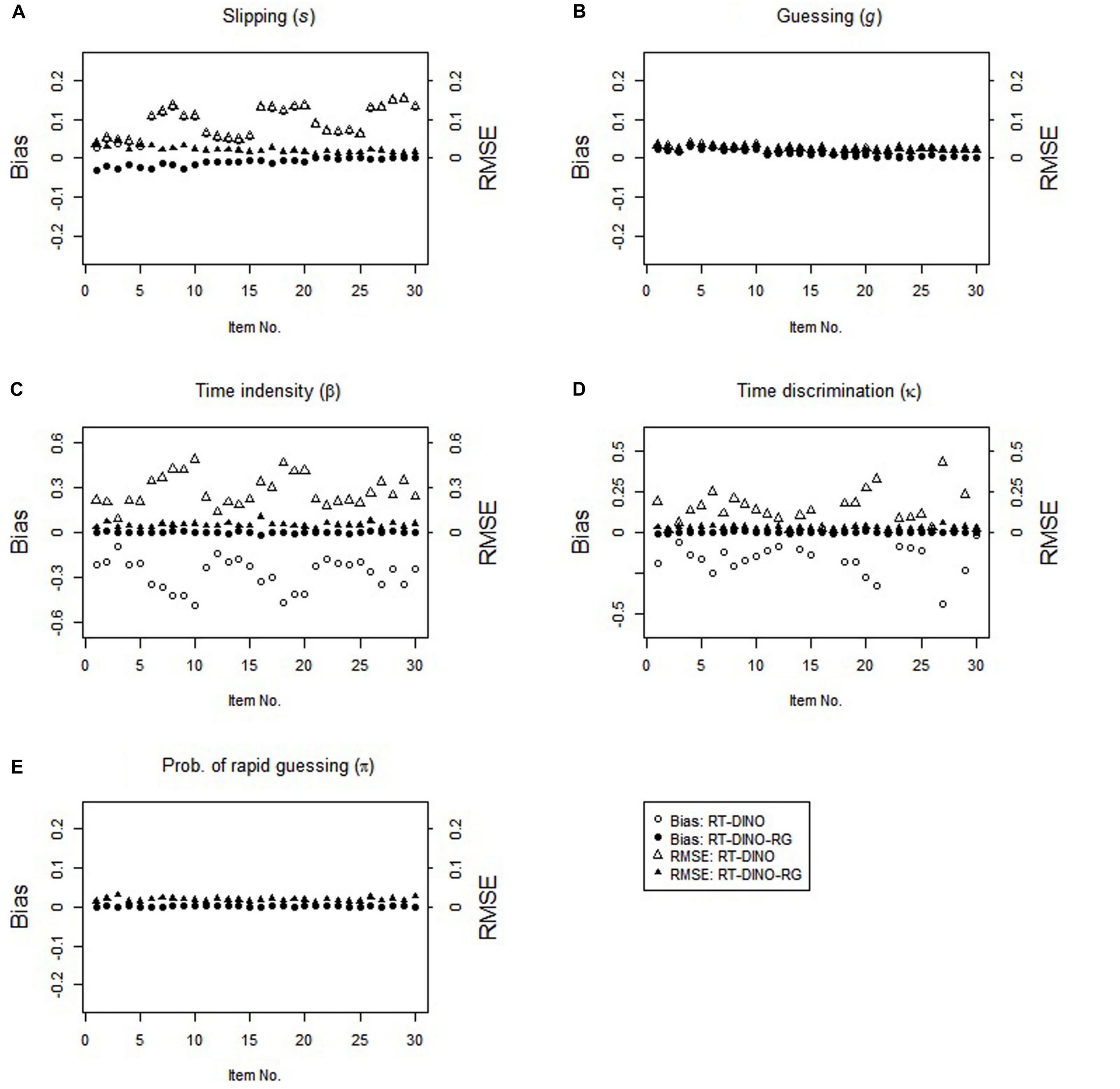

FIGURE 3 | Parameter recovery for the RT-DINO and RT-DINO-RG under the RG condition in simulation 2.

increased the severities of slipping overestimation and time intensity underestimation. The RT-DINO-RG outperformed the RT-DINO on both mean classification accuracy (0.946 > 0.922) across five attributes and reliability of the person speed parameter (*M*: 0.64 > 0.57).

