ABSTRACT

Title of Dissertation: EXPLANATORY COGNITIVE DIAGNOSTIC MODELING INCORPORATING RESPONSE TIMES

Xin Qiao, Doctor of Philosophy, 2021

Dissertation directed by: Professor, Hong Jiao
Measurement, Statistics and Evaluation
Department of Human Development and Quantitative Methodology

The current study proposes the explanatory cognitive diagnostic models (CDMs) incorporating response times (RTs) with item covariates on both the item response side and the RT side. There are two main contributions of the current study. One appealing usage of this model is that scored item covariates can be used to predict item parameters when item calibration is not feasible in diagnostic assessments while the other is that the cognitive theories underlying the test design can be evaluated. Model parameter estimation is explored using the Bayesian Markov chain Monte Carlo (MCMC) method. A Monte Carlo simulation study is conducted to examine the parameter recovery of the proposed model under different simulated conditions in comparison to a few competing models. The results indicate that model parameter could be well recovered using the MCMC approach. Further, the application of the proposed model is illustrated using the Programme for International Student Assessment (PISA) 2012 problem-solving items using both item response and item RT data.

EXPLANATORY COGNITIVE DIAGNOSTIC MODELING INCORPORATING RESPONSE TIMES

by

Xin Qiao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Hong Jiao, Chair
Professor Gregory R. Hancock
Professor Robert W. Lissitz
Professor Yang Liu
Professor Xin He, Dean's Representative

© Copyright by
Xin Qiao
2021

Dedication

To those who have helped, supported, and encouraged me along the way.

Acknowledgement

First, I would like to thank Dr. Jiao and Dr. Lissitz for the opportunity to work at MARC as a research assistant for most of my graduate student life. It has been a pleasure and a valuable experience to have worked with the two professors and many other excellent peer students in the past years. I would also like to express my gratitude to Dr. Jiao who provided me with the opportunities during my research journey at EDMS. Second, I would like to thank Dr. Liu who provided valuable insights and suggestions on my questions from technical perspectives during my doctoral study. Third, I sincerely appreciate that Dr. Hancock provided me with the opportunity to come and study at EDMS which turned out to be an important experience in my life. Last but not least, I would like to thank Dr. He for his generosity for his service on my dissertation committee.

I would also like to express my gratitude to people in the larger EDMS and UMD community. I am fortunate enough to have met Dr. Eric Slud who has greatly influenced me in the study of statistics. I will always remember the period of time when I traveled across the campus to the missing data class with him which was simply life-changing. I am also thankful to Dr. Tracy Sweet who brought us together on an exciting network project that turned out to be a rewarding research experience. Lastly, I would like to thank all the outstanding faculty members and graduate students at EDMS who have helped me along the way of my doctoral study.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Statement of Problem

Cognitive diagnostic assessments have gained increasing popularity and importance in education. The development of these assessments aims at providing more detailed information on the strengths and weaknesses of students' skills, knowledge, and ability (Huff & Goodman, 2007). Such fine-grained diagnostic information can be used as substantial feedback to teachers' instruction and students' learning (e.g., Tang & Zhan, 2020; Wang et al., 2020). The cognitive diagnostic models (CDMs; e.g., Rupp et al., 2010) are an important family of latent variable models that can yield diagnostic information about the students given their item responses. Specifically, CDMs employ discrete latent variables (i.e., attributes) that categorize students into their latent profiles indicating the mastery status of each attribute. This characteristic makes CDMs especially useful in measuring unobservable attributes that are required to solve the tasks.

Recently, latent variable models and test design frameworks pertaining to cognitive theories have been combined in the usage of diagnostic assessments to produce meaningful results. In general, the principled test design frameworks play a crucial role in developing diagnostic assessments with construct validity (e.g., Embretson, 1983, 1998). Traditionally, test development methods mainly consider construct validity as establishing correlated test scores with other measures (nomothetic span). However, contemporary test design frameworks such as the cognitive design system (CDS; Embretson, 1994, 1998) have integrated cognitive theories, task strategies, and knowledge bases in the item design (construct representation). Therefore, test design frameworks bridge the underlying theories and test items so that the interpretations from the diagnostic assessments are meaningful and defensible.

In addition, test design frameworks not only specify item stimulus features but also predict the psychometric properties of the items (Embretsen, 1999). It is desirable to control the item parameters in the assessment development (e.g., Embretson, 1998; Embretson & Yang, 2006; Glas & van der Linden, 2003; Gorin, 2005; Lathrop & Cheng, 2017). In the automatic item generation scenario, a large number of items with known psychometric properties need to be generated (e.g., Embretson & Yang, 2006; Lathrop & Cheng, 2017). For example, the item cloning approach (e.g., Bejar et al., 2003) takes an item with established psychometric properties as an item generation model for new items. Specifically, certain item covariates are treated as variables and can be substituted in the new item. In addition, based on the CDS approach, the item covariates are determined by an information-processing theory and are related to item parameters (Embretson, 1998, 1999). Therefore, it is possible to generate a large number of items with known psychometric properties predicted by item covariates. In the computerized adaptive testing (CAT), it is important to control item parameters to maintain a balanced item pool (e.g., Glas & van der Linden, 2003).

However, it is operationally challenging to write items with targeted characteristics. Built on the item response theory (IRT) models, the explanatory item response models (De Boeck & Wilson, 2004) serve as a methodological framework that accommodates item effects due to item covariates (e.g., test design factors, item stimulus features) that may affect item parameters. From the explanatory perspective, how item response data are generated can be explained by the item covariates (Wilson et al., 2008). One fundamental explanatory item response model is the linear logistic test model (LLTM; Fischer, 1973) where the item difficulty parameter is decomposed into a linear combination of basic parameters that reflect theoretical beliefs that are closely related to test design principles. A plethora of studies have focused on using item

covariates to explain item parameters under the LLTM framework (e.g., De Boeck, 2008; Hohensinn et al., 2008). For example, Hohensinn et al. (2008) examined the item position effect on the item difficulty. Thus, the explanatory item response modeling framework helps item writers and test developers to understand how certain item covariates affect item parameters.

Meanwhile, as the development of computer-based diagnostic assessments, item response times (RTs) became available in the psychometric modeling for cognitive diagnosis purposes. Many RT models have been developed in educational assessment settings, such as the lognormal model (van der Linden, 2006), the gamma model (Maris, 2013), and the Box-Cox model (Klein Entink et al., 2009). Methodologies for simultaneously analyzing item responses and RTs have also been developed, among which the hierarchical modeling framework (van der Linden, 2007; Zhan et al; 2018) is the most commonly used approach.

RTs are useful test data and contain rich information about students' cognitive processes in solving the tasks. For example, the RTs have been used to assess students' learning progression (Wang et al., 2018), to model learner heterogeneity (Zhang & Wang, 2018), and to assess learning outcomes (Wang et al., 2020). Furthermore, explanatory RT models (e.g., Klein Entink et al., 2009; van Breukelen, 2005) that incorporate item covariates have been proposed to evaluate cognitive theories and explain students' cognitive processes. It has been shown that item parameters from both the item response side and the RT side can be explained by item covariates related to test design theories. For example, Embretson and Yang (2007) stated that item RTs are crucial dependent variables to be obtained together with item responses to assess the underlying test design theory. The RTs have been analyzed using a mixed regression model with item covariates specified as fixed effects (Daniel & Embretson, 2010; Embretson & Daniel, 2008). In this way, RTs can provide additional information about the item characteristics with respect to

test design. Therefore, developing explanatory RT models is of theoretical and practical importance in diagnosis assessments.

Recently, several studies have investigated explanatory CDMs (e.g., Liao & Jiao, 2018; Park et al., 2018). In Liao and Jiao (2018), the item parameters in the CDMs are decomposed in the LLTM manner. Specifically, the slipping and guessing parameters were decomposed by linguistic item covariates extracted from the item stimulus in a mathematics assessment. Similar to the explanatory item response models, explanatory CDMs can help improve the development of diagnostic assessments, provide information for item generation, and shed light on students' problem-solving processes through the estimation of item effects from the item covariates. However, RT information has been neglected in the development of explanatory CDMs. It is necessary to incorporate RT in the explanatory CDMs to facilitate the test development process.

1.2 Purpose

The current study proposes the explanatory CDM incorporating RTs. Specifically, the proposed model incorporates item covariates into the item parameters in the CDM and the RT model using the hierarchical modeling approach (van der Linden, 2007; Zhan et al., 2018). The proposed model aims to use known item covariates to predict item parameters in the CDM and RT model, respectively. Based on existing studies (e.g., Embretson, 1999; Liao & Jiao, 2018), the item covariates can be grouped into three major categories: 1) test design covariates (e.g., item types); 2) linguistic covariates (e.g., grammatical density, vocabulary density); 3) cognitive covariates (e.g., steps required to solve the item). Figures 1 and 2 provide examples of an easy item and a difficult item from the PISA 2012 problem-solving items. Several cognitive covariates reflect the difficulty differences. First, the easy item is from a familiar context (i.e., a public traffic map) for the respondents, while the difficulty item is from a less familiar context

(i.e., a virtual ticket machine). Second, the easy item provides a small amount of explicit information on the task. It asks the respondents to select the quickest route between two locations that takes 31 minutes. However, the difficult item asks the respondents to buy the cheapest ticket without providing any specific guidance, thereby requiring more steps to find out the best solutions. Third, the easy item has less possible actions, while the difficult item has more possible actions. From the CDM perspective, the difficult item may be harder to guess but easier to slip. From the RT perspective, the difficult item may take a longer time to complete. Therefore, these cognitive covariates can be quantified and used as item covariates that explain the item parameters in CDMs and RT models.

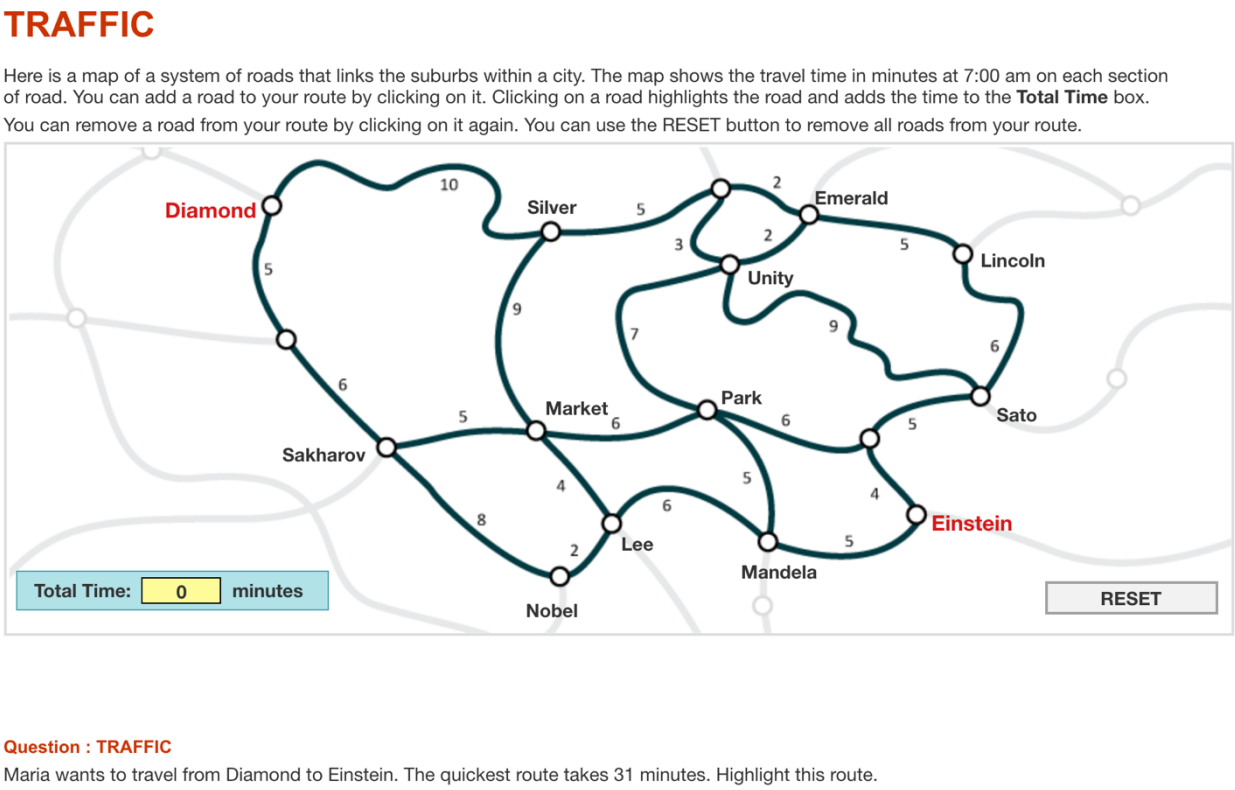


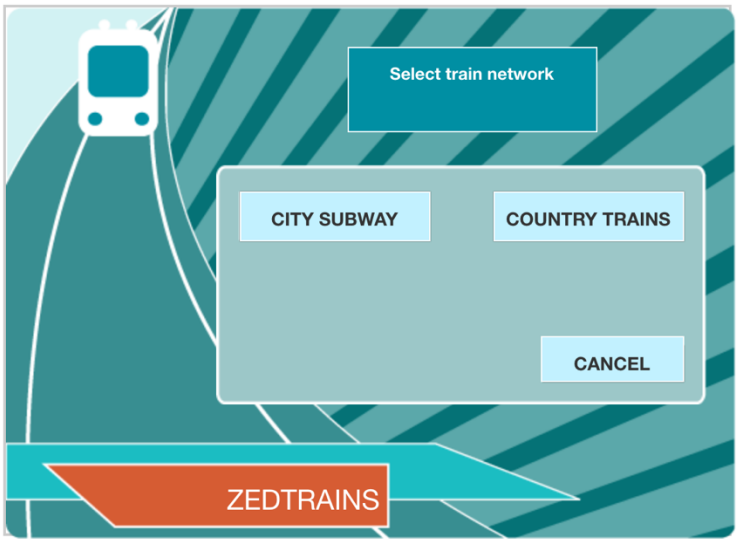
Figure 1. An easy PISA 2012 problem-solving item.

TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- Choose the train network you want (subway or country).
- Choose the type of fare (full or concession).
- Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.



Question TICKETS

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY.

Once you have pressed BUY, you cannot return to the question.

Figure 2. A difficult PISA 2012 problem-solving item.

The purpose of proposing the explanatory CDM incorporating RTs is three-fold. First, it can provide item writers with information about the targeted item characteristics (e.g., slipping, guessing, time intensity). Second, it can facilitate automatic item generation especially when a large number of items are needed. Third, it can be used to evaluate respondents' cognitive processes through the estimation of the item effects.

A Monte Carlo simulation study is conducted to examine the parameter recovery of the proposed model under different conditions based on the real test design structure of the PISA 2012 problem-solving items. In addition, the simulation study compares the performance of the proposed model and that of several competing models under different simulation conditions. The competing models include models with item covariate misspecifications, separate models for RA

and RT, and models without item covariates but using the two-step estimation to obtain the regression coefficient estimates of the item covariates (e.g., Green & Smith, 1987).

An empirical data analysis is conducted as an application of the proposed model to the PISA 2012 problem-solving items (OECD, 2012) using both item responses and item RT data. The empirical data analysis intends to demonstrate the interpretations the item effects of the item covariates in a cognitive diagnosis context.

Specifically, the current study aims to answer the following questions. The first three questions are to be answered by conducting the simulation study, while the last two questions are to be answered using an empirical data analysis:

1) How do the relative model fit indices perform in identifying the proposed model as the best data-fitting model in comparison with alternative models with misspecifications?
2) How accurate is the parameter recovery of the proposed model under different simulation conditions (i.e., sample size, test length, predictive power of the item covariates)?
3) How do the model misspecifications (e.g., the two-step estimation method and the one-step estimation) affect parameter recovery compared to that of the proposed model?
4) According to the empirical data analysis, do the item covariates have significant effects on the item parameters from CDM and the RT model, respectively?
5) How is the consistency of the item parameter estimates and attribute estimates among the data-fitting models?

1.3 Significance of the Study

In this study, the operational and methodological contributions of the development of the explanatory CDMs incorporating RTs are as follows. From the operational perspective, explanatory CDMs and RT models can help item writers to create items with targeted item parameters (e.g., slipping and guessing) in diagnostic assessments. One appealing usage of this model is that item covariates can be used to predict item parameters when item calibration is not feasible in diagnostic assessments. For example, in automatic item generation, test items generated for the respondents are on-the-fly and no pre-calibration of the items is conducted, but the item parameters can be predicted by the information or values from item covariates (Embretson & Yang, 2006). Although there are only a few relevant studies, the automatic item generation applications are needed in diagnostic assessment settings. Another scenario is the cognitive diagnostic computerized adaptive testing (CD-CAT) where a large number of items with a broad range of psychometric properties are needed for test security reasons. However, items targeted for respondents on the two ends of the ability continuum are relatively more challenging to develop. With fewer targeted items, the ability estimates for these respondents could be less accurate. Thus, using item covariates is helpful in the item generation for CAT (e.g., Glas & van der Linden, 2003). Additionally, using item covariates to predict time-related parameters in the RT models is helpful in setting time limits in computer-based tests (e.g., van der Linden, 2011). Under these situations, the proposed model is beneficial in the prediction and development of items with controlled psychometric properties.

Another operational advantage of using the explanatory CDMs incorporating RTs is that the cognitive theories underlying the test design can be evaluated. It is assumed that the item covariates reflect both the differences in the CDM item parameters (e.g., the slipping and

guessing parameters) and the amount of time it takes to answer the items. Therefore, this new model can be used to validate not only the cognitive theories on the CDM item parameters but also on the time-related item parameters.

From the methodological perspective, explanatory DCMs and explanatory RT models are far less investigated than IRT models. The current study advances the state of knowledge of explanatory psychometric models in several ways. First, the proposed model represents a novel method to incorporate item responses and RTs that account for the item covariate effects for cognitive diagnosis. Second, it investigates the parameter recovery of the proposed model that combines the CDM and the RT model under different simulated testing scenarios. This shows the performance of the proposed model and its potential usage. Third, it compares the performance of the proposed model with several competing models. The conclusions can help researchers select appropriate models in practical settings.

In summary, the estimated item effects using the proposed model can provide useful information on the item development and respondents' cognitive processes. In addition, methodological issues (e.g., parameter recovery) related to the proposed model are investigated in the current study.

Chapter 2: Literature Review

Given that the current study is on CDM based joint modeling of response accuracy (RA) and RT using an item explanatory approach, the literature review is categorized into three sub-topics based on the types of observed variables: a) item response models with a focus on CDMs, b) RT models, and c) joint models of RA and RT. Within each area, the studies are further categorized based on whether they used the standard modeling approach or the explanatory modeling approach. Specifically, the literature review section begins with an introduction to the theoretical foundation of the modeling frameworks related to the current study. Then, existing studies related to the three sub-topics (i.e., the item response models, RT models, and joint models of RA and RT) are discussed with an emphasis on the model formulations and assumptions. Lastly, the estimation methods for the existing joint models of RA and RT are summarized.