In the non-RG condition, both RT-DINO and RT-DINO-RG recovered the item parameters well (see **Figure 4**). The bias and RMSE for π-parameter in the RT-DINO-RG model were very small. Also, both models had practically identical classification accuracy (*M* = 98.4%) and reliability of person speed parameter (*M* = 0.71) across replications. In sum, these simulation results supported our three hypotheses.

REAL DATA ANALYSIS

To illustrate a RT-GDINA-RG application, we analyzed a PISA 2015 mathematics test with 22 questions. After screening out students with missing responses, we analyzed 5,158 students' responses. The PISA 2015 mathematics assessment framework (OECD, 2017a,b) and the released computer-based mathematics items' log-file databases covered eight attributes: change and relationships (α_1), quantity (α_2), space and shape (α_3), uncertainty (α_4), occupational (α_5), societal (α_6), scientific (α_7), and personal (α_8). The Q-matrix for the mathematics test shows two cognitive attributes for each item (see **Table 2**). We fit

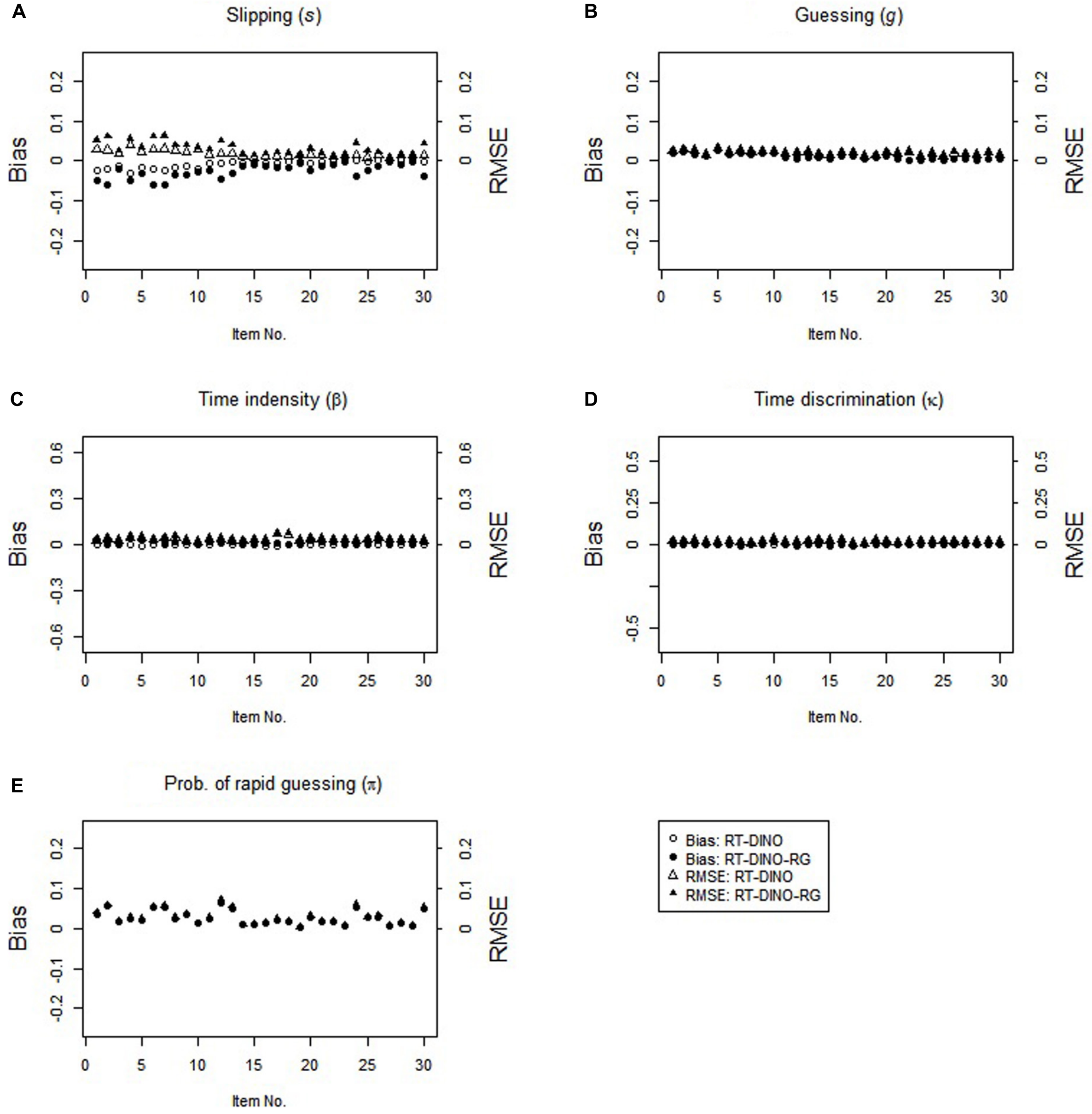

FIGURE 4 | Parameter recovery for the RT-DINO and RT-DINO-RG under the non-RG condition in simulation 2.

the four CDM models (RT-DINA, RT-DINO, RT-DINA-RG, RT-DINO-RG) to these data. Superior models have lower deviance information criteria (DIC; Spiegelhalter et al., 2002).

The results indicate both compensatory attributes and RG. DICs showed that the compensatory models outperformed the non-compensatory ones (RT-DINO < RT-DINA: 1,351,697 < 1,433,173; and RT-DINO-RG < RT-DINA-RG: 1,327,068 < 1,360,978) suggesting that the eight attributes' relationships were more compensatory than non-compensatory. Also, the RG models outperformed the simpler models (RT-DINO-RG < RT-DINO: 1,327,068 < 1,351,697; and RT-DINA-RG < RT-DINA: 1,360,978 < 1,433,173), showing substantial RG. As the data indicated both compensatory attributes and RG, the RT-DINO-RG showed the best fit. Hence, we examine the RT-DINO and RT-DINO-RG results in greater detail.

Like study 2, the RT-DINO estimated higher slipping parameters and lower guessing parameters, compared to the RT-DINO-RG (slipping: $M_{\text{RT-DINO}} > M_{\text{RT-DINO-RG}}$: 0.27 > 0.22; guessing: $M_{\text{RT-DINO}} < M_{\text{RT-DINO-RG}}$: 0.28 < 0.30). Also, the mean discrimination power of RT-DINO-RG exceeded that of RT-DINO ($\text{IDI}_{M(\text{RT-DINO-RG})} > \text{IDI}_{M(\text{RT-DINO})}$: 0.47 > 0.45). Ranging from 0.74 to 0.99, RT-DINO-RG's RG estimates (π) moderately correlated (r = 0.49) with the difference in the slipping parameters of RT-DINO and RT-DINO-RG (see **Figure 5**), supporting the simulation study 2 finding of

TABLE 2 | Specified Q-matrix for the real data.

Item	Label	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
1	CM033Q01S	0	0	1	0	0	0	0	1
2	CM474Q01S	0	1	0	0	0	0	0	1
3	DM155Q02C	1	0	0	0	0	0	1	0
4	CM155Q01S	1	0	0	0	0	0	1	0
5	DM155Q03C	1	0	0	0	0	0	1	0
6	CM155Q04S	1	0	0	0	0	0	1	0
7	CM411Q01S	0	1	0	0	0	1	0	0
8	CM411Q02S	0	0	0	1	0	1	0	0
9	CM803Q01S	0	0	0	1	1	0	0	0
10	CM442Q02S	0	1	0	0	0	1	0	0
11	DM462Q01C	0	0	1	0	0	0	1	0
12	CM034Q01S	0	0	1	0	1	0	0	0
13	CM305Q01S	0	0	1	0	0	1	0	0
14	CM496Q01S	0	1	0	0	0	1	0	0
15	CM496Q02S	0	1	0	0	0	1	0	0
16	CM423Q01S	0	0	0	1	0	0	0	1
17	DM406Q01C	0	0	1	0	0	1	0	0
18	DM406Q02C	0	0	1	0	0	1	0	0
19	CM603Q01S	0	1	0	0	0	0	1	0
20	CM571Q01S	1	0	0	0	0	0	1	0
21	CM564Q01S	0	1	0	0	0	1	0	0
22	CM564Q02S	0	0	0	1	0	1	0	0