2.1 Theoretical Foundation

The latent variable modeling framework, the joint modeling framework of RA and RT, and the explanatory item response modeling framework are briefly introduced as the theoretical foundation of the current study.

Latent variable models. Latent variable models are statistical models that incorporate latent or unobserved variables in the analysis of observed variables. The latent variables usually represent a hypothetical construct being measured, while the observed variables are measured from collected data. Thus, latent variable models serve as a linkage between the latent and observed variables (Spearman, 1904).

In educational measurement settings, the most common observed variables are item responses, which usually are either dichotomous or polytomous. Therefore, some specific latent

variable models have been developed for the analysis of these data, with the latent variables being either continuous or categorical. For example, the item response theory (IRT) models assume continuous latent variables (e.g., math ability), while CDMs assume discrete latent variables (e.g., skill mastery status). The current literature review mainly focuses on CDMs for dichotomous item responses based on the purpose of the current study. However, the CDMs introduced in Section 2.2 can be extended to accommodate polytomous item responses. In addition, as the development of computer-based assessments, continuous observed variables such as item RTs can also be collected. Therefore, the RT models that usually assume continuous latent variables (e.g., work speed) have been developed and are reviewed in Section 2.3.

Joint modeling of RA and RT. Given that the separate models for RA and RT do not capture the relationships between the two variables, researchers have developed the joint modeling framework for RA and RT (e.g., van der Linden, 2007). The joint distribution of the item response and RT for a single item is given by:

$$f(y_i, t_i \mid \theta, \tau; \alpha_i, \beta_i), \tag{2.1.1}$$

where y_i and t_i denote the item response and item RT for item *i*, respectively; α_i and β_i denote the item parameters for item i; θ and τ denote the latent ability and latent speed. It has been shown that different approaches can be taken to specify the joint distribution of the item response and RT for a single item (Bloxom, 1985). As shown in Ranger and Ortner (2012), given the conditional independence (CI) assumption between RA and RT, the expression in equation 2.1.1 can be factored as:

$$f(y_i \mid \theta; \alpha_i) f(t_i \mid \tau; \beta_i), \tag{2.1.2}$$

where θ and τ are assumed to be the latent traits for RA and RT, respectively; α_i and β_i denote the item parameters for RA and RT, respectively. Note that $f(y_i \mid \theta; \alpha_i)$ can be specified

using either the IRT models or CDMs and $f(t_i \mid \tau; \beta_i)$ can be specified using a RT model. Further, when the CI assumption is violated, the expression in the equation 2.1.1 can be factored as:

$$f(y_i \mid \theta; \alpha_i) f(t_i \mid y_i, \tau; \beta_i), \tag{2.1.3}$$

or

$$f(t_i \mid \tau; \beta_i) f(y_i \mid t_i, \theta; \alpha_i), \tag{2.1.4}$$

where the conditional dependence (CD) of RA on RT or the CD of RT on RA is accommodated. The joint models for RA and RT are reviewed in the section 2.4.

Explanatory psychometric models. Explanatory item response models (De Boeck & Wilson, 2004) have been developed to incorporate covariates or covariates into the person side, or item side, or both sides of the IRT models. Similar approaches have been taken in the development of RT models (e.g., Scheiblechner, 1979) and the joint models of RA and RT (e.g., Klein Entink et al., 2009). The current literature review covers the explanatory psychometric models that incorporate both item covariates or person covariates but gives an emphasis on those incorporating item covariates given the goal of the current study. The explanatory item response models are summarized in section 2.2.2. The explanatory response time models are summarized in section 2.3.2. The explanatory joint models are summarized in section 2.4.3.

2.2 Models for Item Responses

2.2.1 CDMs

CDMs are psychometric models that aim at providing fine-grained diagnostic information about respondents' mastery status of several latent attributes measured by assessment items. Essentially, CDMs are restricted latent class models where the latent classes are defined as the attribute profiles. This section first presents the parameterization of the general unrestricted

latent class models. Then, several core CDMs are introduced to show how restrictions are applied on the item parameters.

Using the formal representation of the latent class models, the marginal likelihood of the response pattern of respondent *j* is given as:

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{c=1}^{C} \upsilon_c \prod_{i=1}^{I} \pi_{ic}^{y_{ij}} (1-\pi_{ic})^{1-y_{ij}}, \tag{2.2.1}$$

where $P(\mathbf{Y_j} = \mathbf{y_j})$ denotes the probability of observing a response pattern $\mathbf{y_j}$ from respondent *j*; y_{ij} denotes the item response from respondent *j* to item *i*; π_{ic} denotes the probability of obtaining a correct response to item *i* given that the respondent is in latent class *c*; $\prod(\cdot)$ indicates that the item responses are assumed to be conditional independent across items and the product is the observed likelihood of the response pattern in latent class *c*; υ_c denotes the mixing proportion of the latent class *c* (i.e., the proportion of respondents in the latent class *c*); $\sum(\cdot)$ indicates that the product of the mixing proportion and the observed likelihood is summed across all latent classes.

In CDMs, the latent classes refer to the attribute profiles which can be denoted as $\boldsymbol{\alpha}_c$ = ($\alpha_1,...,\alpha_K$), indicating the latent classes are determined by the mastery status of the *K* latent attributes of respondent *j*. Given the scope of the current study, the mastery status of the attribute is assumed to be binary. There are many ways to specify the distribution of the $\boldsymbol{\alpha}_c$. If every attribute profile is permissible, the total number of latent classes equals 2^K. It is also possible that some attribute profiles are not permissible. Then, the mixing proportions for these attribute profiles are constrained to be 0, while the mixing proportions for the other attribute profiles are estimated (e.g., Liu & Huggins-Manley, 2016). In either case, the mixing proportion υ_c for each

attribute profile can be estimated with the constraint $\sum_{c=1}^{C} \upsilon_c = 1$. In addition, the higher-order structural model (de la Torre & Douglas, 2004) can be specified to take into account the correlation among the attributes by introducing a general latent ability θ_j. In this case, the logit of the probability of mastering latent attribute *k* given θ_j is given by:

$$\text{logit}(P(\alpha_{jk} = 1 \mid \theta_j)) = \gamma_k \theta_j + \lambda_k, \tag{2.2.2}$$

where logit = $\log \frac{x}{1-x}$; γ_k and λ_k are attribute-specific slope and intercept parameters. It is assumed that the attributes are conditional independent given the latent ability. When the higher-order structural model is specified, two parameters per attribute (i.e., γ_k and λ_k) need to be estimated instead of the mixing proportions. Therefore, the higher-order structure reduces the number of parameters to be estimated in the structural component from $2^K - 1$ to 2*K*.

Many CDMs have been developed to model the probability of obtaining a correct response for respondent *j* in latent class *c*, π_{ic}. Some rest on more stringent assumptions, such as the deterministic input, noisy "and" gate (DINA; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model and the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006) model. Others have more general parameterizations, including the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) and the general diagnostic model (GDM; von Davier, 2005). The DINA model is one of the simplest and most commonly used CDM. It is a non-compensatory model that assumes a respondent needs to master all attributes required in an item to answer it correctly. The latent response variable ξ_{ij} is created based on the conjunctive condensation rule (Maris, 1999), which is given as:

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}} \ , \tag{2.2.3}$$

where q_{ik} is an element in the *Q*-matrix (Tatsuoka, 1983, 1985) that connects item *i* to the attributes it measures. The *Q*-matrix reflects the cognitive specification of tests (Leighton et al., 2004). Specifically, $q_{ik} = 1$ if the attribute *k* is measured by item *i* and $q_{ik} = 0$ otherwise; α_{jk} indicates the mastery status of attribute *k* of the respondent *j*; $\prod(\cdot)$ indicates the conjuctive rule that all attributes measured by an item need to be mastered to obtain a correct response. The probability of obtaining a correct response in the DINA model for respondent *j* with attribute profile $\boldsymbol{\alpha}_j$ is further determined by two item parameters (i.e., slipping s_i and guessing g_i), which is given as:

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_j) = (1 - s_i)^{\xi_{ij}} g_i^{1-\xi_{ij}} \ , \tag{2.2.4}$$

where $1 - s_i$ is the probability of obtaining a correct response when $\xi_{ij} = 1$; g_i is the probability of obtaining a correct response when $\xi_{ij} = 0$. Conventionally, $1 - s_i > g_i$ is constrained to ensure that respondents who have mastered all attributes have larger probability to obtain a correct response than those who lack at least one attribute (Junker & Sijtsma, 2001). The DINO model is the compensatory counterpart of the DINA model. It assumes that the respondent *j* needs to master at least one attribute to answer the item correctly. The latent response variable ξ_{ij} is created based on the disjunctive condensation rule (Maris, 1999), which is given as:

$$\xi_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{jk})^{q_{ik}} \ , \tag{2.2.5}$$

while the probability of obtaining a correct response in the DINO model has the same form as shown in equation 2.2.4. Both the DINA model and the DINO model have stringent assumptions.

For example, the DINA model assumes that respondents who lack one or more attributes to one item have the same probability of obtaining a correct response; the DINO model assumes that respondents who master one or more attributes to one item have the same probability of obtaining a correct response. However, these assumptions may not hold in real-world situations.

The LCDM (Henson et al., 2009) is a more generalized CDM where the probability of obtaining a correct response is given by:

$$\text{logit}\, P(X_{ij} = 1 \mid \boldsymbol{\alpha}_j) = \lambda_{i,0} + \boldsymbol{\lambda}_i^T h(\boldsymbol{\alpha}_j, q_i), \tag{2.2.6}$$

where $\lambda_{i,0}$ denotes the intercept term that is interpreted as the logit of the probability of obtaining a correct response to item *i* when respondent *j* does not master any of the attributes; $\boldsymbol{\lambda}_i^T h(\boldsymbol{\alpha}_j, q_i)$ denotes a linear function of the main effects and the interaction effects of the required attributes of item *i*, which is shown as:

$$\boldsymbol{\lambda}_i^T h(\boldsymbol{\alpha}_j, q_i) = \sum_{k=1}^{K} \lambda_{i,1,(k)} \alpha_{jk} q_{ik} + \sum_{k_1 < k_2} \lambda_{i,2,(k_1,k_2)} \alpha_{jk_1} \alpha_{jk_2} q_{ik_1} q_{ik_2} + \ldots, \tag{2.2.7}$$

where $\lambda_{i,1,(k)} \alpha_{jk} q_{ik}$ indicates the main effects of attribute k; $\lambda_{i,2,(k_1,k_2)} \alpha_{jk_1} \alpha_{jk_2} q_{ik_1} q_{ik_2}$ indicates the two-way interaction effect of the attribute k_1 and k_2. In its saturate form, all possible main effects and two or higher order interaction effects can be estimated in the LCDM. Alternatively, some of the main effects and interaction effects of the attributes can be constrained to be 0. In fact, the DINA model is a constrained LCDM where only the highest-order interaction effect of the required attributes is retained (Rupp et al., 2010). The identification of the LCDM requires that all main effects to be positive and the interaction terms are constrained such that respondents who master more attributes have higher probability of obtaining a correct response (Lao, 2016).

The GDM (von Davier, 2005, 2014) is a more generalized CDM that include the LCDM as a special case. The GDM can accommodate discrete and continuous latent attributes and

response data with two or more response categories. For a more comprehensive review of the CDMs, readers can refer to Rupp and Templin (2008).

2.2.2 Explanatory Item Response Models

Explanatory item response models (De Boeck & Wilson, 2004) have been developed to explain item response data by incorporating covariates from person side, item side, or both sides (i.e., person or item covariates) into the statistical models. The purpose for such explanatory approach is to investigate how the respondents' task performance relates to the external factors such as test design elements (e.g., Embretson, 1985). In addition, incorporating auxiliary information from the persons and items can improve the estimation precision and stability of the item parameter estimation (Mislevy, 1987; Mislevy, 1988; Mislevy & Sheehan, 1988; Mislevy & Sheehan, 1989). Meanwhile, explanatory CDMs (Liao & Jiao, 2018; Park & Lee, 2014; Park et al., 2018) have also been proposed for similar purposes. The current study focuses on psychometric models with covariates on the item side. Therefore, these models are reviewed in the following paragraphs.

Explanatory IRT models. For a unidimensional IRT model, the marginal probability of the observed response pattern of respondent j (j = 1, 2,…, J) to test items (i = 1, 2, …, I) is written as (assuming local item independence):

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \int \prod_{i=1}^{I} P(Y_{ij} = y_{ij} \mid \theta) f(\theta) d\theta \,, \tag{2.2.8}$$

where $f(\theta)$ defines the probability distribution function of the latent ability θ, and $P(Y_{ij} = y_{ij} \mid \theta)$ is given by a specific IRT model and defines the probability of a response category given the ability level. One of the most basic unidimensional IRT models is the Rasch

model (Rasch, 1960), in which the probability of a correct response of respondent *j* to an item *i* is given by:

$$P(Y_{ij} = 1 \mid \theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \tag{2.2.9}$$

where θ_j and b_i denote the latent ability for respondent *j* and item difficulty for item *i*, respectively. The Rasch model does not provide information on what manifest features of the items may affect the item difficulty of the items. Alternatively, the linear-logistic test model (LLTM; Fischer, 1973), which is nested within the Rasch model, decomposes the item difficulty parameters such that:

$$P(Y_{ij} = 1 \mid \theta_j, \eta_k, q_i) = \frac{\exp(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k)}{1 + \exp(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k)}, \tag{2.2.10}$$

where η_k (k = 1, 2, …, K; $K < I$) denotes the basic parameters that represent the difficulty of the item covariates; q_{ik} denotes the design matrix that links the I items to the K item covariates. Therefore, the LLTM can be used to evaluate the difficulty of specific aspects of the items that are usually defined based on psychological and cognitive theories.

Both the Rasch model and the LLTM make the stringent assumption that all items have the same discriminatory power across respondents. However, this assumption can be relaxed by using the two parameter logistic (2PL) model (Lord & Novick, 1968), in which the probability of a correct response of respondent *j* to an item *i* is given by:

$$P(Y_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))}, \tag{2.2.11}$$

where a_i denotes the item discrimination parameter for item *i*, while other notations are consistent with those in the Rasch model. Thus, the 2PL model allows the estimation of both the item discrimination and item difficulty parameters. Embrestson (1999) further extended the LLTM to the 2PL-constrained model, which is given by:

$$P(Y_{ij}=1\,|\,\theta_j,\eta_k,\xi_k,q_i)=\frac{\exp\left[(\sum_{k=1}^{K}q_{ik}\xi_k)(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k)\right]}{1+\exp\left[(\sum_{k=1}^{K}q_{ik}\xi_k)(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k)\right]}, \tag{2.2.12}$$

where ξ_k denotes the basic parameters that represent the discriminatory power of the item covariates, while other notations are consistent with those in the LLTM. It can be seen that the 2PL-constrained model is nested within the 2PL model by decomposing both the item discrimination and item difficulty parameters. In this way, the 2PL-constrained model can be used to evaluate both the discriminatory power and difficulty of specific cognitive aspects of the items.

Both the LLTM and the 2PL-constrained model impose a strong assumption that the item covariates can predict the item parameters in the Rasch model or the 2PL model perfectly. This implies that the items share the same structure in the design matrix and would have exactly the same difficulty or/and discrimination. However, it has been found that item difficulties may have considerable variability even after accounting for the design structure (Embretson, 1998). To take such variability into account, a random error term has been introduced into the LLTM (Janssen et al., 2004; De Boeck, 2008), where the item difficulty is decomposed as:

$$b_i=\sum_{k=1}^{K}q_{ik}\eta_k+\varepsilon_i\,, \tag{2.2.13}$$

where ε_i follows the normal distribution, $\varepsilon_i \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$. A similar approach can be generalized to the 2PL-constrained model where both the item discrimination and difficulty parameters are allowed to vary among items that share the same design structure (e.g., Glas & van der Linden, 2003; De Jong et al., 2007).

The model with restrictions on the item difficulty (MIRID; Butter et al., 1998) is closely related to the LLTM. Similar to the LLTM, the MIRID model decomposes the item difficulty parameter in the Rasch model into a linear composite and assumes the perfect prediction of the item difficulty. The design matrix used in the LLTM is known a' priori. However, the MIRID model is applied in the scenario where a composite item consisting of several component items exists in the test. The weights in the design matrix used in the MIRID model are to be estimated using the item responses to the component items. Specifically, the probability of obtaining a correct response in the MIRID model is given by:

$$P(Y_{ij} = 1 \mid \theta_j, \eta_{ik}, \sigma_k) = \frac{\exp(\theta_j - \sum_{k=1}^{K} \sigma_k \eta_{ik} + \tau)}{1 + \exp(\theta_j - \sum_{k=1}^{K} \sigma_k \eta_{ik} + \tau)}, \tag{2.2.14}$$

where σ_k is the weight of the component difficulty η_{ik} in the composite item *i*; τ denotes the scaling intercept; θ_j denotes the person ability of respondent *j*. In the MIRID model, σ_k, η_{ik} and τ all need to be estimated, while only the basic parameters η_k need to be estimated in the LLTM. Since the current study focused on explanatory psychometric models with item covariates, the applications and the extensions of the MIRID model are not introduced.

Explanatory CDMs. Several studies have been conducted to incorporate covariates into CDMs to facilitate understanding of how such variables relate to the item parameters or the mastery of fine-grained latent attributes (Ayers et al., 2013; Liao & Jiao, 2018; Park & Lee,

2014, 2019; Park et al., 2018). Ayers et al. (2013) and Park and Lee (2014) have associated observed person covariates to the response probability or a single attribute of the DINA model using a logistic function. Similarly, Park et al. (2018) and Park and Lee (2019) further incorporated both observed and latent covariates into the DINA model to explain the attribute profiles. They showed how various covariates can influence the mastery of the latent attributes. However, studies using covariates to explain item parameters in the CDMs are few. Liao and Jiao (2018) is the first attempt to extend the LLTM into the CDM framework. Specifically, they decomposed the slipping and guessing parameters in the DINA model using linguistic features extracted from the item stimulus. The decomposition considered both the LLTM and the LLTM + ε_i formulations:

$$\text{logit}(s_i) = \gamma_0 + \sum_{k=1}^{K} \gamma_k X_{ik} , \tag{2.2.15}$$

and

$$\text{logit}(s_i) = \gamma_0 + \sum_{k=1}^{K} \gamma_k X_{ik} + \varepsilon_i . \tag{2.2.16}$$

The guessing parameter was decomposed in the same fashion. They found some linguistic features have significant effects on the item parameters of the DINA model.

2.3 Response Time Modeling

In educational assessments, RT has been shown to be a useful data source in various applications, including cheating detection (e.g., Marianti et al., 2014; Sinharay & Johnson, 2019; Sinharay, 2020), setting test time limit (e.g., van der Linden, 2011a), speededness control in adaptive testing (e.g., van der Linden et al., 1999; van der Linden, 2009; van der Linden, 2011b), and improving ability estimation precision (e.g., Bolsinova, & Tijmstra, 2018; van der Linden, 2007). At the same time, RT models have been developed in formative assessment settings to

better understand students learning behavior (e.g., Wang et al., 2018) and to improve the estimation accuracy in latent profile classifications (e.g., Zhan et al., 2018). The current section first reviews the statistical distributions that have been proposed to model RT. Then, RT models using the explanatory approach are summarized.