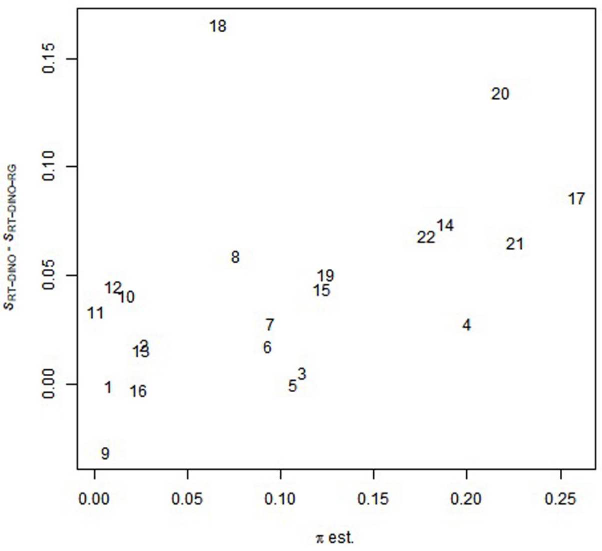

FIGURE 5 | Relationship between π-parameter in the RT-DINO-RG and the difference in the slipping parameters between the RT-DINO the RT-DINO-RG models. Numbers are item identifiers; πEst is RG estimates; and $S_{\text{RT-DINO}} - S_{\text{RT-DINO-RG}}$ is the difference in the slipping parameters of RT-DINO and RT-DINO-RG.

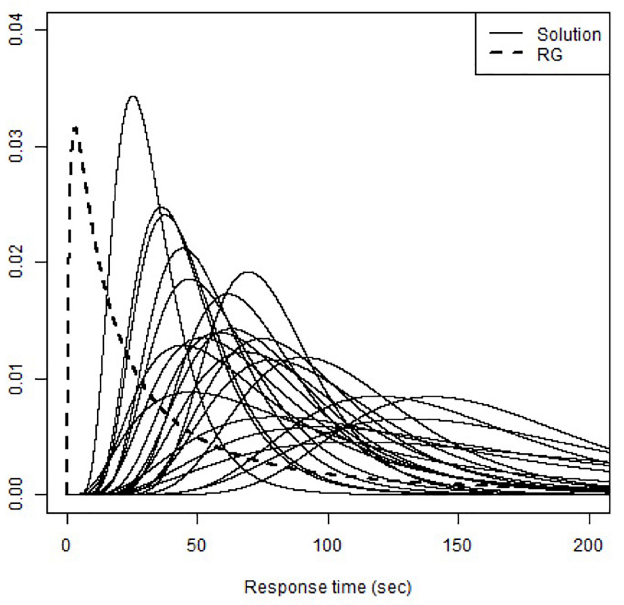

FIGURE 6 | Probability density function of RT for the RT-DINO-RG.

overestimated slipping parameters when ignoring RGs. Also, the πs of items 1–11 were generally lower than those of items 12–22. If these items appeared on the test in this sequence (item position information was not publicly available), these π results suggest that test-taker accuracy depended on their completion speed (*speededness*).

The RT-DINO-RG also uses response time to recognizes RGs and solution attempts, showing estimated mean time intensity (β_0) of 3.21 and time discrimination (κ_0) of 0.70. The various probability density functions of response time for RGs and solution attempts in the RT-DINO-RG (see **Figure 6**) suggest that students used varied answering strategies to spend more time on some items and less time on others (including RGs). The RT-DINO and RT-DINO-RG did not consistently classify mastery of the eight attributes [Cohen's κ ranged from 0.48 (quantity) to 0.98 (occupational), see **Table 3**]. Notably, few students had knowledge of the third attribute (space and shape). The simulation studies suggest that the RT-DINO-RG classifications are more reliable than the RT-DINO ones.

DISCUSSION

CDMs assess whether test-takers have the needed skills (*attributes*) to answer each test question and give suitable diagnostic feedback, but they have not adequately modeled RG vs. solution attempts with reaction times. Hence, this study

TABLE 3 | Mastery of attributes for the RT-DINO and RT-DINO-RG.

	Attributes							
	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
RT-DINO	0.090	0.381	0.057	0.133	0.173	0.235	0.366	0.349
RT-DINO-RG	0.234	0.359	0.036	0.151	0.168	0.166	0.254	0.361
Cohen's *k*	0.483	0.861	0.644	0.899	0.975	0.762	0.721	0.885

developed a new class of CDMs based on the G-DINA model (de la Torre, 2011), namely RT-GDINA-RG, with a latent indicator to jointly utilize both item responses and RTs to model RG and solution attempts to enhance cognitive diagnosis. We propose two models based on the DINA and DINO models, namely RT-DINA-RG and RT-DINO-RG.

The RT-DINA-RG and RT-DINO-RG were evaluated via (a) simulation studies with Markov chain Monte Carlo methods in JAGS and (b) real data analysis by analyzing the PISA 2015 computer-based mathematics test. Complementing Wang and Xu (2015) person-level manipulation of RG, this study manipulated RG at the item level (Wang et al., 2018). The simulation results and real data analysis showed that the RT-DINA-RG and RT-DINO-RG recovered parameters well and assessed test-takers' diagnostic results more accurately. In contrast, ignoring RGs by fitting simpler models yielded biased parameters, less reliable person speed parameter, and less classification accuracy of test results.

Hence, this study extends research showing how analyses of RT improves cognitive assessments of test-takers (e.g., van der Linden, 2008; Lee and Chen, 2011; Wang and Xu, 2015). When test-takers rapidly guess, the RT-GDINA-RG yields greater accuracies in person parameters, item parameters, and cognitive results. Therefore, researchers or users should use the RT-GDINA-RG to depict a data if RGs might occur. The choice of RT-GDINA-RG model (i.e., RT-DINA-RG or RT-DINO-RG) depends on the nature of the test items. If a test item's needed underlying constructs can compensate for one another, then RT-DINO-RG is suitable. If the underlying constructs cannot compensate for one another, then RT-DINA-RG is suitable.

Moreover, the person and item parameters of the RT-GDINA-RG were assumed to be, respectively, independent in this study. As attributes and item parameters of a CDM are often related in practice, we capture the relations between them with correlational structures (e.g., van der Linden, 2007). Note that the commonly-used multivariate normal distribution to specify the relations among person parameters is not feasible for the discrete feature of attributes in CDMs. Following Zhan et al. (2018), one can address this problem by using a higher-order latent trait to link the correlated attributes (de la Torre and Douglas, 2004), and then assuming that the person parameters (i.e., the higher-order latent trait and person speed) follow a bivariate normal distribution.

In addition, this study assumes the same probability of correctly answering an item by a RG as by guessing with none of the required attributes for the sake of simplicity. Such a naïve assumption can be further explored as in Wang and Xu (2015). Further, the RT-GDINA-RG distinguishes between solution attempt and RG for cognitive diagnosis via a latent indicator. In addition to RG, RT-GDINA-RG can be easily extended to adapt diverse test-taking behaviors and various tests' requirements. For example, we can extend CDMs to include other test-taking behaviors such as *prior knowledge/pre-knowledge* (Wang et al., 2018; Man et al., 2019) or *nonresponses* (Ulitzsch et al., 2019) if and only if the probabilities of a correct response from different latent indicators (or classes) can be clearly defined. In a high-stakes test, individuals often use pre-knowledge to correctly answer items with extremely short RT (unlike solution attempts with relatively long RT and unlike RGs with often wrong answers and short RT). Furthermore, we can adapt the functions for depicting RT and item response to the testing contexts, such as linear transformation (Wang et al., 2013), a gamma distribution to depict RT for mental rotation items (Maris, 1993), etc. (De Boeck and Jeon, 2019). Also, the item response function can be replaced by other CDMs, such as the GDM (von Davier, 2005) or the linear logistic model (Maris, 1999). Future studies can investigate these approaches.

In addition, ignoring RGs can harm the development and application of cognitive assessments (for both high- and low-stakes tests), distort test results, or invalidate inferences. For example, greater precision of test parameters via the RT-GDINA-RG ensures the quality of item bank construction and assembly of tests, especially for large-scale assessments. Their greater precision also reduces the number of necessary test items to accurately assess a test-taker's domain knowledge, thereby enabling more subdomains to be assessed. The RT-GDINA-RG results regarding time can also inform designers of timed tests regarding the time needed for different solution approaches to a test question. For example, for a timed test, items have frequent RG might because test-takers perceive that they lack sufficient time to attempt a solution. Thus, such information can provide the users of test scores to set a suitable time (e.g., increasing the response time) for completing the test. In addition, greater accuracy in the estimation of test scores increases users' confidence in the results and their subsequent inferences.