2.3.1 Standard RT Models

In various application settings, RT distributions are usually positively skewed. Therefore, many studies chose to normalize the RT distributions using the log-transformation (e.g., Thissen, 1983; van der Linden, 2006) in statistical modeling. The lognormal model proposed by van der Linden (2006) is one of the most commonly used RT models, which is presented as:

$$\log(T_{ij}) = \zeta_i - \tau_j + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.3.1}$$

in which the RT person *j* spends on item *i* is log-transformed to $\log T_{ij}$; τ_j is the speed parameter for person *j*, indicating the speed with which the person *j* works throughout the assessment. The larger τ_j is, the faster the person worked on the items; ζ_i is the item intensity for item *i*, indicating the labor required for that item; and α_i^{-2} is referred to as the item-specific time discrimination parameter. A smaller value for α_i^{-2} indicates that the item better discriminates the RT distribution produced by people with different levels of speed. Since the raw RT is non-negative and positively skewed, the log-transformation of RT approximates the normal distribution of RT and makes the lognormal model a success in various applications (e.g., Schnipke & Scarms, 1999). Note that a slope parameter can be included to capture the differential effects of the items on person speed (e.g., Fox et al., 2007; Klein Entink et al., 2009).

However, the lognormal model cannot always remove the skewness in the raw RT data. Therefore, the Box-Cox normal model was proposed under such scenarios (Klein Entink et al., 2009):

$$T_{ij}^{(v)} = \zeta_i - \tau_j + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.3.2}$$

where the parameters are interpreted the same as those in the lognormal model, except that $T_{ij}^{(v)}$ indicates the Box-Cox transformation (Box & Cox, 1964) of RTs instead of the log transformation, and the parameters are all on the Box-Cox transformed scale. Specifically, $T_{ij}^{(v)}$ is expressed as:

$$T^{(v)} = \begin{cases} \dfrac{T^v - 1}{v}, v \neq 0 \\ \log T, v = 0 \end{cases}, \tag{2.3.3}$$

where $v \in \mathbb{R}$ is a parameter that determines the degree to which the RT is transformed. Based on the above equation, the Box-Cox model is more flexible and the lognormal model is a special case of it when $v = 0$. However, the major disadvantage of the Box-Cox transformation is that it puts item parameters on different scales and makes parameter interpretation more difficult.

As an alternative to the Box-Cox model, the Cox proportional hazards (PH) model (Ranger & Ortner, 2011; Wang et al., 2013) is shown as:

$$h_{ij}(t_{ij} \mid \tau_j) = h_{0j}(t_{ij}) \exp(\gamma_i \tau_j), \tag{2.3.4}$$

where t_{ij} denotes RT for item *j* and person *i*, h_{0j} is the baseline hazard function, τ_j is the speed parameter for person *j*, and γ_i the slope parameter for item *i* that controls the increase of the hazard rate. Hazard rate in the testing scenario represents how intensively the respondent is working at the moment. In psychological terms, hazard rate indicates the respondent's relative mental processing ability in a unit of time (Ranger & Ortner, 2011; Wenger & Gibson, 2004).

Therefore, γ_i is analogous to the item discrimination parameters in the item response models. The semiparametric nature (i.e., involving both unknown and known probability density functions) of the Cox PH model provides with it two important advantages over the Box Cox model: it allows the usage of various RT distributions and item-specific transformations that preserve the interpretability of the item parameters.

A more generalized and flexible modeling approach is the linear transformation model (Ranger & Kuhn, 2012a, 2013; Wang, et al., 2013):

$$H_i(t_{ij}) = \gamma_i \tau_j + \varepsilon_{ij}, \tag{2.3.5}$$

where τ_j and γ_i are interpreted the same as those in the cox PH model. This model indicates that, after some order-preserving transformation $H_i(t_{ij})$, RTs become linearly related to the item parameters plus Gaussian errors ε_{ij}. It can be algebraically shown that various other models, such as the Box-Cox model and the Cox PH model, are special cases of the linear transformation model (Wang et al., 2013).

Researchers have proposed the usage of other distribution functions to model RT data, including the linear exponential model (Scheiblechner, 1979), the gamma model (Verhelst et al., 1997) and the Weibull model (Roskam, 1997; Rouder et al., 2003) to model RT exclusively. The linear exponential model (Scheiblechner, 1979) is presented as follows:

$$f(RT_{ij}) = (\tau_j + \gamma_i)\exp[-(\tau_j + \gamma_i)RT_{ij}], \tag{2.3.6}$$

where τ_j and γ_i are person and item intensity parameters, respectively. Note that γ_i is interpreted differently in the Box Cox model and the linear transformation model. In the linear exponential model, γ_i represents a linear combination of the item intensities of several item covariates that represent different latent cognitive processes.

A more generalized model to the linear exponential model is the gamma model (Maris, 1993; Verhelst et al., 1997):

$$f(t_{ij}) = \frac{\tau_j^{\lambda_i}}{\Gamma(\lambda_i)} t_{ij}^{\lambda_i - 1} e^{-\tau_j t_{ij}} , \tag{2.3.7}$$

where τ_i and λ_j are interpreted similarly as those in the linear exponential model. The two parameters indicate the precision and speed that respondents perform in a test, altogether forming a measure of mental power. The Weibull model (Roskam, 1997), however, models the test completion time *t* instead of item RTs. Its density function is written as:

$$f(t) = \lambda t \exp(-\frac{\lambda}{2} t^2) , \tag{2.3.8}$$

where $\lambda = \frac{\theta_j}{b_i \delta_j}$, θ_j and δ_j are the mental speed and persistence of the person *j*, b_i is the item difficulty for item *i*. Rouder et al. (2003) further suggested a 3-parameter Weibull distribution that easily captures any shift, scale, and shape of the item-level RT distributions. The usage of the Weibull distribution is beneficial because it naturally accommodates the positive skewness of the RT distributions and has broad applications in statistics (Rouder et al., 2013).

To summarize, various statistical models have been proposed for the modeling of RTs: the lognormal model (van der Linden, 2006), the Box-Cox model (Klein Entink et al., 2009), the Cox PH model (Ranger & Ortner, 2011; Wang et al., 2013), the linear transformation model (Ranger & Kuhn, 2012a, 2013; Wang, et al., 2013), the Weibull model (Roskam, 1997; Rouder et al., 2003), the gamma model (Maris, 1993; Verhelst et al., 1997), and the linear exponential model (Scheiblechner, 1979). Among these RT models, the lognormal model is the most widely applied for several reasons. First, the log-transformed RTs are assumed to have the normality that can be easily adapted to factor analytic models. Second, it has been found that the lognormal

model has adequate model-data fit in various application scenarios (e.g., Schnipke & Scrams, 1999, 2002). However, the other RT models also have their own benefits and values in the analysis of RTs, while they may differ in terms of statistical assumptions and the interpretation of time-related parameters. For a comprehensive review of RT models, readers can refer to Schnipke and Scrams (2002) and Lee and Chen (2011).

2.3.2 Explanatory RT models

Similar to the decomposition modeling approaches for the item responses, there has been a tradition to decompose RTs based on hypothesized problem-solving processes in cognitive psychology (Donders, 1869; Sternberg, 1969). The most notable work was Sternberg (1997a, 1997b) that focused on analogy items and Sternberg (1980, 1986) that focused on deductive reasoning items. For example, Sternberg (1997a) used a multiple linear regression model regressing RT on several item covariates that represent the number of cognitive operations for the items. The model is written as:

$$RT = X_0 + aX_1 + bX_2 + cX_3 + \varepsilon \,, \tag{2.3.9}$$

where X_0 is the intercept, $X_i (i = 1,2,3)$ denotes the item covariates, ε is the Gaussian residual. Using this model, Sternberg (1997a) was able to estimate the RT required for each hypothesized cognitive process for each respondent.

More recently, researchers in the area of educational assessment have developed explanatory RT models that can be put into three categories: models for RT only (Scheiblechner, 1979; Maris, 1993), separate models for RA and RT (Bejar & Yocom, 1991; Embretson, 1998; Gorin, 2005; Mulhoolland et al., 1980; Primi, 2001), and joint models for RA and RT (Klein Entink, et al., 2009; van Breukelen, 2005). Scheiblechner (1979) proposed the linear exponential

model, as mentioned in the previous section, in which the RT is assumed to have an exponential density function:

$$f(RT_{ij}) = (\tau_j + \gamma_i)\exp[-(\tau_j + \gamma_i)RT_{ij}], \tag{2.3.10}$$

where τ_j is person speed parameter and γ_i is the item intensity parameter. Scheiblechner (1979) further suggested that γ_i can be decomposed using the approach analogous to the LLTM:

$$\gamma_i = \sum_{k=1}^{K} q_{ik}\eta_k \ , \tag{2.3.11}$$

where η_k indicates the intensity of the cognitive process *k* embedded in item *i* and q_{ik} indicates whether item *i* is associate with cognitive process *k*. Similarly, Maris (1993) incorporated item covariates in his gamma model for RTs. He showed that the linear exponential model of Scheiblechner (1979) is a special case of one formulation of the gamma model. Nevertheless, these models did not take RA into consideration.

Several studies have been conducted to analyze RA and RT separately in an explanatory way. A shared characteristic of these analyses is that the RT was evaluated based on item covariates. Specifically, Gorin (2005) used the LLTM to analyze item responses and then regressed the log-transformed RTs on the item covariates using reading comprehension items. Similarly, Embretson (1998) and Primi (2001) decomposed RTs using regression models with figural reasoning items. Mulholland et al. (2001), on the other hand, made prediction of RTs using item covariates through analysis of variance for correct and incorrect responses separately. Using a more descriptive approach, Bejar and Yocom (1991) compared the item difficulty parameters and RT distributions of item isomorphs (i.e., items with the same underlying design matrix) from two test forms.

2.4 Joint Modeling of RA and RT

Models reviewed in sections 2.2 and 2.3 are developed for the modeling of RA or RT exclusively. Separate analysis of RA and RT can provide some information on both variables, but the relation between the two variables remains unknown. To explicitly model the relation between RA and RT, researchers have developed methods for the joint modeling of these two variables. As mentioned in section 2.1, the joint distribution of RA and RT may rest on the assumption of CI or accommodates the CD between RA and RT. According to van der Linden and Glas (2010), three CI assumptions exist in the joint modeling of RA and RT: 1) independence between responses; 2) independence between RTs; 3) independence between responses and RTs. The CI assumption discussed in this section mainly refers to the third type. The CD of the two variables, on the other hand, has two possible directions: 1) the RT distribution depends on the item response categories (i.e., correct or incorrect); 2) the RA distribution depends on the RT characteristics (i.e., fast or slow). It is to be noted that the methods of joint modeling of RA and RT are mainly developed using the IRT models for the RA, while several recent extensions have been made using the CDMs for the RA. Therefore, this section reviews methods that jointly model RA and RT, and categorize the studies into three subsections: joint models with IRT models for the RA, joint models with CDMs for the RA, and explanatory joint models of RA and RT. For the joint models with IRT models for the RA, these models are further categorized based on their assumptions: joint models that assume independent RA and RT, joint models that accommodate CD of RA on RT, and joint models that accommodate CD of RT on RA.

2.4.1 IRT models for the RA

2.4.1.1 CI between RA and RT

A straightforward assumption in the joint modeling of RA and RT is that the item response and item RT are locally independent given the constant latent traits. The drift diffusion model (Ratcliff, 1978) is one such approach in the simultaneous analysis of RA and RT. The drift diffusion model assumes a diffusion process where the information accumulates over time and hits a boundary when the decision is made. It is mostly used in the experimental psychology settings where data from only one respondent is collected and analyzed. However, Tuerlinckx and De Boeck (2005) have shown that the drift diffusion model can be extended to incorporate a random effect for the persons and be used in the cross-sectional data consisting of both item responses and item RTs from a sample of respondents. Researchers have made further efforts to adapt the drift diffusion model for the estimation of person ability and item parameters (Molenaar et al., 2015c; van der Maas et al., 2011; Vandekerckhove et al., 2011; Wagenmakers, 2009). In fact, the drift diffusion model has been shown to be related to the 2PL model (Luce, 1986; Tuerlinckx & De Boeck, 2005). Therefore, the drift diffusion model can be applied in educational measurement settings as well (e.g., van Rijn & Ali, 2017).

An alternative modeling approach that assumes CI is the hierarchical modeling framework of RA and RT (van der Linden, 2007), which has been a prominent approach to jointly analyze RA and RT in educational and psychological measurement settings. As its name suggests, the hierarchical model consists of two levels: the first level models specify the probability distribution functions for RA and RT, respectively; the second level models allow the person and item parameters from both RA and RT sides to covary. van der Linden (2007) adopted a three-parameter normal-ogive (3PNO) model for the RA and the lognormal model (van der Linden, 2006) for the RT as the first level models. At the second level, the multivariate

normal distribution is assumed for the person parameters and item parameters, respectively. The item parameters follow a multivariate normal distribution:

$$\mathbf{\Psi}_i = \begin{pmatrix} a_i \\ b_i \\ c_i \\ \zeta_i \end{pmatrix} = N\left(\begin{pmatrix} \mu_a \\ \mu_b \\ \mu_c \\ \mu_\zeta \end{pmatrix}, \mathbf{\Sigma}_{Item}\right), \mathbf{\Sigma}_{Item} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\zeta} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\zeta} \\ \sigma_{ac} & \sigma_{bc} & \sigma_c^2 & \sigma_{c\zeta} \\ \sigma_{a\zeta} & \sigma_{b\zeta} & \sigma_{c\zeta} & \sigma_\zeta^2 \end{pmatrix}, \tag{2.4.1}$$

where *a*, *b* and c denote the item discrimination, difficulty, guessing parameters in the 3PNO model, respectively; ζ denotes the item intensity parameter in the lognormal RT model. The latent ability parameter θ in the 3PNO model and the latent speed parameter τ in the lognormal RT model follow a bivariate normal distribution, which is given by:

$$\mathbf{\Theta}_j = \begin{pmatrix} \theta_j \\ \tau_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \mathbf{\Sigma}_{person}\right), \mathbf{\Sigma}_{person} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{2.4.2}$$

This modeling framework is flexible in that the measurement models for the RA and the RT can be easily modified based on the practical needs.

Some extensions of van der Linden's model (2007) have been made by using alternative RT models under the same hierarchical framework (Klein Entink et al., 2009; Molennaar & Bolsinova, 2017; Ranger & Kuhn, 2014a; Ranger et al., 2015; Wang, et al., 2013; Wang et al., 2013). Specifically, the Box-Cox model (Klein Entink et al., 2009), the Cox PH model (Ranger et al., 2015; Wang et al., 2013), and the linear transformation model (Wang et al., 2013) have been adopted as the RT model in the hierarchical modeling framework. In addition, Ranger et al. (2015) proposed the race model for the RT, which considers the responding mechanism that is determined by the stochastic process for specific response options.

In addition to the specification of different RT models, extensions of the van der Linden (2007) model have been made to accommodate the non-normal distribution of the log-transformed RT and multiple latent dimensions due to multiple sources for RA and RT. Molenaar and Bolsinova (2017) stated that the assumption of the normality of the log-transformed RTs may not hold in reality. Therefore, they modified the RT model to accommodate the non-normality due to either the heteroscedastic residual variances or the non-normal latent speed. Fox et al. (2014) considered the impact of feedback behavior on students' performance. They proposed a multivariate joint model of RA and RT with four latent dimensions to accommodate multiple response sources (i.e., item responses and feedback behavior). Similarly, Zhan et al. (2018) considered the multidimensionality issue in both RA and RT models using the hierarchical modeling framework of van der Linden (2007). Man et al., (2019) considered the compensatory multidimensionality issue in the RA model using the hierarchical modeling framework of van der Linden (2007). These extensions have shown the flexibility of the hierarchical modeling framework in the simultaneous analysis of RA and RT.

Some researchers have advocated the use of the 2PL model for the RA (Fox, 2010, chapter 8; Fox et al., 2007; Molenaar et al., 2015; Molenaar et al., 2015; Ranger & Ortner, 2012; Ranger, 2013) and a different specification of the lognormal model with an additional slope parameter for the latent speed (Fox, 2010, chapter 8; Fox et al., 2007; Klein Entink et al., 2009) in the hierarchical model. The lognormal model with the slope parameter is given by:

$$\log(T_{ij}) = \zeta_i - \varphi_i \tau_j + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_i^2), \tag{2.4.3}$$

where the notations are consistent with the lognormal model (van der Linden, 2006) presented in the equation 2.3.1, except that φ_i is the time discrimination parameter modeled as the slope parameter for the latent speed τ_j and σ_i^2 is the item specific residual variance. This

specification is analogous to the 2PL model for item responses and allows the relation between the expected RT and the latent speed to vary across items. Such model specification makes it possible to extend the hierarchical modeling approach to the generalized linear factor model approach (Molenaar et al., 2015b). Molenaar et al. (2015b) simplified the hierarchical model of RA and RT developed by van der Lindon (2007), Fox et al. (2007), Klein Entink et al. (2009), and Glas and van der Linden (2010) by omitting the second level model on the item side. That is, the item parameters were treated as fixed effects rather than random effects in the model of van der Linden (2007). The benefit of this extension enables the use of flexible latent variable modeling software to estimate the model, although the inclusion of the second level model on the item side should be considered for substantive reasons if necessary.

Built upon the generalized linear factor model approach, Molenaar et al. (2015a) proposed an even more flexible bivariate generalized linear IRT modeling framework for the joint analysis of RA and RT. This modeling framework allows both linear and nonlinear cross relation functions between latent ability and latent speed. In addition, many existing models, including the hierarchical modeling framework (van der Linden, 2007) and some models for RA only or RT only, fall under the bivariate generalized linear IRT modeling framework. According to Molenaar et al. (2015a), for dichotomous responses and continuous RTs:

$$E(Y_{ij}) = \Phi(a_i\theta_j + b_i), \tag{2.4.4}$$

$$E(\log(T_{ij})) = \zeta_i + \varphi_i\tau_j + f(\theta_j;\rho), \tag{2.4.5}$$

where the notations are consistent with equations 2.2.4 and 2.3.1, except that θ_j and τ_j are orthogonal and the relation between ability and speed is captured by the cross-relation function $f(\theta_j;\rho)$. Note that the sign before $\varphi_i\tau_j$ changes from negative to positive and, therefore, the

interpretation of the τ_j now indicates slowness rather than speed or fastness. The cross-relation function indicates relations that can be linear, quadratic, interaction or other types of relations. The model of van der Linden (2007) assumes a linear relation between latent ability and speed and the expectation of the log-transformed RT is given by:

$$E(\log(T_{ij})) = \zeta_i + \varphi_i \tau_j - \varphi_i \rho_1 \theta_j . \tag{2.4.6}$$

Let $\tau' = \tau_j - \rho_1 \theta_j$ represent the latent speed parameter as specified in van der Linden (2007). Then, it can be seen that ρ_1 indicates the correlation between ability and speed in the model of van der Linden (2007). Additionally, several models for RT only (Ferrando & Lorenzo-Seva, 2007a; 2007b; Gaviria, 2005; Ranger & Kuhn, 2012a; Thissen, 1983) and models for RA only (Roskam, 1987; Wang & Hanson, 2005) fall under the modeling framework of Molenaar et al. (2015a). For example, the model of Thissen (1983) also assumes a linear relation between ability and speed. The expectation of the log-transformed RT is given by:

$$E(\log(T_{ij})) = \mu + \zeta_i + \tau_j - \rho_1 (a_i \theta_j - b_i) , \tag{2.4.7}$$

where μ is a general intercept; a_i and b_i are item discrimination and difficulty in the IRT model, respectively; ρ_1 indicates the regression coefficient between RT and latent ability. The terms μ and $\rho_1 b_i$ can be absorbed by ζ_i and the equation above can be rewritten as:

$$E(\log(T_{ij})) = \zeta_i + \tau_j - a_i \rho_1 \theta_j . \tag{2.4.8}$$

Similarly, ρ_1 in the above equation captures the linear relation between RT and latent ability. In addition, Ferrando & Lorenzo-Seva (2007a; 2007b) used the absolute value $\left| a_i \theta_j - b_i \right|$ in the equation 2.4.7 and, thus, specified a nonlinear regression of IRT parameters on RTs. Given that

the current study focuses on the joint modeling of RA and RT, the specifications of the other models for RT only or RA only can be found in Molenaar et al. (2015a).

In summary, the drift diffusion model (Ratcliff, 1978), the model of van der Linden (2007), and their variants are fundamental modeling frameworks that rest on the CI assumption of RA and RT. That is, given the constant latent ability and latent speed, the item responses and item RTs are identically and independently distributed. This assumption makes the statistical modeling straightforward and convenient. However, this assumption may not always hold in real-world settings. For example, the speed level and ability level of a respondent may fluctuate during the test due to various reasons such as strict time limit (Bolsinova et al., 2017). Therefore, the violation of the CI assumption may occur and jeopardize the statistical inference of the focal parameters. To evaluate the impact of the within-person variations on model misfit, methods have been developed to assess person fit (Fox & Marianti, 2017; Marianti et al., 2014), model fit (Marianti, 2015; Ranger & Kuhn, 2014b; Ranger et al., 2017), and the CI assumption (van der Linden & Glas, 2010; Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016). The studies reviewed in the next two sections 2.4.1.2 and 2.4.1.3 focus on the modeling approaches that accommodate CD in the joint modeling of RA and RT. Specifically, section 2.4.1.2 focuses on models that accommodate CD of RA on RT, while section 2.4.1.3 focuses on models that accommodate CD of RT on RA.

2.4.1.2 CD of RA on RT

Conceptually, the model of van der Linden (2007) can be conceived as a simple structure hierarchical model that assumes the residuals of RA and RT are uncorrelated. However, as mentioned in section 2.1, when CD occurs, the joint distribution of RA and RT can be factored as

$$f(t_i \mid \tau; \beta_i) f(y_i \mid t_i, \theta; \alpha_i), \tag{2.4.9}$$

that implies that the measurement model for RA may depend on RT even when the latent ability and the latent speed are correlated. In other words, the residual RT from either a faster or slower response can be related to the measurement model for the RA in some ways. Thus, it is crucial to accommodate such CD to improve the model for RA (e.g., Bolsinova et al., 2017).

To examine whether fast or slow responses indicate different response processes, one straightforward way is using the IRTree modeling approach (De Boeck & Partchev, 2012). IRTree models are IRT models for item response data with a tree structure (De Boeck & Partchev, 2012). In the case of joint modeling of RA and RT, the tree structure can be formed by splitting the item RTs into fast and slow categories using median split within persons or within items (De Boeck & Partchev, 2012; DiTrapani et al., 2016; Partchev & De Boeck, 2012). Common IRT models (e.g., 1PL and 2PL) can be used for these item responses, but different latent variables and item parameters can be assumed for fast responses (fast and correct, or fast and incorrect) and slow responses (slow and correct, or slow and incorrect), respectively. Although the estimation and interpretation of these models are convenient, the drawbacks of the IRTree modeling approach include: 1) the assignment of item responses into the fast or slow categories is deterministic; 2) dichotomization of item RTs lead to information loss in the estimation of latent ability (Molennar & De Boeck, 2018).

An appealing alternative approach is to use the mixture models that allow different latent classes with various person and item properties for different speed (Marianti, 2015, chapter 3; Molenaar et al., 2016; Molenaar et al., 2016; Molenaar, et al., 2016; Molenaar & De Boeck, 2018; Molenaar et al., 2019; Wang & Xu, 2015; Wang et al., 2018). Molenaar and De Boeck (2018) proposed the response mixture model that used the log-RTs as predictors to identify the

mixing proportions. Molenaar et al. (2016) further extended the response mixture model in two ways: 1) incorporate two latent variables for faster and slower responses, respectively; 2) allow the latent class sizes to vary across items. Different from traditional mixture modeling approaches, the response mixture model classifies each item response rather than each response vector into fast or slow classes. While the response mixture model incorporated the mixture component in the RA model only, Wang and Xu (2015) incorporated mixture components for both RA and RT in the scenario where rapid guessing occurs. The model of Wang and Xu (2015) specified different IRT model and RT model for the solution behavior and rapid guessing behavior. Therefore, different variabilities in the RT are modeled for different response processes (i.e. heteroscedasticity). Wang et al. (2018) further developed a two-stage procedure to flag aberrant behavior using a Bayesian residual index in addition to the mixture modeling. Similarly, Marianti (2015, chapter 3) proposed a mixture model that differentiates stationary and non-stationary speed over blocks of items. However, like traditional mixture models, latent variables are assumed to be independent in these models (Marianti, 2015, chapter 3; Wang & Xu, 2015; Wang et al., 2018), which may not always hold (Molenaar et al., 2016). Thus, Molenaar et al. (2016) proposed the hidden markov IRT models for RA and RT that allow transitioning latent variables to be dependent across items. In addition, the parametric assumption of the RT distribution can be too stringent. Thus, Molenaar et al. (2016) proposed the semi-parametric mixture model for RA and RT using categorized RTs to relax the parametric assumption of the RT. Molenaar et al. (2019) further proposed the heteroscedastic hidden Markov mixture model that accommodate the heteroscedastic variances of RT in different response processes, dependence of latent states between subsequent items, and the semi-parametric distribution of the RT at the same time.

Both the IRTree models and mixture models treat the CD of RA on RT as discrete. However, researchers have argued that the impact of RT on the RA model can be continuous (e.g., Bolsinova et al., 2017; Fox & Marianti, 2016). Fox and Marianti (2016) and Marianti (2015, chapter 4) adopted the latent growth modeling techniques to reparameterize the latent speed parameter to relax the constant speed assumption in the model of van der Linden (2017). Apart from this, most of the studies incorporated observed RTs as predictors in the RA model to capture the CD of RA on RT. The generalized linear mixed modeling framework was used to incorporate both fixed and random effects of the log-transformed RTs on RA (Goldhammer et al., 2014; Goldhammer & Kroehne, 2014; Goldhammer et al., 2015). Goldhammer et al. (2017) further extended the previous approach by decomposing the log-transformed RTs as person average, item average and residual RTs as predictors of the RA model to examine RT effects in both within- and between- subject conditions. Similarly, De Boeck et al. (2017) incorporated the log-RTs in the RA model to explore the RT effects under spontaneous and imposed speed conditions. Rather than using the lognormal RT model (van der Linden, 2006), Wang (2006) proposed a joint model using the one-parameter Weibull model for the RT and specified the RA model conditional on RT. Building upon Wang (2006), Ingrisone (2008) specified the marginal distribution of RT using the two parameter Weibull model, while the conditional distribution of RA as the 1PL model conditional on RT.

Although incorporating the observed RTs in the RA model is straightforward, residual RTs have been advocated in the modeling of conditional distribution of RA (e.g., Bolsinova et al., 2017). Several studies have extended the hierarchical model by incorporating different forms of the RT residual dependence on the RA model. Ranger and Ortner (2012) allowed item-specific correlation parameters between RA and RT that capture any dependence that cannot be

explained by the correlation between latent ability and speed. Although it is straightforward, this approach does not explicitly explain how the item response function changes as a function of the RT residual and only allows the RT residual to be associated with the item intercept parameter of the item response function. Thus, Bolsinova et al. (2017) proposed a hierarchical model that allows both the item slope and item intercept parameters in the 2PL model (i.e., the measurement model for RA) to depend on the standardized RT residuals. Both Ranger and Ortner (2012) and Bolsinova et al. (2017) took into account the item-specific effects of the residual RT on RA, but none of them considered the possible CD from the differences between persons. Meng et al. (2015) specified the product of a person-specific component and an item-specific component for the residual correlation. However, the item-specific component was constrained to be non-negative, which is contradictory to existing findings that the signs of the CD can be either positive or negative (e.g., Bolsinova et al., 2016). To develop a more flexible and general framework for the modeling of the CD, Bolsinova et al. (2016) proposed the response moderation models that consider the CD differences between items and between persons and allow differences in the sign and magnitude of the CD. Further, Bolsinova and Molenaar (2018) extended the response moderation models by incorporating quadratic effects of the residual RT on RA.

In conclusion, studies have modeled the CD of RA on RT from either a discrete point of view using the IRTree modeling approach and mixture modeling or from a continuous point of view incorporating RTs or the residual RTs in the RA model. It has been found that these modeling approaches have improved the model-data fit (Bolsinova et al., 2017; Molenaar et al., 2016). In addition, these studies revealed either negative or positive CD between RA and RT (e.g., Bolsinova et al., 2017).

2.4.1.3 CD of RT on RA

As mentioned in section 2.1, the joint distribution of RA and RT can also be factored as

$$f(y_i \mid \theta; \alpha_i) f(t_i \mid y_i, \tau; \beta_i), \tag{2.4.10}$$

which implies that the measurement model for RT may depend on RA. That is, different RT distributions may exist for correct and incorrect responses, respectively.

Three statistical testing procedures have been developed to examine the CD of RT on RA: the Lagrange Multiplier (LM) test (Glas & van der Linden, 2010; van der Linden & Glas, 2010); the non-parametric Kolmogorov-Smirnov (KS) test (Bolsinova & Maris, 2016); and the posterior predictive checks (PPCs; Bolsinova & Tijmstra, 2016). van der Linden and Glas (2010) and Glas and van der Linden (2010) evaluated the CI assumption under the hierarchical modeling framework (van der Linden, 2007). They specified the RT model under CI as the lognormal model (van der Linden, 2006):

$$\log(T_{ij}) = \zeta_i - \tau_j + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.4.11}$$

where the notations are the same as in the equation 2.3.1. This model is considered as the null hypothesis and tested against a parametric alternative:

$$\log(T_{ij}) = \zeta_i - \tau_j + \lambda_i y_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.4.12}$$

where y_{ij} denotes the correct or incorrect item responses; λ_i is an item specific location shift parameter that captures the difference of location of the RT distribution for the correct or incorrect responses. The limitation of the LM test used in this scenario is that it can only assess possible location shift of the RT distribution but cannot assess other potential types of CD. In addition, the LM test requires that the parametric distribution of RT is correctly specified. Bolsinova and Maris (2016) proposed a non-parametric KS test approach that does not require the specification of a certain type of CD. However, it requires that the RA model is an

exponential family model and a large sample to achieve adequate power (Bolsinova & Tijmstra, 2016). To overcome the limitations of the LM test and the KS test, Bolsinova and Tijmstra (2016) proposed the PPC procedure using three discrepancy measures that indicate different types of CD. Their results showed that the PPC procedure is a flexible and robust method in detecting CD.

The CI assumption between RA and RT indicates that no cross-loadings are allowed in the hierarchical model of van der Linden (2007). Bolsinova and Tijmstra (2017) relaxed this assumption by allowing cross-loadings between latent ability and the log-transformed RTs:

$$\log(T_{ij}) = \zeta_i - \varphi_i \tau_j + \lambda_i \theta_j + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_i^2), \tag{2.4.13}$$

where λ_i denotes the linear effect of ability on the log-RT of item *i* and other notations are consistent with those in the equation 2.4.3. They adopted the orthogonal rotation of the factors for the model identifiability purpose. They stated that this extended hierarchical model improves the estimation of ability and also provides an extra measure of item performance based on the cross-loading estimates. Similarly, Magnus et al. (2017) used a hierarchical model with such a structure on an empirical dataset and also found that the inclusion of RT improved the measurement precision of RA.

Apart from allowing cross-loadings between RA and RT, Bolsinova and Tijmstra (2019) proposed an extended hierarchical model that account for the difference between correct and incorrect responses directly. Specifically, the conditional distribution of log-RTs is specified as:

$$\log(T_i) = \zeta_{iy_i} - \varphi_{iy_i} \tau_{y_i} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_{iy_i}^2), \tag{2.4.14}$$

where $y_i = 1$ denotes the correct item response and $y_i = 0$ denotes the incorrect item response for item *i*; two sets of the item intensity parameter, latent speed parameter and residual variance

parameter exist or the correct and incorrect responses, respectively: $\left\{\zeta_{i1}, \varphi_{i1}, \tau_1, \sigma_{i1}^2\right\}$ and $\left\{\zeta_{i0}, \varphi_{i0}, \tau_0, \sigma_{i0}^2\right\}$. Then, three latent variables: latent ability, latent speed for correct responses, latent speed for incorrect responses are assumed to follow the multivariate normal distribution. Their results from the empirical data application suggested that difference in the RT models between correct and incorrect responses existed and the two latent speed dimensions were strongly correlated but bear different interpretations.

Based on the empirical evidence from the existing studies, the CD of RT on RA is prevalent. Bolsinova and Maris (2016) found that the CI assumption was violated for a majority of items in an arithmetic test using the KS test. van der Linden and Glas (2010) also found that the CI assumption was violated for more than half of the items from a multistage test using the LM test. Bolsinova and Tijmstra (2016) found that the CI assumption was violated for 22, 32, and 8 items out of 38 items based on three discrepancy measure using the PPC procedure, respectively. Therefore, it is important to account for the CD of RT on RA in practical settings.

2.4.2 CDMs for the RA

RT has been used as collateral information in the CDMs to improve measurement precision (e.g., Minchen, 2017; Zhan et al., 2018) or to shed light on respondents' learning behaviors (e.g., Wang et al., 2018). Similar to the studies reviewed in the previous section using IRT models for RA, the joint modeling of RA and RT using CDMs also assumed either CI or CD between RA and RT. However, most of the studies that use CDMs for RA focused on issues related to learning due to the nature of the diagnostic assessments where CDMs are usually applied. For example, the studies modeled the change of respondents' latent attribute pattern over time and the fluency of applying the latent attributes using the joint models (e.g., Wang et al., 2018; Wang et al., 2019; Wang & Chen, 2020; Wang et al., 2020; Zhang & Wang, 2018).

Several hierarchical models using the CDMs for item responses assuming CI between RA and RT have been developed (Minchen, 2017; Zhan et al., 2018). Both Minchen (2017) and Zhan et al. (2018) made a direct adaptation of the model of van der Linden (2007) by using the higher-order DINA model (de la Torre & Douglas, 2004) as the measurement model for RA. However, Minchen (2017) omitted the correlations among the items, while Zhan et al. (2018) imposed both correlational structures on the item side and the person side. Further, Zhan et al. (2018) considered the local dependence for RA and RT and incorporated testlet structures in both the RA model and the RT model. The shared purpose of these models is to improve measurement precision of RA.

Also assuming CI between RA and RT, Wang et al. (2018) proposed a joint model that allows the change of latent attribute pattern overtime and the growth of latent speed due to the change of the latent attribute pattern and the increase of respondents' fluency in applying mastered attributes. Similar to the hierarchical model (van der Linden, 2007; Zhan et al., 2018), measurement models for RA and RT are specified, respectively. The measurement model for RA is the higher-order Hidden Markov CDM (Wang et al., 2018) that consists of the DINA model and a transition model that describes how respondents' latent attribute profile change over time. The measurement model for RT is the lognormal model (van der Linden, 2006) with a latent covariate specified as a function of the latent attributes:

$$\log(T_{ij}) = \zeta_i - \tau_j - \phi G_{ij}\,(\boldsymbol{\alpha}_j) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i^2), \tag{2.4.15}$$

where the $G(\cdot)$ is a function of the attribute trajectory of respondent j; the fixed slope parameter ϕ captures the rate of the latent attribute change; other notations are consistent with equation 2.4.3. Different from the model of Zhan et al. (2017), the two measurement models in Wang et

al. (2018) are related through the latent covariate $G(\cdot)$ in the RT model rather than the correlation between latent ability and latent speed.

Several extensions have been made based on Wang et al. (2018) retaining the CI assumption. Wang et al. (2019) incorporated observed and latent covariates on the persons such as practice and intervention effects in the transition model of the latency change. This model with person covariates was applied in the development of a multidimensional diagnostic assessment with learning tools (Wang et al., 2020). Other extensions, however, considered CD between RA and RT. Zhang and Wang (2018) used the mixture model to capture the heterogeneity of the learning processes in the RA and RT based on the model of Wang et al. (2018). This mixture modeling approach is similar to the models reviewed in 2.4.1.2 (e.g., Molenaar et al. 2016). Wang and Chen (2020) proposed a joint model in which the RT model is conditional on correct or incorrect responses, which is related to the models reviewed in 2.4.1.3 (e.g., Bolsinova & Tijmstra, 2019). Specifically, the RT information from correct responses is used to further model fluency using a similar model formulation as in the equation 2.4.15, while the RT distribution for the incorrect responses is the lognormal model (van der Linden, 2006).