When using RT-GDINA-RG to estimate more precise person and item parameters during RG, Q-matrix is an essential component in CDM contexts. An identifiable Q-matrix ensures the consistency of a CDM estimation. In this study, the simulation studies used an identifiable Q-matrix (Xu and Zhang, 2016; Xu, 2017), and the real data analysis adopted a partially identifiable Q-matrix (Gu and Xu, 2020). To enable consistent CDM estimation, checking the identifiability of the Q-matrix in advance is crucial. Besides, for ease of use, a tutorial to introduce the RT-GDINA-RG in JAGS can be developed in future work (cf. Curtis, 2010; Zhan et al., 2019).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors contributed to the article and approved the submitted version.

REFERENCES

Curtis, S. M. (2010). BUGS code for item response theory. *J. Stat. Softw.* 36, 1–34. doi: 10.18637/jss.v036.c01

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

Gu, Y., and Xu, G. (2020). Partial identifiability of restricted latent class models. *Ann. Statist.* 48, 2082–2107. doi: 10.1214/19-AOS1878

Henson, R., and Douglas, J. (2005). Test construction for cognitive diagnostics. *Appl. Psychol. Meas.* 29, 262–277. doi: 10.1177/0146621604272623

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258—-272. doi: 10.1177/01466210122032064

Lee, Y.-H., and Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychol. Test Assess. Model.* 53, 359–379.

Lee, Y.-S., de la Torre, J., and Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. *Asia Pac. Educ. Rev.* 13, 333–345. doi: 10.1007/s12564-011-9196-3

Man, K., Harring, J. R., Ouyang, Y., and Thomas, S. L. (2018). Response time based nonparametric kullback-leibler divergence measure for detecting aberrant test-taking behavior. *Int. J. Test.* 18, 155–177. doi: 10.1080/15305058.2018.1429446

Man, K., Harring, J. R., and Sinharay, S. (2019). Use of data mining methods to detect test fraud. *J. Educ. Meas.* 56, 251–279. doi: 10.1111/jedm.12208

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables and their application to psychometric models for response times. *Psychometrika* 58, 445–469. doi: 10.1007/BF02294651

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Appl. Psychol. Meas.* 34, 521–538. doi: 10.1177/0146621609355451

OECD (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving*. Paris: OECD.

OECD (2017b). *PISA 2015 Technical Report*. Paris: OECD.

Plummer, M. (2017). *JAGS version 4.3 User Manual*.

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Appl. Psychol. Meas.* 14, 271–282. doi: 10.1177/014662169001400305

Rupp, A. A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–616. doi: 10.1111/1467-9868.00353

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20, 345—-354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin, J., and Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

Ulitzsch, E., von Davier, M., and Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *Br. J. Math. Stat. Psychol.* 2019:e12188. doi: 10.1111/bmsp.12188

van der Linden, W. J. (2006). A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *J. Educ. Behav. Stat.* 33, 5–20. doi: 10.3102/1076998607302626

von Davier, M. (2005). *A General Diagnostic Model Applied to Language Testing Data*. ETS Research Report RR-05-16. Princeton, NJ: Educational Testing Service.

Wang, C., Chang, H., and Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

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., Shang, Z., and Kuncel, N. (2018). 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

Wise, S. L., and DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *J. Educ. Meas.* 43, 19–38. doi: 10.1111/j.1745-3984.2006.00002.x

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *Ann. Stat.* 45, 675–707. doi: 10.1214/16-AOS1464

Xu, G., and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika* 81, 625–649. doi: 10.1007/s11336-015-9471-z

Zhan, P., Jiao, H., and Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *Br. J. Math. Stat. Psychol.* 71, 262–286. doi: 10.1111/bmsp.12114

Zhan, P., Jiao, H., Man, K., and Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: a tutorial. *J. Educ. Behav. Stat.* 44, 473–503. doi: 10.3102/1076998619826040

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.