2.4.3 Explanatory Joint Modeling of RA and RT

Several explanatory joint models of RA and RT have developed by incorporating person and/or item covariates in the RA model and the RT model, respectively (Klein Entink, Kuhn, Hornke, & Fox, 2009; Klein Entink, Fox, & van der Linden, 2009; Loeys et al., 2011; van Breukelen, 2005). van Breukelen (2005) proposed a bivariate mixed logistic regression model for the log-transformed RTs and the logit of the probability of correct responses. In this model, persons were treated as random, while items fixed. Specifically, log-transformed RTs and the log-odds of correct responses were decomposed in the LLTM fashion, which is specified as:

$$\text{logit}(p_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + ... + \beta_{pj}X_{pij}, \quad (2.4.16)$$

$$\log(RT_{ij}) = \gamma_{0j} + \gamma_{1j}X_{1ij} + \gamma_{2j}X_{2ij} + ... + \gamma_{pj}X_{pij} + e_{ij}, \quad (2.4.17)$$

where p_{ij} is the probability of obtaining a correct response from respondent *j* to item; X_1 through X_p indicate either person or item covariates; β_0 through β_p and γ_0 through γ_p indicate random person effects for item responses and item RTs, respectively. The disadvantage of the model of van Breukelen (2005) is that no parameters were specified to reflect the item intensity and the person speed. Loeys et al. (2011) also adopted the mixed logistic regression modeling technique and incorporated both person and item covariates in the RA and the RT model. However, they treated both person and item as random effects that indicated the person speed variation and the item intensity variation:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2i} + \theta_{1j} + \tau_{1i} \quad (2.4.18)$$

$$T_{ij} = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2i} + \theta_{2j} + \tau_{2i} + e_{ij}, \quad (2.4.19)$$

where θ_{1j} denotes person ability; τ_{1i} denotes item difficulty; θ_{2j} denotes person speed; τ_{2i} denotes item intensity; β_0 through β_p and γ_0 through γ_p indicate fixed effects for item responses and item RTs, respectively; X_{1j} and X_{1i} denote a person covariate and an item covariate, respectively. Loeys et al. (2011) allowed the RT to follow either the lognormal distribution or the Weibull distribution. In addition, their model followed the hierarchical model of van der Linden (2007) to allow both person and item parameters from the RA model and the RT model to be correlated. Instead of using the mixed regression modeling framework, Klein Entink, Fox, and van der Linden (2009) decomposed the item parameter in the hierarchical model of van der Linden (2007) using item covariates (e.g., test design factors) in the LLTM

fashion. Further, Klein Entink, Kuhn, Hornke, and Fox (2009) extended the model of van der Linden (2007) to a multilevel multivariate model and incorporated person covariates (e.g., gender, citizenship, test effort, SAT scores, and age) to explain the variance in speed and accuracy between persons and groups.

2.5 Model Estimation of the Joint Models

The marginal maximum likelihood (MML; Bock & Aitkin, 1981) and the Bayesian Markov Chain Monte Carlo (MCMC) method are two major estimation methods for the joint models of RA and RT. These two estimation methods differ in their assumptions and optimization methods in obtaining the parameter estimates. The MML is a frequentist method that assumes the model parameters as fixed values. The parameter estimates are obtained by finding the parameter values that maximize the likelihood function. The Expectation-Maximization (EM) algorithm is commonly used to find the optimal parameter values (e.g., Molenaar & Bolsinova, 2017; Ranger & Ortner, 2012). The benefit of the MML is its efficiency in terms of both statistical precision and computation time. The Bayesian MCMC approach, on the other hand, assumes the model parameters are random variables. Specifically, the goal of Bayesian estimation is to obtain the posterior distributions of the model parameters based on the prior beliefs and the likelihood from the data. Usually, the posterior mean and standard deviation of the posterior distribution are taken as the parameter estimates and standard error estimates, respectively (e.g., Zhan et al., 2018).

The choice between the MML and Bayesian MCMC estimation methods depends on several factors. First, MML was favored in the generalized linear modeling approach where item parameters were treated as fixed effects (e.g., Molenaar et al., 2015; Molenaar & Bolsinova, 2017). One benefit of using MML is that currently available software for latent variable models

can be readily used by the practitioners. However, the Bayesian MCMC estimation was usually used when the item parameters were treated as random effects (e.g., van der Linden, 2007). Second, MML becomes infeasible when the number of latent dimensions exceeds five (Wood et al., 2002). This is when the Bayesian approach becomes an alternative. Therefore, studies that investigated the multidimensionality issues in the joint model of RA and RT adopted the Bayesian MCMC estimation (e.g., Zhan et al., 2018). Lastly, the Bayesian MCMC estimation was preferred when the distribution of the variables was assumed to be complex (e.g., the Box-Cox model of the RT; Klein Entink et al., 2009). The Gibbs sampler (Gelman & Gelman, 1984) is the most commonly used MCMC algorithm (e.g., Bolsinova et al., 2017; Fox et al., 2007). However, when there was no conjugate prior, the Metropolis-Hastings (MH; Chib & Greenberg, 1998) was used in Klein Entink et al. (2009).

Most studies reviewed in this chapter used either of the two estimation methods. The software Mplus (Muthén & Muthén, 2007) and the latentGOLD (Vermunt & Magidson, 2013) have been used for the MML estimation. JAGS (Plummer, 2015), WinBUGS (Spiegelhalter et al., 2003), Stan (Gelman et al., 2015) and OpenBUGS (Thomas et al., 2006) have been used for the Bayesian MCMC estimation.

2.6 Summary of Literature Review

This chapter reviewed studies on CDMs, explanatory item response models, (explanatory) RT models, and (explanatory) joint models of RA and RT. Existing studies have explored ways to incorporate item covariates in the estimation of item parameters in CDMs, RT models, and joint models of RA and RT. However, two gaps are identified in the literature. First, current joint modeling approaches of RA and RT that incorporated item covariates have focused on IRT models. Specifically, most of the explanatory models focused on the item difficulty

parameter in the Rasch model (e.g., De Boeck & Wilson, 2008), while most of the explanatory joint models of RA and RT used the mixed regression approach or the IRT-based approach. In fact, the explanatory modeling approaches are originated in the IRT framework. However, these models do not consider the item characteristics (e.g., slipping and guessing) in the diagnostic assessments. In addition, IRT models do not provide fine-grained diagnostic classification information on the respondents.

Second, existing explanatory CDMs do not incorporate RT information. Given that computer-based assessments are prevalent nowadays, RT information is important in the test development (e.g., Embretson & Yang, 2007). Without the joint modeling of RA and RT, certain research questions cannot be asked. For example, the average time required by an item in the population is hard to know. This piece of information could be necessary in the test design of tests with time limits. In addition, the relationship between time-related item parameters and CDM item parameters cannot be revealed. It is hypothesized that items that are harder to guess may also take longer time to complete. However, such hypothesis cannot be tested if the RT information is not incorporated in the psychometric modeling.

Given the limitations in the literature, the current study proposes the explanatory CDMs incorporating RTs that makes two contributions. First, the proposed model can provide guidance for the item writers in the development of diagnostic assessments. By incorporating the item covariates that characterize specific item features, the proposed model can estimate the effects of these covariates. Therefore, based on the parameter estimates, item writers can control item characteristics such as slipping or guessing probabilities. They can also know what item features may make the items more time intensive. Then, specific instructions and guidance can be provided during the item writing procedure.

Second, the current study advances the state of knowledge of explanatory psychometric models by combining explanatory DCMs and explanatory RT models. First of all, the proposed model utilizes both item responses and RTs and accounts for the item covariate effects for cognitive diagnosis. In addition, it shows the performance and potential usage of the proposed model through the investigation of parameter recovery of the proposed model under different simulated testing scenarios. Lastly, the performance of the proposed model is compared with several competing models to provide guidance to researchers in the selection of appropriate models in practical settings.

Chapter 3: Methodology

3.1 The Proposed Model

The current study proposes the explanatory CDM incorporating RTs, which extends the hierarchical cognitive diagnostic modeling incorporating RTs (Zhan et al., 2018) by introducing item covariates into the item parameters. In this section, the first level measurement models for RA and RT are introduced. Then, the higher level population models for the person and item parameters are presented. In particular, the approaches of incorporating item covariates into this hierarchical framework are elaborated. Specifically, reparameterized DINA model (e.g., DeCarlo, 2011) is adopted as the measurement model for item responses. The DINA model is chosen due to its flexibility and its wide application in the field. The lognormal model (van der Linden, 2006) is chosen as the measurement model for RTs for several reasons. First, it has been shown that the lognormal model provides satisfactory model-data fit in various applications (e.g., Schnipke & Scrams, 1999). Second, the normality assumption of the log-transformed RTs is a major advantage over the other RT models especially in the simultaneous analysis of item responses and RTs (Klein Entink et al., 2009). In fact, the normality assumption makes the joint distribution of the latent response and RT variables a bivariate normal one, which makes the hierarchical model specification feasible. Third, the statistical inference is simplified given that the statistical properties of multivariate normal distribution are well known (Anderson, 1984).

3.1.1 Level 1 CDM

Denote Y_{ij} as the binary item response from respondent j (j = 1, 2, …, J) to item i (i = 1, 2, …, I). In the DINA model, the probability of obtaining a correct answer ($Y_{ij} = 1$) is given by:

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_j) = g_i + (1 - s_i - g_i)\prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}, \tag{3.1.1}$$

where s_i and g_i are slipping and guessing parameters that characterize item-level aberrant responses, respectively; $\boldsymbol{\alpha}_j = (\alpha_{j1},...,\alpha_{jK})'$ denotes the attribute pattern for respondent *j* given all *K* attributes measured; *Q*-matrix (Tatsuoka, 1983) is an *I*-by-*K* confirmatory matrix that specifies the relationship between items and binary latent attributes. Specifically, $q_{ik} = 1$ indicates attribute *k* is required to correctly answer item *i* and $q_{ik} = 0$ otherwise. Furthermore, it is assumed that the item probabilities in the DINA model are subject to the order constraint (i.e., $1 - s_i > g_i$) such that a respondent who has mastered all the attributes always has a higher change to answer the item *i* correctly than a respondent who lacks at least one of the attributes (Rupp et al., 2010).

However, the item parameters in the DINA model are on the probability scale which makes it inconvenient to specify the higher level variance and covariance structure among the item parameters in the hierarchical model (e.g., van der Linden, 2007) and incorporating item covariates in the current study. Therefore, the logit scale parameterization for the DINA model (DeCarlo, 2011; Henson et al., 2009; von Davier, 2014) is used for the item responses. This model can be considered as a special case of the G-DINA model using a logit link (de la Torre & Douglas, 2011). Mathematically, the logit of the probability of obtaining a correct response ($Y_{ij} = 1$) under such parameterization is given as:

$$\text{logit}\, P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_j) = \beta_i + \delta_i \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}, \tag{3.1.2}$$

where β_i and δ_i denote the item intercept and item interaction parameters, respectively. The item parameters (i.e., β_i and δ_i) in this logit scale parameterization are related to the slipping and guessing parameters (i.e., s_i and g_i) in the DINA model in the following ways:

$$\beta_i = \text{logit}\,(g_i), \tag{3.1.3}$$

$$\delta_i = \text{logit}\,(1 - s_i) - \text{logit}(g_i). \tag{3.1.4}$$

Furthermore, it is reasonable to form the higher-order latent structural model to account for the correlation among latent attributes (de la Torre & Douglas, 2004; Templin, 2004):

$$\text{logit}\,(P(\alpha_{jk} = 1 \mid \theta_j)) = \gamma_k \theta_j + \lambda_k, \tag{3.1.5}$$

where θ_j is the person-specific general latent ability; $P(\alpha_{jk} = 1 \mid \theta_j)$ denotes the probability of mastering attribute *k* by respondent *j* given ability θ_j; γ_k and λ_k are attribute-specific slope and intercept parameters. The slopes are assumed to be positive given that higher ability leads to higher probability of mastering an attribute. There are three benefits from incorporating the higher-order structure: 1) taking in to account the correlations among the attributes; 2) allowing the estimation of continuous general latent abilities in addition to categorical latent attribute profiles, which can further facilitate the formulation of the variance and covariance structure between person parameters in the hierarchical model; 3) reducing the number of structural parameters associated with attributes from $2^K - 1$ to $2K$ (Zhan et al., 2018).

3.1.2 Level 1 RT Model

The log-transformed RTs (i.e., item response times taken natural logarithm), denoted as $\log(T_{ij})$, are modeled by the lognormal model (van der Linden, 2006), which is essentially a linear model specified as:

$$\log(T_{ij}) = \zeta_i - \tau_j + \varepsilon_{ij}, \tag{3.1.6}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon i}^{\ 2}). \tag{3.1.7}$$

Parameters ζ_i and τ_j denote the intensity for item *i* and the latent speed for respondent *j*, respectively; ε_{ij} is the residual term. The latent speed τ_j, analogous to the general latent ability θ_j, is taken as the latent construct for the RTs that captures the heterogeniety of the work speed among the respondents. This model assumes that 1) the respondents have a constant speed during the test; 2) the RTs to a set of items are locally independent given the constant speed level. This local independence assumption is also analogous to that of the measurement model for item responses. In addition, the item intensity ζ_i indicates that items may vary in terms of the time required to reach the solutions. For example, items with longer stimulus or items that require more cognitive steps may require more time to solve. Finally, the expectation of the $\log(T_{ij})$ equals $\zeta_i - \tau_j$. Note that it is possible to include a discrimination parameter in the lognormal model (e.g., Fox et al., 2007; Klein Entink et al., 2009) to control the decrease in $E(\log(T_{ij}))$ as the speed τ_j increases. However, this study does not include the discrimination parameter for the current exploration.

3.1.3 Level 2 Model for the Person Parameters

Following the hierarchical approach (van der Linden, 2007; Zhan et al., 2018), the latent ability θ_j and the latent speed τ_j from the level 2 models are further assumed to follow a bivaraite normal distribution in the population:

$$\mathbf{\Theta}_j = \begin{pmatrix} \theta_j \\ \tau_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \mathbf{\Sigma}_{person} \right), \mathbf{\Sigma}_{person} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{3.1.8}$$

It is assumed that the item responses and RTs are locally independent conditional on constant ability and speed levels. That is, if a respondent works with constant speed and ability, all

variations within an item are assumed to be captured by these parameters. This local independence assumption is analogous to those in the level 1 measurement model. The covariance $\sigma_{\theta\tau}$ allows us to examine the relationship between the two person parameters. For example, a negative $\sigma_{\theta\tau}$ may indicate that a respondent with higher ability may spend longer time to complete the items.

3.1.4 Level 2 Model for the Item Parameters

The same hierarchical approach is applied to the item side of the model. It is assumed that the items are sampled from an item domain, just as persons are sampled from the population (van der Linden, 2007). Therefore, the item parameters are treated as random effects just as the person parameters.

The current study aims at predicting the item parameters using item covariates that are related to test design elements or item stimulus features. The item covariates can be either categorical (e.g., test design factors) or continuous (e.g., number of words in the item stem). Given that the item parameters are treated as random effects in the current study, the item parameters are decomposed using the LLTM + ε approach (De Boeck, 2008). In the true data generating model, the item parameters are assumed to be a linear function of both a continuous and a dichotomous item covariate. Take the item intercept parameter β_i for example:

$$\beta_i = \mathrm{A}_{0(\beta)} + \mathrm{A}_{1(\beta)} C_i + \mathrm{A}_{2(\beta)} D_i + \varepsilon_{\beta i},$$

where C_i and D_i denote the continuous and dichotomous item covariates, respectively. Let $\mathbf{A} = (\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2)$ and $\mathrm{B}_i = (1, C_i, D_i)'$, then the model can be presented in the matrix form:

$$\beta_i = \mathbf{A}_{(\beta)} \mathbf{B}_i + \varepsilon_{\beta i},$$

$$\delta_i = \mathbf{A}_{(\delta)} \mathbf{B}_i + \varepsilon_{\delta i},$$

$$\zeta_i = \mathbf{A}_{(\zeta)}\mathbf{B}_i + \varepsilon_{\zeta i}.$$

The vector **A** contains the parameters of interest: regression coefficients of the item covariates. The intercept term indicates the base level of the item parameter and the regression coefficients indicate the effects from the item covariates (e.g., Liao & Jiao, 2018). The interpretation of these parameters are analogous to those in a linear regression model. The error terms are assumed to follow a multivariate normal distribution, $\varepsilon \sim N(0, \mathbf{\Sigma}_{Item\varepsilon})$:

$$\begin{pmatrix} \varepsilon_{\beta i} \\ \varepsilon_{\delta i} \\ \varepsilon_{\zeta i} \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma}_{Item\varepsilon} \right), \mathbf{\Sigma}_{Item\varepsilon} = \begin{pmatrix} \sigma^2_{\varepsilon_\beta} & \sigma_{\varepsilon_{\beta\delta}} & \sigma_{\varepsilon_{\beta\zeta}} \\ \sigma_{\varepsilon_{\beta\delta}} & \sigma^2_{\varepsilon_\delta} & \sigma_{\varepsilon_{\delta\zeta}} \\ \sigma_{\varepsilon_{\beta\zeta}} & \sigma_{\varepsilon_{\delta\zeta}} & \sigma^2_{\varepsilon_\zeta} \end{pmatrix}. \tag{3.1.9}$$

The proposed model is referred to as the JRT-DINA-LLTM.

There are two advantages of the proposed model over the hierarchical model (referred to as the JRT-DINA model) proposed by Zhan et al. (2018): 1) the proposed model allows the evaluation of cognitive theories because we can test the effects of the item covariates that are usually related to hypothesized cognitive processes; 2) the scored item covariates can be used to predict the item parameters of the new items in item generation and provide guidance for item writers.

3.1.5 Model Identification and Constraints

For the scale identifiability purposes, constraints are set as $\mu_\theta = \mu_\tau = 0$ and $\sigma^2_\theta = 1$. In addition, the attribute-specific slope parameter is restricted to be $\gamma_k > 0$. It is reasonable to assume that higher latent ability θ_j is associated with a higher probability of mastering a latent attribute α_k, although this may not be true if one of the latent attributes refers to misconception rather than skill (see Bradshaw & Templin, 2014).

Furthermore, it is suggested to set item interaction parameter $\delta_i > 0$ to ensure the order constraint (i.e., $1 - s_i > g_i$) in the DINA model (e.g., Culpepper, 2015; DeCarlo, 2012; Henson et al., 2009; Junker & Sijtsma, 2001). This is implemented differently for models with covariates and models without covariates. For models with covariates (e.g., the proposed model), the item interaction parameter is constrained to be $\delta_i > 0$ by assuming that the error terms in the Equation 3.1.9 are uncorrelated for the following reasons: 1) it is difficult to implement truncated multivariate normal distribution in JAGS (Levy & Mislevy, 2016); 2) the correlations among the error terms are not focal parameters and constraints on them help facilitate the computation effciency of the parameter estimation; 3) it is reasonable to assume that the correlations among item parameters come from the shared variability due to the same set of item covariates. For models without covariates, the current study does not constrain $\delta_i > 0$, but instead constraining $\mu_\delta > 0$ to ensure that a majority of the items satisfy the order constraint for the following reasons: 1) truncated multivariate normal distribution is difficult to implement in JAGS (Levy & Mislevy, 2016); 2) it has been shown that the constraint $\delta_i > 0$ is not necessary for model identification (e.g., de la Torre, 2011; de la Torre & Douglas, 2004; Li, 2008; Zhan et al., 2018); 3) preliminary analysis on the simulated and empirical datasets does not reveal any violation of the $\delta_i > 0$ constraint using this relaxed constraint. Detailed model specifications are elaborated in the section 3.2.

3.1.5 True Model and Alternative Models

The proposed model is used as the data generating model in the simulation study, which is demonstrated in Figure 3. The true model and five competing models (CMs) with model misspecifications are used to fit the generated datasets, which are presented in the Table 1. The

five CMs are: the JRT-DINA model with no covariates, the model with only the dichotomous item covariate (JRT-DINA-LLTM-D), the model with only the continuous item covariate (JRT-DINA-LLTM-C), the seperate models with no covariates (i.e., the DINA model and the lognormal model), the separate models with item covariates (i.e., the DINA-LLTM and the lognormal-LLTM). The two-step estimation is conducted when fitting models without covariates, which means the item parameters are first estimated from the models and then regressed on the item covariates to obtain the regression coefficient estimates. These estimates can be further compared to those obtained from the proposed model.

The purpose of fitting the alternative misspecified models is to examine the parameter recovery under model misspecifications. Specifically, three types of model misspecifications are considered: 1) item covariates misspecifications, including the the model with only the dichotomous item covariate (JRT-DINA-LLTM-D), the model with only the continuous item covariate (JRT-DINA-LLTM-C) (Green & Smith, 1987); 2) ignoring the second level models, including the DINA-LLTM + the lognormal-LLTM (Loeys et al., 2011); 3) ignoring item covariate specifications and using the two-step estimation, including the JRT-DINA and the DINA model + the lognormal model (Green & Smith, 1987; Liao & Jiao, 2018; Park et al., 2018).

In the current study, all the item covariates are assumed to be independent from each other in the true model. In addition, the covariates are assumed to be linearly related to the item parameters. However, other regression functions can be used to link the covariates to the item parameters, such as the curvlinear functions. It is also possible to include more covariates, covariates of other distributions (e.g., ordinal or nominal) or interaction terms in the model.

However, these possibilities are not explored in the simulation study to make the scope of the study managable.

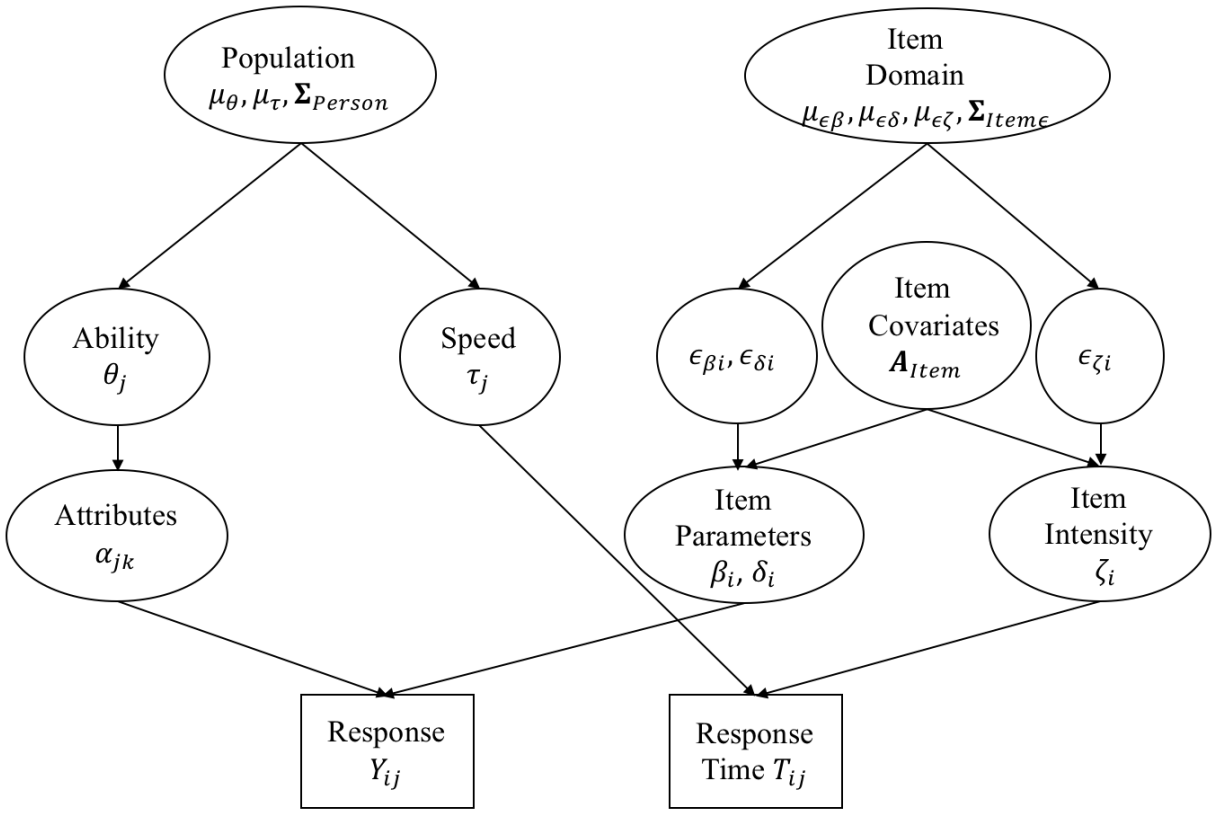


Figure 3. A graphical representation of the JRT-DINA-LLTM.

Table 1. *True data-generating model and alternative models.*

Model Type	Model Specification
True model (TM)	the JRT-DINA-LLTM with $\beta_i = \mathrm{A}_{0(\beta)} + \mathrm{A}_{1(\beta)}C_i + \mathrm{A}_{2(\beta)}D_i + \varepsilon_{\beta_i}$ $\delta_i = \mathrm{A}_{0(\delta)} + \mathrm{A}_{1(\delta)}C_i + \mathrm{A}_{2(\delta)}D_i + \varepsilon_{\delta_i}$ $\zeta_i = \mathrm{A}_{0(\zeta)} + \mathrm{A}_{1(\zeta)}C_i + \mathrm{A}_{2(\zeta)}D_i + \varepsilon_{\zeta_i}$ $\varepsilon \sim N(0, \Sigma_{Item\varepsilon})$
Competing Models 1. CM-N (joint model with no covariates)	the JRT-DINA model without covariates (two-step estimation)
2. CM-D (joint model with only the dichotomous covariate)	the JRT-DINA-LLTM with $\beta_i = \mathrm{A}_{0(\beta)} + \mathrm{A}_{2(\beta)}D_i + \varepsilon_{\beta_i}$ $\delta_i = \mathrm{A}_{0(\delta)} + \mathrm{A}_{2(\delta)}D_i + \varepsilon_{\delta_i}$ $\zeta_i = \mathrm{A}_{0(\zeta)} + \mathrm{A}_{2(\zeta)}D_i + \varepsilon_{\zeta_i}$ $\varepsilon \sim N(0, \Sigma_{Item\varepsilon})$
3. CM-C (joint model with only the continuous covariate)	the JRT-DINA-LLTM with $\beta_i = \mathrm{A}_{0(\beta)} + \mathrm{A}_{1(\beta)}C_i + \varepsilon_{\beta_i}$ $\delta_i = \mathrm{A}_{0(\delta)} + \mathrm{A}_{1(\delta)}C_i + \varepsilon_{\delta_i}$ $\zeta_i = \mathrm{A}_{0(\zeta)} + \mathrm{A}_{1(\zeta)}C_i + \varepsilon_{\zeta_i}$ $\varepsilon \sim N(0, \Sigma_{Item\varepsilon})$
4. CM-SN (separate models with no covariates)	the G-DINA model and the lognormal model (two-step estimation)
5. CM-S (separate models with covariates)	the G-DINA model and the lognormal model with $\beta_i = \mathrm{A}_{0(\beta)} + \mathrm{A}_{1(\beta)}C_i + \mathrm{A}_{2(\beta)}D_i + \varepsilon_{\beta_i}$ $\delta_i = \mathrm{A}_{0(\delta)} + \mathrm{A}_{1(\delta)}C_i + \mathrm{A}_{2(\delta)}D_i + \varepsilon_{\delta_i}$ $\zeta_i = \mathrm{A}_{0(\zeta)} + \mathrm{A}_{1(\zeta)}C_i + \mathrm{A}_{2(\zeta)}D_i + \varepsilon_{\zeta_i}$

3.2 Model Parameter Estimation

In the present study, the R2jags package (Su & Yajima, 2015) is used to interface with JAGS (version 4.3.0; Plummer, 2015), which will be used to estimate the parameters using the full Bayesian approach with the Markov chain Monte Carlo (MCMC) method. A Gibbs sampler (Gelfand & Smith, 1990) was implemented in JAGS.

Prior specifications. The statistical inference is based on the posterior distribution of the model parameters when the Bayesian estimation is utilized. The posterior distribution is derived from the prior distribution assigned to the model parameters and the observed likelihood based on observed data. Therefore, the specifications of the priors are crucial in the estimation of model parameters. In the current study, the priors of the model parameters are specified based on previous studies on the joint modeling of RA and RT using CDMs (Minchen, 2017; Zhan et al., 2018) and preliminary analysis of the empirical dataset. Assuming conditional independence, item responses Y_{ij} and log-transformed $\log(T_{ij})$ are conditionally and independently distributed as $Y_{ij} \sim Bernoulli(P(Y_{ij}=1|\alpha_j,\beta_i,\delta_i)$ and $\log(T_{ij}) \sim N(\zeta_i - \tau_j, \sigma_{\varepsilon i}^2)$.

In the data generating model, the item parameters are decomposed using a linear function of observed item covariates. For each item parameter, one intercept term ($\mathbf{A}_0$), two regression coefficients ($\mathbf{A}_1$, $\mathbf{A}_2$) are estimated. Noninformative priors for the intercept terms are specified as:

$$\begin{aligned} \mathbf{A}_{0(\beta)} &\sim Normal(0,10^6) \\ \mathbf{A}_{0(\delta)} &\sim Normal(0,10^6) \\ \mathbf{A}_{0(\zeta)} &\sim Normal(0,10^6) \end{aligned} . \tag{3.2.1}$$

As in Liao and Jiao (2018), the priors for the two regression coefficients for one dichotomous covariate and a continuous covariate are specified as noninformative for all three item parameters:

$$\begin{aligned} \mathbf{A}_1 &\sim Normal(0,10^6) \\ \mathbf{A}_2 &\sim Normal(0,10^6) \end{aligned} . \tag{3.2.2}$$

The *Normal* in the above equations indicates the normal distribution. The two numbers in the parentheses after *Normal* indicate the mean and variance of the normal distribution.

The priors for item parameters are assumed to follow normal distributions which are specified as:

$$\beta_i \sim N(\mathbf{A}_{(\beta)}\mathbf{B}_i, \sigma_\beta^2),$$

$$\delta_i \sim N(\mathbf{A}_{(\delta)}\mathbf{B}_i, \sigma_\delta^2)\ T(0,+\infty)\ ,$$

$$\zeta_i \sim N(\mathbf{A}_{(\zeta)}\mathbf{B}_i, \sigma_\zeta^2).$$

The inverse-gamma is chosen as the hyperpriors for the variances of the item parameters (i.e., σ_β^2, σ_δ^2, σ_ζ^2) and the prior for the variance of the log-RTs (σ_ε^2):

$$\sigma_\beta^2 \sim \text{InvGamma (1, 1)},$$

$$\sigma_\delta^2 \sim \text{InvGamma (1, 1)},$$

$$\sigma_\zeta^2 \sim \text{InvGamma (1, 1)},$$

$$\sigma_\varepsilon^2 \sim \text{InvGamma (1, 1)}.$$

The inverse-gamma is chosen because it is a conjugate prior to the variance components in the normal distribution conditional on the means. Although Gelman (2006) discussed the disadvantages of using the inverse-gamma as the prior to the variance components, the issue arises only when the variance is near zero.

The priors for the person parameters in the data generating model and alternative models are specified as:

$$\begin{pmatrix} \theta_j \\ \tau_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \mathbf{\Sigma}_{person}\right), \mathbf{\Sigma}_{person} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix},$$

where $\mu_\theta = \mu_\tau = 0$ and σ_θ^2 = 1 for the scale identification. The inverse-Wishart prior cannot be used as the prior for $\boldsymbol{\Sigma}_{person}$ due to the constraint σ_θ^2 = 1. Therefore, the current study follows Zhan et al. (2018) in reparameterizing $\boldsymbol{\Sigma}_{person} = \boldsymbol{\Delta}_{person}\boldsymbol{\Delta}'_{person}$ using the Cholesky decomposition, where

$$\Delta_{person} = \begin{pmatrix} 1 & 0 \\ \varphi & \psi \end{pmatrix}, \tag{3.2.3}$$

is a lower triangular matrix with ψ being positive while φ being unrestricted entries. Specifically, the priors for these two elements are $\varphi \sim N(0,1)$ and $\psi \sim Gamma(1,1)$; Δ'_{person} is the conjugate transpose of Δ_{person}. The priors for the higher-order structural parameters are

$$\gamma_k \sim Normal(0,4)T(0,+\infty),$$

$$\lambda_k \sim Normal(0,4). \tag{3.2.4}$$

$T(0,+\infty)$ indicates that the values are truncated to be positive. Therefore, the attribute-slope parameter γ_k is specified as following a normal distribution with mean = 0 and variance = 4, but only positive values are considered. This restriction is due to the assumption that a respondent with higher latent ability is more likely to master more attributes. Further, attributes $\boldsymbol{\alpha}_{jk}$ are conditionally independent given θ. The prior distribution of the attributes is further specified as:

$$\alpha_{jk} \mid \theta_j, \gamma_k, \lambda_k \sim Bernoulli\left(\frac{\exp(\gamma_k\theta_j + \lambda_k)}{1+\exp(\gamma_k\theta_j + \lambda_k)}\right). \tag{3.2.5}$$

For the models with model misspecifications, most prior specifications remain the same except for the ones listed as follows. Specifically, for the JRT-DINA model without covariates (CM-N), the priors for the item parameters are specified as:

$$\begin{pmatrix} \beta_i \\ \delta_i \\ \zeta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\zeta \end{pmatrix}, \Sigma_{item} \right), \Sigma_{item} \sim \text{ InvWishart } (\mathbf{R}_{\text{item}}, 3),$$

where $\mathbf{R}_{\text{item}}$ is a three-dimensional identity matrix. The inverse-Wishart is chosen because it is a conjugate prior to the multivariate normal distribution. Further, the hyper priors are specified as $\mu_\beta \sim N(-2.197,2)$, $\mu_\delta \sim N(4.394,2)I(\mu_\delta > 0)$, $\mu_\zeta \sim N(4,2)$. The hyper priors μ_β and μ_δ follow Zhan et al. (2017), assuming the means of the guessing probability and the slipping probability equal 0.1. The hyper prior μ_ζ is set based on the mean log-transformed RT in the empirical dataset (i.e., 4.03).

For the competing model with only dichotomous item covariate (CM-D) and the competing model with only continuous item covariate (CM-C), one prior of the standard normal distribution is assigned to the regression coefficient, as shown in the Equation 3.2.2.

For the separate DINA model and lognormal model without covariates (CM-S), the priors of the item parameters in the DINA model are assumed to follow the bivariate normal distribution:

$$\begin{pmatrix} \beta_i \\ \delta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \end{pmatrix}, \Sigma_{item} \right), \Sigma_{item} \sim \text{ InvWishart } (\mathbf{R}_{\text{item}}, 2),$$

the hyper priors are specified as $\mu_\beta \sim N(-2.197,2)$, $\mu_\delta \sim N(4.394,2)I(\mu_\delta > 0)$. The item parameters ($\beta_i$ and δ_i) are assumed to be correlated to capture the within-item characteristic dependency, as in Zhan et al. (2019). Note that Zhan et al. (2019) specified the within-item characteristic dependency as the correlation between logit(s_i) and logit(g_i), while the current study specifies it as the dependency between β_i and δ_i to make the parameterization consistent

among different models. The prior for the higher-order ability θ is the standard normal distribution. The prior for the latent speed parameter τ is a normal distribution $N(0,\sigma_\tau^2)$ with the hyper parameter $\sigma_\tau^2 \sim$ InvGamma(1,1). The prior of the time intensity parameter follow a normal distribution $N(\mu_\zeta,\sigma_\zeta^2)$ with the hyper priors $\mu_\zeta \sim N(4,2)$ and $\sigma_\zeta^2 \sim$ InvGamma(1,1). For the separate DINA model and lognormal model with covariates (CM-S), the priors for the item intercept terms and the regression coefficients are the same as shown in Equations 3.2.1 and 3.2.2. The priors of the item parameters in the DINA model are assumed to follow the normal distribution $\beta_i \sim N(\mathbf{A}_{(\beta)}\mathbf{B}_i,\sigma_\beta^2)$, $\delta_i \sim N(\mathbf{A}_{(\delta)}\mathbf{B}_i,\sigma_\delta^2)$, $\zeta_i \sim N(\mathbf{A}_{(\zeta)}\mathbf{B}_i,\sigma_\zeta^2)$. Hyper priors for the variances are inverse-gamma, $\sigma_\beta^2 \sim$ InvGamma (1, 1), $\sigma_\delta^2 \sim$ InvGamma (1, 1) and $\sigma_\zeta^2 \sim$ InvGamma (1, 1).

All sampled model parameters in the proposed model are given in the *S* shown as:

$$S = \left\{\alpha_k,\theta_j,\tau_j,\lambda_k,\gamma_k,\mathbf{\Sigma}_{Item\varepsilon},\mathbf{\Sigma}_{Person},\sigma_\varepsilon{}^2,\mathbf{A}\right\}. \tag{3.2.6}$$

Given the priors and the *S*, the joint posterior probability of the proposed model is written as:

$$P(S|\mathbf{Y},\log(\mathbf{T})) \propto L(\mathbf{Y},\log(\mathbf{T})|\boldsymbol{\alpha},\mathbf{A},\boldsymbol{\tau},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2)$$

$$\times P(\boldsymbol{\alpha}\,|\,\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\theta})P(\boldsymbol{\lambda})P(\boldsymbol{\gamma})P(\boldsymbol{\theta},\boldsymbol{\tau}\,|\,0,\mathbf{\Sigma}_{Person})P(\mathbf{\Sigma}_{Person})P(\mathbf{A})P(\boldsymbol{\varepsilon}_{(\boldsymbol{\beta})},\boldsymbol{\varepsilon}_{(\boldsymbol{\delta})},\boldsymbol{\varepsilon}_{(\boldsymbol{\zeta})}\,|\,0,\mathbf{\Sigma}_{Item\boldsymbol{\varepsilon}})P(\mathbf{\Sigma}_{Item\boldsymbol{\varepsilon}})P(\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2)$$

where

$$L(\mathbf{Y},\log(\mathbf{T})|\boldsymbol{\alpha},\mathbf{A},\boldsymbol{\tau},\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2) = \prod_{j=1}^{J}\prod_{i=1}^{I} P(Y_{ij}\,|\,\boldsymbol{\alpha}_{jk},\mathrm{A}_{(\boldsymbol{\beta},\boldsymbol{\delta})})\; f(\log(T_{ij})\,|\,\mathbf{A}_{(\zeta)},\tau_j,\sigma_{\varepsilon i}^2), \tag{3.2.7}$$

is the likelihood of the proposed model. Finally, the posterior mean and posterior mode are used as estimates for continuous parameters (e.g., τ_j and θ_j) and categorical parameters (i.e., attributes), respectively.