APPENDIX

JAGS Code for the RT-DINA-RG and RT-DINO-RG Models

```
###################
### RT-DINA-RG###
###################
RT-DINA-RG.model
{
    for (i in 1:N) {
      for (k in 1:5) { # five attributes
            alpha[i,k] ~ dbern(ap[k])}
      tau[i] ~ dnorm(0, Inv_sigma2)
      for (j in 1:J) {
            xi[i,j] ~ dbern(pii[j]) # 0 = RG, 1 = normal
            eta[i,j] < -
                      pow(alpha[i,1],Q[j,1])*pow(alpha[i,2],Q[j,2])*pow(alpha[i,3],Q[j,3])*p
                      ow(alpha[i,4],Q[j,4])*pow(alpha[i,5],Q[j,5])
            prob[i,j] < - pow(1-slip[j],eta[i,j]*xi[i,j])*pow(guess[j],1-eta[i,j]*xi[i,j])
            r[i,j] ~ dbern(prob[i,j])
            rt.mu[i,j] < - (1-xi[i,j])*beta.0 + xi[i,j]*(beta[j] - tau[i])
            rt_kappa[i,j] < - (1-xi[i,j])*kappa2.0 + xi[i,j]*kappa2[j]
            RT[i,j] ~ dlnorm(rt.mu[i,j], rt_kappa[i,j])}}
    # Priors
    for (k in 1:5) {
      ap[k] ~ dunif(0, 1)}
    for (j in 1:J) {
      pii[j] ~ dunif(0, 1)
      slip[j] ~ dunif(0, 0.5)
      guess[j] ~ dunif(0, 0.5)
      beta[j] ~ dnorm(3, 0.1)
      kappa2[j] ~ dgamma(0.1, 0.1)
      kappa[j] < - sqrt(kappa2[j])}
    Inv_sigma2 ~ dgamma(0.1, 0.1)
    sigma < - 1/sqrt(Inv_sigma2)
    beta.0 ~ dnorm(0, 0.1) %_% I(,min(beta))
    kappa2.0 ~ dgamma(0.1, 0.1)
    kappa.0 < - sqrt(kappa2.0)}
###################
### RT-DINO-RG###
###################
RT-DINO-RG.model
{
    for (i in 1:N) {
      for (k in 1:5) { # five alphaibutes
            alpha[i,k] ~ dbern(ap[k])}
    tau[i] ~ dnorm(0, Inv_sigma2)
    for (j in 1:J) {
            xi[i,j] ~ dbern(pii[j]) # 0 = RG, 1 = normal
            eta[i,j] < - 1-pow(1-alpha[i,1],Q[j,1])*pow(1-alpha[i,2],Q[j,2])*pow(1-
                    alpha[i,3],Q[j,3])*pow(1-alpha[i,4],Q[j,4])*pow(1-
                    alpha[i,5],Q[j,5])
            prob[i,j] < - pow(1-slip[j],eta[i,j]*xi[i,j])*pow(guess[j],1-eta[i,j]*xi[i,j])
            r[i,j] ~ dbern(prob[i,j])
            rt.mu[i,j] < - (1-xi[i,j])*beta.0 + xi[i,j]*(beta[j] - tau[i])
            rt_kappa[i,j] < - (1-xi[i,j])*kappa2.0 + xi[i,j]*kappa2[j]
            RT[i,j] ~ dlnorm(rt.mu[i,j], rt_kappa[i,j])}}
    # Priors
    for (k in 1:5) {
      ap[k] ~ dunif(0, 1)}
    for (j in 1:J) {
      pii[j] ~ dunif(0, 1)
      slip[j] ~ dunif(0, 0.5)
      guess[j] ~ dunif(0, 0.5)
      beta[j] ~ dnorm(3, 0.1)
      kappa2[j] ~ dgamma(0.1, 0.1)
      kappa[j] < - sqrt(kappa2[j])}
    Inv_sigma2 ~ dgamma(0.1, 0.1)
    sigma < - 1/sqrt(Inv_sigma2)
    beta.0 ~ dnorm(0, 0.1) %_% I(,min(beta))
    kappa2.0 ~ dgamma(0.1, 0.1)
kappa.0 < - sqrt(kappa2.0)}
```