The number of MCMC chains, iterations, and convergence check. In the Bayesian estimation, two Markov chains are run with random starting points. The thinning interval is set as two. The number of iterations per chain, the number of burn-in and resulting posterior sample sizes are presented in Table 2. The decisions are made based on model type and simulation conditions which are described in section 3.3.1. The potential scale reduction factor (PSRF; Brooks & Gelman, 1998), also known as $\hat{R}$, is computed for each parameter. $\hat{R}$ is a ratio between the estimated pooled variance of the MCMC draws (i.e., sum of the weighted means of the within-chain variance and the between-chain variance) and the estimated within-chain variance. In general, $\hat{R}$ smaller than 1.1 or 1.2 is considered as the convergence criterion for the parameter estimation (e.g., Zhan et al., 2018; de la Torre & Douglas, 2004); the current study adopts the convergence criterion as $\hat{R}$ smaller than 1.1. In addition, the trace plots of the parameter estimates are examined to check the convergence. Trace plots display the parameters drawn from each iteration of the MCMC chains. Specifically, if the traces from different MCMC chains are stable and converge to the same location, the convergence is achieved. Otherwise, if the traces from different MCMC chains separate apart and do not converge to the same location, it is claimed that the convergence is not achieved. In addition, the effective sample size (ESS) is used to indicate the precision of the Bayesian parameter estimates. Specifically, ESS smaller than 400 is considered as lack of precision of the Bayesian estimation (Zitzmann & Hecht, 2019). Preliminary simulation runs showed that the convergence was well achieved.

Table 2. MCMC Iterations in Simulation Study.

Condition	Model	Number of Iterations	Number of Burn-in	Posterior Sample Size
$J = 500$	JRT-DINA-LLTM	40000	20000	20000
$I = 20$	JRT-DINA	40000	20000	20000
	DINA	40000	20000	20000
	DINA-LLTM	40000	20000	20000
	Lognormal	12000	6000	6000
	Lognormal-LLTM	12000	6000	6000
	JRT-DINA-LLTM-C	40000	20000	20000
	JRT-DINA-LLTM-D	40000	20000	20000
$J = 500$	JRT-DINA-LLTM	30000	15000	15000
$I = 40$	JRT-DINA	30000	15000	15000
	DINA	30000	15000	15000
	DINA-LLTM	30000	15000	15000
	Lognormal	20000	10000	10000
	Lognormal-LLTM	20000	10000	10000
	JRT-DINA-LLTM-C	30000	15000	15000
	JRT-DINA-LLTM-D	30000	15000	15000
$J = 1000$	JRT-DINA-LLTM	30000	15000	15000
$I = 20$	JRT-DINA	30000	15000	15000
	DINA	30000	15000	15000
	DINA-LLTM	30000	15000	15000
	Lognormal	16000	8000	8000
	Lognormal-LLTM	16000	8000	8000
	JRT-DINA-LLTM-C	30000	15000	15000
	JRT-DINA-LLTM-D	30000	15000	15000
$J = 1000$	JRT-DINA-LLTM	20000	10000	10000
$I = 40$	JRT-DINA	20000	10000	10000
	DINA	20000	10000	10000
	DINA-LLTM	20000	10000	10000
	Lognormal	20000	10000	10000
	Lognormal-LLTM	20000	10000	10000
	JRT-DINA-LLTM-C	20000	10000	10000
	JRT-DINA-LLTM-D	20000	10000	10000

Note. J = Sample Size; I = Test Length.

3.3 Simulation Design

The proposed model and the Bayesian estimation method are presented in the previous two sections. To examine the performance of the proposed model, a simulation study is carried out and the detailed simulation design is delineated in this section. First, the manipulated factors

and fixed factors are presented and justified. Then, the outcome measures for evaluating parameter recovery and overall model fit indices are introduced. The purpose of the simulation study is: 1) to examine the parameter recovery of the proposed model under several simulated conditions; 2) to compare the performance of the proposed model and that of the alternative models; 3) to evaluate how relative model fit indices perform in identifying the proposed model as the best data fitting model; and 4) to compare the accuracy and efficiency of the regression coefficient estimates obtained from the proposed model and those from the competing models.

3.3.1 Manipulated Factors

In this simulation study, the proposed JRT-DINA-LLTM is used as the true data generating model. The proposed model and alternative models listed in the Table 1 are used as data fitting model. The manipulated factors and corresponding levels are determined based on empirical data analysis findings from previous studies (e.g., De Boeck, 2008; Hartig et al., 2012; Liao & Jiao, 2019) and the simulation study design. Specifically, the manipulated factors in the simulation study include: sample size (500, 1000), test length (20, 40), the absolute values of the regression coefficients for the items parameters (CDM: 0.4, 0.8; RT model: 0.2, 0.4), the correlation between person ability and person speed $\rho_{\theta\tau}$ (0.2, 0.5, 0.8). Specific levels of the manipulated factors are presented in Table 3. The four manipulated factors are fully crossed, yielding 24 simulation conditions, as summarized in Table 4.

Table 3. *Summary of manipulated factors.*

	Manipulated Factors			
Levels	Sample Size	Test Length	Regression coefficient	$\rho_{\theta\tau}$
1	500	20	$A_{1(\boldsymbol{\beta})} = A_{2(\boldsymbol{\beta})} = -0.4$ $A_{1(\boldsymbol{\delta})} = A_{2(\boldsymbol{\delta})} = 0.4$ $A_{1(\zeta)} = A_{2(\zeta)} = 0.2$	0.2
2	1000	40	$A_{1(\boldsymbol{\beta})} = A_{2(\boldsymbol{\beta})} = -0.8$ $A_{1(\boldsymbol{\delta})} = A_{2(\boldsymbol{\delta})} = 0.8$	0.5

	$A_{1(\zeta)} = A_{2(\zeta)} = 0.4$	
3		0.8

Note. $\rho_{\theta\tau}$ = Correlation between ability and speed.

Table 4. *Summary of simulation conditions.*

Condition No.	Manipulated Factors			
	Sample Size	Test Length	Regression Coefficient	$\rho_{\theta\tau}$
1	500	20	Small	0.2
2	500	20	Small	0.5
3	500	20	Small	0.8
4	500	20	Large	0.2
5	500	20	Large	0.5
6	500	20	Large	0.8
7	500	40	Small	0.2
8	500	40	Small	0.5
9	500	40	Small	0.8
10	500	40	Large	0.2
11	500	40	Large	0.5
12	500	40	Large	0.8
13	1000	20	Small	0.2
14	1000	20	Small	0.5
15	1000	20	Small	0.8
16	1000	20	Large	0.2
17	1000	20	Large	0.5
18	1000	20	Large	0.8
19	1000	40	Small	0.2
20	1000	40	Small	0.5
21	1000	40	Small	0.8
22	1000	40	Large	0.2
23	1000	40	Large	0.5
24	1000	40	Large	0.8

Note. $\rho_{\theta\tau}$ = Correlation between ability and speed.

Sample size, test length, and Q-matrices. The sample size and test length are manipulated to examine their impact on parameter recovery. Sample size can influence the estimation accuracy and precision of the item parameters, which in turn may affect the estimation of the regression coefficients. Sample size is manipulated at two levels: 500 and 1000. These two levels are typically observed in applied research and methodological studies. In the CDM-based joint modeling literature, most simulation studies include sample size levels around

500 and 1000. For example, Minchen (2017) used sample sizes of 500, 1000, and 2000; Wang et al. (2020) used sample sizes of 500 and 1000; Zhang and Wang (2018) used sample sizes of 585, 1000, and 3000. When using a fixed sample size in the simulation study or conducting empirical data analysis, Zhan et al. (2018) used a sample size of 1000 and Wang et al. (2018) used 351, respectively. Therefore, the current study chose the sample size of 500 to represent a small sample condition and 1000 to represent an adequately large sample size condition in the investigation of parameter recovery.

Test length can influence the recovery of person parameters, including person ability, speed, and latent attributes. In addition, in the item explanatory modeling scenario, test length (i.e., number of items) may affect the recovery of the regression coefficient estimation. Specifically, it is expected that smaller number of items may jeopardize the estimation accuracy and precision of the regression coefficients. Therefore, test length is manipulated at two levels: 20 and 40. A diagnostic assessment with test length of 40 is typical considered as a long test in the empirical datasets from diagnostic assessments. For example, Wang et al. (2020) developed and evaluated a multidimensional diagnostic assessment consisting of 40 items that measured 4 attributes. Zhang and Wang (2018) also fixed the test length at 40 in the simulation study that evaluated the parameter recovery of a mixture hierarchical model of RA and RT that tracked the longitudinal transition of the mastery status of 4 attributes. Further, Wang et al. (2020) simulated 20 and 40 items as two levels of the test length to evaluate the parameter recovery of the hierarchical model that considered CD of RT on RA. Thus, the current study simulated 20 and 40 items as two levels of the test length to measure 4 latent attributes, where 20 items and 40 items are considered as a short test scenario and a long test scenario, respectively.

The simulated Q-matrices are presented in Table 5. The Q-matrices in both test length conditions include identity matrices that ensure the completeness of the Q-matrix (Chiu et al., 2013). It also satisfies the identifiability of the DINA model (Gu & Xu, 2019).

Table 5. Q-matrices Used in the Simulation Study.

Item	α_1	α_2	α_3	α_4
1	**1**			
2		**1**		
3			**1**	
4				**1**
5	**1**	**1**		
6		**1**	**1**	
7			**1**	**1**
8	**1**			**1**
9	**1**	**1**	**1**	
10		**1**	**1**	**1**
11	**1**			
12		**1**		
13			**1**	
14				**1**
15	**1**	**1**		
16		**1**	**1**	
17			**1**	**1**
18	**1**			**1**
19	**1**	**1**	**1**	
20		**1**	**1**	**1**
21	1			
22		1		
23			1	
24				1
25	1	1		
26		1	1	
27			1	1
28	1			1
29	1	1	1	
30		1	1	1
31	1			
32		1		
33			1	
34				1
35	1	1		
36		1	1	
37			1	1
38	1			1

39	1	1	1	
40		1	1	1

Note. Blank means "0"; the first 20 rows in bold face indicates the Q-matrix used in the test length = 20 condition; the 40 rows indicates the Q-matrix used in the test length = 40 condition.

Regression Coefficients. The magnitude of the regression coefficients is manipulated to reflect strong and weak relations between the item covariates and the item parameters. Specifically, absolute values of the regression coefficients are set at .400 or .800 for the item intercept and item interaction parameters in the CDM to indicate strong and weak relations, respectively; the regression coefficient is set at .200 or .400 for the item intensity parameter to indicate strong and weak relations, respectively. The regression coefficients for the item intercept parameter are set at negative values, opposite from the item interaction and item intensity parameters, based on empirical findings that items easy to guess are usually hard to slip and take longer to complete (e.g., Zhan et al., 2018). The resulting variance explained R^2 values in the item parameters in the weak and strong relation conditions are .200 and .800, respectively. These R^2 values are reasonable to reflect a low predictive power scenario and a strong predictive power scenario based on the empirical findings in the literature. Specifically, based on the empirical results using the LLTM $+\varepsilon$ model, Hartig et al. (2012) found that the R^2 value in their empirical study to be .396, while De Boeck (2008) reported a R^2 value of .875. Based on the joint modeling using the Rasch model and the lognormal RT model, Klein Entink, Kuhn, Hornke, et al. (2009) reported the R^2 of the item difficulty parameter and item intensity parameter were .340 and .370, respectively. In the item explanatory CDM study, Liao and Jiao (2018) found that the R^2 values were .260 and .300 for the item intercept and item interaction parameters, respectively.

The detailed variance decomposition and the corresponding item parameter vairances in the strong and weak relation conditions are showed in Table 6. The resulting variances for the item intercept ($\boldsymbol{\beta}_i$) and interaction ($\boldsymbol{\delta}_i$) parameters in the CDM equal 1.000 in the two R^2 conditions. These values are consistent with the empirical results in the CDM studies where the variances of the item intercept and interaction parameters are around 1.000 (Zhan et al., 2018). The resulting variance for the time intensity ($\boldsymbol{\zeta}_i$) parameter in the RT model equals .250 in both R^2 conditions, which is consistent with empirical findings in Zhan et al. (2017) and Zhan et al. (2018). Given the fixed means of the item parameters, the corresponding means of the item intercept ($\boldsymbol{\beta}_i$) parameter equal -1.997 and -1.797, respectively; the corresponding means of interaction ($\boldsymbol{\delta}_i$) parameter equal 4.194 and 3.994, respectively; the corresponding means of time intensity ($\boldsymbol{\zeta}_i$) parameter equal 3.900 and 3.800, respectively. Therefore, the slipping (s_i) and guessing (g_i) probabilities range from 0 to .650, indicating that the simulated items consist of both high-quality (1-*s*-*g* >= 0.650) and low-quality (1-*s*-*g* < 0.650) items (de la Torre, 2007). In addition, the time required to solve an item is in general within one to two minutes as in the real-world settings (Lee, 2007).

Table 6. The simulated regression coefficients and residual variances.

Item Parameter	Weak Relation					Strong Relation				
	Total variance	$\mathbf{A}_1$	$\mathbf{A}_2$	$\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2$	R^2	Total variance	$\mathbf{A}_1$	$\mathbf{A}_2$	$\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2$	R^2
β_i	1.000	.400	.400	.800	.200	1.000	.800	.800	.200	.800
δ_i	1.000	.400	.400	.800	.200	1.000	.800	.800	.200	.800
ζ_i	.250	.200	.200	.200	.200	.250	.400	.400	.050	.800

Note. $\mathbf{A}_1$ and $\mathbf{A}_2$ denote the regression coefficients for the continuous and dichotomous item covariates, respectively. $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2$ denote the variances of the residual terms of the item parameters.

Correlation between person ability and speed. The magnitude of the correlation between person ability and speed typically affects the measurement precision of the person ability. That is, larger correlation between person ability and speed usually improves the meansurement precision and the attribute classifiication accuracy. Given that separate models are considered as one type of competing models, it is meaningful to see how large the impact would be in the item explanatory modeling scenario when the second-level structual component on the person side is omitted. The choice of the $\rho_{\theta\tau}$ levels depends on empirical findings and simulation study designs of existing estudies. Based on the empirical results from hierarchical models based on CDMs, the correlation person ability and speed was estimated as -.20 (Zhan, Liao, & Bian, 2018) and -.50 (Zhan et al., 2018). The empirical findings from hierarchical models based on IRT models revealed similar results. For example, van der Linden (2009) found that the $\rho_{\theta\tau}$ ranged from -.65 to .30. In simulation studies, $\rho_{\theta\tau}$ larger than 0.6 was usually simulated to indicate larger correlation. For example, Klein Entink (2009) set $\rho_{\theta\tau}$ as 0, .25, .75, and 1 to mimick low to high person parameter correlations. Patton (2015) considered a similar range of values of $\rho_{\theta\tau}$ = 0, .30, .60, .90. Therefore, the current simulation study chose to manipulate .20, .50, .80 to represent small, moderate, and large correlation between person ability and person speed. Only positive values are considered because the sign does not affect its impact on parameter estimation accuracy.

3.3.2 Fixed Factors

In the current simulation study, the parameter recovery of the regression coefficients and their associated standard errors is of major interest, along with latent attribute and profile classification accuracies. Therefore, factors that have been extensively studied in the hierarchical modeling studies (e.g., Zhan et al., 2018) or are not expected to affect the parameter recovery are

fixed, as listed in Table 7. The fixed factors include: distribution of latent ability, distribution of latent speed, distribution of the item covariates, distribution of the log-RT variance, correlation between ability and speed, number of latent attributes, attribute slope and intercept parameters, and means of the item parameters. In addition, the CDM and the RT model used in the current study are the DINA model and the lognormal model, respectively.

Table 7. *Summary of fixed factors.*

Factor	Fixed Value
Distribution of the ability	$N(0, 1)$
Distribution of the speed	$N(0, .250)$
Number of attributes	4
Attribute slope parameter γ_k	$\gamma_k = 1$
Attribute intercept parameter λ_k	$\lambda_1 = 1,\ \lambda_2 = .500,\ \lambda_3 = -.500,\ \lambda_4 = -1$
Means of the item parameters ($\mu = \mathrm{A}_0 + \sum_{m=1}^{M} \mathrm{A}_m \mathrm{B}_{mi}$)	$\mu_{(\boldsymbol{\beta})} = -2.197,\ \mu_{(\delta)} = 4.394,\ \mu_{(\zeta)} = 4$
Distribution of dichotomous item covariates	Bernoulli(p = 0.5)
Distribution of continuous item covariates	$N(0, 1)$
Variance of the RT residual (σ_ε^2)	0.250
CDM	the DINA model
RT model	the lognormal model

Higher-order structural parameters and latent variables. Each test measures 4 latent attributes in the current study. The true attribute profile of each respondent is thereby randomly chosen from the 16 possible latent attribute profiles with equal probability. The attribute intercept parameters λ_k (k = 1, 2, 3, 4) are fixed at 1, .5, -.5, -1, indicating that the probability of the mastery of each attribute equals .730, .620, .380 and .270 when latent ability is 0, respectively. The attribute slope parameter is fixed at 1 for all four attributes. Thus, no intersection exists among the four probability curves of the attribute mastery. In other words, the probabilities of mastering the four attributes increase at the same rate as latent ability θ increases. The latent ability θ is assumed to be drawn from the standard normal distribution,

following the convention from numerous studies in IRT literature and the higher-order structure CDM literature (e.g., de la Torre & Douglas, 2004; Zhan et al., 2018). The latent speed is drawn from a normal distribution with mean 0 and variance .250, which is chosen based on empirical studies (e.g., Klein Entink et al., 2009; Molenaar & Bolsinova, 2017; Zhan et al., 2018).

Residual correlations and means of the item parameters and the log-RT variance. In the current study, item parameters are decomposed into a linear function of item covariates and residual terms. The correlations among the item parameter residual terms are set at 0. As mentioned in previous section, it is reasonable to assume that the correlations among item parameters come from the same set of item covariates. In addition, correlations among the residual terms are not focal parameters and constraints on these parameters improve the computation efficiency of the estimation of the proposed model. Further, the means of the item parameters in the linear regression functions are fixed at -2.197, 4.394, and 4 based on Zhan et al. (2017). This indicates that the average item guessing probability equals .100; the average item slipping probability equals .100; and the average time required to complete an item equals 54.6 seconds for a respondent with speed 0 to mimic real-world scenarios.

The variance of the log-RT ($\sigma_{\varepsilon i}^2$) is fixed at .25 across items. This is decided based on the empirical findings using the lognormal model that the difference among $\sigma_{\varepsilon i}^2$ is negligible and $\sigma_{\varepsilon i}^2$ is generally around .25 (e.g., Liao, 2018; Zhan et al., 2018).

Distribution of the item covariates. The distribution of the continuous item covariate is assumed to follow a standard normal distribution. Commonly seen continuous item covariates in the literature include word count in the item stimulus, sentence length, number of mathematics terms, word count of the key, and word counts of the options (e.g., Ferrara et al., 2011; Liao & Jiao, 2018; McLeod et al., 2015; Rowe et al., 2006). These variables may have wide ranges in

real-world settings. They are standardized in the linear regression function of the item parameters, as in Liao and Jiao (2018). The distribution of the dichotomous item covariate is set at 50% ones and 50% zeros, which mimics the empirical dataset used in the current study. This indicates that the mean of the dichotomous item covariate equals 0.500 and the variance equals 0.250.

A total of $2 \times 2 \times 2 \times 3 = 24$ simulated conditions exist in the current simulation study. Thirty replications are run to evaluate the parameter recovery of the proposed model. The replication number of thirty is chosen based on previous CDM-based joint modeling of RA and RT (Zhan et al., 2018). A total of 1440 datasets are generated for all conditions. In each condition, the proposed model is used as the data generating model. The alternative models including the JRT-DINA model with no covariates (CM-N), the seperate models (i.e., the DINA model and the lognormal model; CM-SN), the model omitting the continuous item covariate (CM-D), the model omitting the dichotomous item covariate (CM-C), the separate models (i.e., the DINA model and the lognormal model) with item covariates (CM-S) and the proposed model are data fitting models. Therefore, there are a total of $6 \times 24 = 144$ simulation cells and $144 \times 30 = 4320$ replications. R version 3.3.3 (R Core Team, 2017) is used to generate data. The R2jags package is used to interface with JAGS to evaluate model performance.

3.3.3 Data Generating Procedure

The proposed model JRT-DINA-LLTM is used as the data generating model. The data are simulated according to the following steps:

1) Simulate true person ability and speed parameters ($\boldsymbol{\theta}$ and $\boldsymbol{\tau}$). The person ability and speed parameters are drawn from a multivariate normal distribution , with mean

vector being $\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and the variance-covariance matrix being

$\boldsymbol{\Sigma}_{person} = \begin{pmatrix} 1 & \\ \sigma_{\theta\tau} & 1 \end{pmatrix}$, where $\sigma_{\theta\tau}$ = 0.100, 0.250, or 0.400 depending on the simulatoin conditions.

2) Simulate true item parameter residual variance-covariance matrix $\boldsymbol{\Sigma}_{item\varepsilon}$, which is represented as $\begin{pmatrix} \sigma^2_{\varepsilon\beta} & & \\ \sigma_{\varepsilon\beta\delta} & \sigma^2_{\varepsilon\delta} & \\ \sigma_{\varepsilon\beta\zeta} & \sigma_{\varepsilon\delta\zeta} & \sigma^2_{\varepsilon\zeta} \end{pmatrix}$. The $\boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}}$ are specified as in Table 7 depending on the strong or weak predictive power conditions. Accordingly, the residual correlations were set to be $\sigma_{\varepsilon\beta\delta}$ = $\sigma_{\varepsilon\beta\zeta}$ = $\sigma_{\varepsilon\delta\zeta}$ = 0.

3) Simulate true item parameters ($\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{\zeta}$). The item parameters are generated using a linear function of observed covariates (i.e., one continuous covaraite C_i and one dichotomous covariate D_i), i.e., $\mathbf{AB}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{A} = (\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2)$ and $\mathrm{B}_i = (1, C_i, D_i)'$. The $\mathbf{A}_1 = \mathbf{A}_2$ = (-0.4, 0.4, 0.2) or (-0.8, 0.8, 0.4) depending on the strong or weak predictive power conditions. The $\boldsymbol{\varepsilon}_i$ are drawn from the $\boldsymbol{\Sigma}_{item\varepsilon}$ obtained from the previous step. The means of the item parameters are fixed at μ_β = -2.197, μ_δ = 4.394, μ_ζ = 4 as shown in Table 7.

4) Simulate true higher-order structual parameters as specified in Table 3. Specificaly, the higher-order structual parameters include the attribute slope parameters γ_k = 1 (k = 1,…, 4) and attribute intercept parameters λ_1 = 1, λ_2 = .5, λ_3 = -.5, λ_4 = -1.

5) Simulate attribute mastery status parameter α. The attribute mastery status parameters are simulated from the bernoulli distribution for which the probability functions are obtained according to Equation 3.1.5 using the person ability parameters and higher-order strutual parameters simulated from previous steps.
6) Simulate response data using Equation 3.1.2 by plugging in true item parameters, attribute master status parameters, and Q-matrix as specified in Table 5.
7) Simulate response time data using Equation 3.1.6 by plugging the person speed parameters and time intensity parameters ζ. Specifically, the log-tranformed item response time $\log(T_{ij})$ follows a normal distribution with a mean of $\tau_j - \zeta_i$ and variance of 0.25.

3.3.4 Evaluation Criteria and Analysis Procedure

The outcome measures and statistical analysis of the simulation outcomes are introduced in this section to answer the research questions related to the parameter recovery of the proposed model and the performance of the relative model fit indices in identifying the proposed model as the best data fitting model.

The absolute model fit of the data fitting models is assessed by the posterior predictive *p*-value (PPP). The PPP is obtained based on the posterior predictive model check (PPMC) method. Specifically, the PPMC method assesses the fit of the model by examining whether the observed data ($\boldsymbol{y}$) appear extreme with respect to the replication of the posterior distribution of the replicated data ($\boldsymbol{y}^{\text{rep}}$), which is generated using the same model and the same model parameter values ($\boldsymbol{\Theta}$) as $\boldsymbol{y}$. The discrepancy between $\boldsymbol{y}^{\text{rep}}$ and $\boldsymbol{y}$ is quantified by the discrepancy measures $D(\boldsymbol{y}^{\text{rep}}|\boldsymbol{\Theta})$ and $D(\boldsymbol{y}|\boldsymbol{\Theta})$. The PPP is the probability that $\boldsymbol{y}^{\text{rep}}$ could be more extreme than $\boldsymbol{y}$ as measured by the discrepancy measures. The PPP values near 0.5 indicate adequate model-data

fit, while extreme PPP values (> .95 or < .05) indicate systematic differences between observed and predicted data. The current study followed Zhan et al. (2018) to evaluate the absolute model-data fit of RA and RT separately. The sum of the squared Pearson residuals for respondent *j* and item *i* (Yan et al., 2003) was used as the discrepancy measure for the RA model, which is written as:

$$D(Y_{ij};\alpha_j,\beta_i,\delta_i)=\sum_{j=1}^{J}\sum_{i=1}^{I}\left(\frac{Y_{ij}-P(Y_{ij}=1)}{\sqrt{P(Y_{ij}=1)(1-P(Y_{ij}=1))}}\right)^2, \tag{3.3.1}$$

where $P(Y_{ij}=1)$ has the same distribution as that in the Equation 3.1.2. The sum of the standardized residual function of $\log(T_{ij})$ for respondent *j* and item *i* (Marianti et al., 2014) was used as the discrepancy measure for the RT models:

$$D(\log T_{ij};\tau_j,\zeta_i,\varepsilon)=\sum_{j=1}^{J}\sum_{i=1}^{I}(\varepsilon(\log T_{ij}-(\zeta_i-\tau_j)))^2. \tag{3.3.2}$$

Assessing the performance of model fit indices. Relative model fit indices, including the Akaike information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978), and the deviance information criterion (DIC; Spiegelhalter et al., 2002) are used for model selection. The performance of the model fit indices is quantified by the frequency of replications that one particular model fit index correctly identifies true data generating model as the best fitting model (i.e., the proposed model has the smallest model fit index compared to alternative models with misspecifications). According to Congdon (2003), when used in Bayesian MCMC estimation, the AIC and BIC model fit indices are specified as:

$$AIC=\bar{D}+p, \tag{3.3.3}$$

$$BIC=\bar{D}+p(\log J-1), \tag{3.3.4}$$

where $\bar{D}$ denotes the posterior mean of the deviance; p denotes the number of parameters; J denotes the sample size. Both the AIC and the BIC penalize models with more parameters, but the BIC has heavier penalization on the more complex model.

The DIC is a generalization of the AIC and the BIC to be used in the Bayesian MCMC method and is calculated as:

$$DIC = \bar{D} + p_D, \tag{3.3.5}$$

where $\bar{D}$ denotes the posterior mean of the deviance; p_D is the effective number of parameters, which can be estimated by $\mathrm{var}(D)/2$ (Gelman et al., 2014; Su & Yajima, 2015), that is, half of the posterior variance of the deviance. The smaller values of the AIC, the BIC and the DIC indicate better model fit.

The model selection was further quantified based on Anderson (2008). First, the difference between the value of a fit index and its minimum value among R fitted models was computed, denoted as $\Delta_r (r = 1,\ldots,R)$. Second, the likelihood of model r given data x was computed and denoted as:

$$L(g_r \mid x) \propto \exp(-\Delta_r / 2), \tag{3.3.6}$$

This likelihood represents the relative strength of evidence among competing models and is different from the likelihood of the data given a statistical model. Third, these model likelihoods were normalized to a set of Akaike weights ω_r that added up to 1, which describe the probability of model r is the expected best fitting model:

$$\omega_r = \frac{\exp(-\Delta_r / 2)}{\sum_{r=1}^{R} \exp(-\Delta_r / 2)}, \tag{3.3.7}$$

Lastly, the evidence ratio between any two competing models i and j was computed as:

$$E_{i,j} = \frac{L(g_i \mid x)}{L(g_j \mid x)} = \frac{\omega_i}{\omega_j}. \tag{3.3.8}$$

According to Anderson (2008), evidence ratios larger than 55 usually indicate that the two models have significantly different model fit.

Assessing the model parameter recovery. To assess the parameter recovery of the continuous parameters (e.g., ability and speed), the bias, the empirical standard error (SE) and the root mean squared error (RMSE) are calculated for each parameter. The bias, the SE and the RMSE are calculated as:

$$Bias(\vartheta) = \frac{1}{R}\sum_{r=1}^{R} \hat{\vartheta} - \vartheta_{true}, \tag{3.3.9}$$

$$SE(\vartheta) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\vartheta} - \frac{\sum_{r=1}^{R}\hat{\vartheta}}{R}\right)^2}, \tag{3.3.10}$$

$$RMSE(\vartheta) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\vartheta} - \vartheta_{true})^2}, \tag{3.3.11}$$

where ϑ is the parameter of interest; ϑ_{true} and $\hat{\vartheta}$ denote the true parameter value and the point estimate, respectively; R is the number of total replications. Posterior mean is used as the point estimate for all continuous parameters. These three measures assess different aspects of the parameter recovery. Specifically, the bias evaluates the systematic errors, while the SE measures the random errors in the parameter estimates. The RMSE, however, assess both the systematic and random errors of the parameter estimates. In fact, the square of the RMSE equals the sum of the square of the bias and the square of the SE, that is,

$$RMSE(\vartheta)^2 = Bias(\vartheta)^2 + SE(\vartheta)^2. \tag{3.3.12}$$

Smaller bias, SE, and RMSE indicate better parameter recovery. These measures are first compared descriptively across different data-fitting models. Then, these measures are analyzed using the ANOVA is used to test the significance of the impact of the manipulated factors.

The standard error biases of the regression coefficient estimates are also obtained from Equation 3.3.9. Specifically, the true standard error value is the standard deviation of the point estimate $\hat{\vartheta}$ across replications, while $\hat{\vartheta}$ is the posterior standard deviation associated with the point estimate in each replication.

For the discrete parameters, including the latent attributes and latent attribute profiles, the classification accuracy is assessed by the attribute correct classification accuracy (ACCR) and the pattern correct classification rate (PCCR). The ACCR is the proportion of correct classification for each attribute in the simulation sample; the PCCR is the proportion of the correct classification of the attribute profile in the simulation sample. The ACCR for each attribute *k* and the PCCR are calculated as:

$$ACCR_k = \frac{1}{J}\sum_{j=1}^{J} I(\hat{\alpha}_{jk} = \alpha_{jk}), \tag{3.3.13}$$

$$PCCR = \frac{1}{J}\sum_{j=1}^{J} I(\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j), \tag{3.3.14}$$

where *J* denotes the simulated sample size; α_{jk} and $\boldsymbol{\alpha}_j$ denote the true attribute mastery status for attribute *k* and true attribute profile for respondent *j*; $I(\hat{\alpha}_{jk} = \alpha_{jk})$ is an indicator function that determines whether the estimated mastery status of the attribute *k* of respondent *j* equal to its corresponding true value; $I(\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j)$ is an indicator function that determines whether the estimated attribute profile of respondent *j* equal to its corresponding true value. Posterior mode is used as the point estimate for all discrete parameters.

In summary, three relative model-fit indices are selected and compared in terms of their performance of identifying the true data generating model. In addition, the bias, the SE, and the RMSE are used to evaluate the parameter recovery of the continuous parameters in the data-fitting models, while the ACCR and PCCR are used to examine the classification accuracy of the latent attributes and attribute profiles.

Inferential statistics on outcome measures. To answer the research questions on how the manipulated factors affected the performance of the proposed model compared with the competing models, the outcome measures of the parameters obtained from all data fitting models are inspected on each level of the manipulated factors. When the number of parameter in the parameter type is no less than 20, the mixed-effect analyses of variance (ANOVA) is to be conducted to examine the statistical and practical significance of the effects of the data fitting model type and its interactions with the manipulated factors. These parameter types include person ability parameter (θ_j, j = 1, 2, 3, …, J), person speed parameter (τ_j, j = 1, 2, 3, …, J), item intercept parameter (β_i, i = 1, 2, 3, …, I), item interaction parameter (δ_i, i = 1, 2, 3, …, I), and item intensity parameter (ζ_i, i = 1, 2, 3, …, I). In the mixed-effect ANOVA, each parameter is treated as a "subject". Data fitting model types are treated as "within-subject factors" and manipulated factors are treated as "between-subject factors". The outcome measures (i.e., bias, SE, RMSE) are taken as repeated measurements on each "subject". Thus, there are $6 \times 24 = 144$ cells of the design in the mixed-effect ANOVAs, each cell contains the parameter estimates from one data-fitting model under one of the 24 simulated conditions.

Three assumptions exist for the mixed-effect ANOVA, including the normality of residuals, the homogeneity of residual variance and sphericity. Normality of residual is assessed by testing whether the outcome measure in each cell is normally distributed with the Shapiro-

Wilk test of normality (Shapiro & Wilk, 1965). Nevertheless, since the *F*-test in the ANOVA is robust to the violation of the normality assumption (Pearson, 1931; Tiku, 1964), the ANOVA will be carried out as long as the nonnormality is not extreme. The homogeneity of residual variances means that the residual variances are equal across groups of the between-subject factors, which is tested by the Levene's test of equality of error variances (Levene, 1960). The violation of the homogeneity of residual variances may affect the Type I error rate of the ANOVA (Box, 1954; Horsnell, 1953). However, the ANOVA was found to be robust to this violation when the sample sizes are approximately equal across groups (Kohr & Games, 1974). It is also suggested to use equal sample size design to prevent the violation of the homogeneity assumption (Maxwell & Delaney, 1990). Therefore, the mixed-effect ANOVAs are carried out separately for simulated conditions of different sample size and test length levels. The sphericity assumption means that the variances of the differences between levels of the within-subject factor are equal. Mauchly's test of sphericity is used to test this assumption (Mauchly, 1940). If the sphericity assumption is violated, the *p*-values in the ANOVA results will be corrected by adjusting the degrees of freedom with the Greenhouse-Geisser Procedure (Greenhouse & Geisser, 1959). In addition, observations are assumed to be identically and independently distributed in ANOVA. The violation of the independence assumption would result in an inflated Type I error rate in the ANOVA (e.g., Kenny & Judd, 1986).

In terms of statistical inference, a *p*-value smaller than 0.05 is considered statistically significant. In addition, the effect size quantified by the partial η^2 (Cohen, 1965) is used as a measure of practical significance of the results. According to Cohen (1988), the criterion of the magnitude of the effect size is: partial $0.01 \leq \eta^2 < 0.06$ for a small effect, $0.06 \leq \eta^2 < 0.14$ for a medium effect, $\eta^2 \geq 0.14$ for a large effect. In the result section, only effects that are both

statistically significant (i.e., $p < 0.05$) and have at least a small effect size ($\eta^2 \geq 0.01$) are reported.

To evaluate the impact of manipulated factors on the parameter recovery of the proposed model JRT-DINA-LLTM, the three-way ANOVA was conducted for the same group of parameter types include person ability parameter (θ_j, j = 1, 2, 3, …, J), person speed parameter (τ_j, j = 1, 2, 3, …, J), item intercept parameter (β_i, i = 1, 2, 3, …, I), item interaction parameter (δ_i, i = 1, 2, 3, …, I), and item intensity parameter (ζ_i, i = 1, 2, 3, …, I). Each manipulated factor is treated as a factor of the ANOVA design. All the assumptions mentioned in the previous paragraph are examined except for the sphericity assumption which is not required in the three-way ANOVA.

3.4 Empirical Data Analysis

3.4.1. Data and Research Questions

The empirical data analysis uses public datasets from the PISA 2012 problem-solving items. This dataset is suitable for the current study because it includes both the item response and response time data from the respondents. In addition, the test design information and sample items are accessible on the OECD website and technical report (OECD, 2014).

Public datasets are available for eleven problem-solving items. According to OECD (2014), these items assess four latent attributes: exploring and understanding (α_1), planning and executing (α_2), monitoring and reflecting (α_3), representing and formulating (α_4). However, based on Gu and Xu (2019), only the two attributes (i.e., α_1, α_2) measured by enough items were used in the data analysis to ensure the identifiability of the DINA model. Therefore, a maximum of 7 items were chosen (i.e., CP002Q07, CP002Q08, CP007Q01, CP007Q02,

CP025Q02, CP038Q01, and CP038Q02) for the current data analysis. The Q-matrix used in the data analysis is shown in Table 8. Respondents with not reached responses (coded as 7) were removed from the sample. Polytomous responses were dichotomized with full credit as correct response (= 1), while partial credit and no credit were coded as incorrect responses (= 0). RT data were obtained from the log files from the same set of problem-solving items. Specifically, item RTs were calculated as the difference between the time of the first action and the time stamp of the last action for each person on each item. Respondents with mistakenly recorded time stamps (negative RTs) were removed. The RTs were originally in seconds and then log-transformed. To create a sample of realistic size, respondents from Australia, Denmark, and Turkey were retained in the final sample (N = 2173). Respondents from these three countries are representative for a sample of a wide spectrum of problem-solving competence according to the performance report from the OECD (2014). In addition, PISA items do not advantage a particular country or language group (OECD, 2014). However, not all respondents answered all 7 items due to the sampling design. The least number of respondents who answered an item was 1067. The missing data mechanism was ignorable and can be handled properly by the Bayesian estimation used in the current study (Gelman et al., 2014).

Table 8. *The Q-matrix Used in the Empirical Data Analysis.*

Items	α_1	α_2
CP002Q07	1	
CP002Q08	1	
CP007Q01		1
CP007Q02		1
CP025Q02		1
CP038Q01	1	
CP038Q02		1

Note. α_1 = exploring and understanding, α_2 = planning and executing. Blank means "0".

The empirical analysis aims to address the following research questions:

1) Do the categorical and continuous item covariates have significant effects on the item parameters from CDM and the RT model, respectively?
2) How is the consistency of the item parameter estimates and attribute estimates among the data-fitting models?

3.4.2. Data Analysis Procedure

3.4.2.1. Item Covariates and Q-matrix

Given the limited number of available items, two item covariates were used in the empirical data analysis as a demonstration of the item explanatory model. Specifically, one dichotomous item was determined based on the test design description in OECD (2014). It describes the *familiarity of information and communication technology* of the items which can be either familiar (= 1) or unfamiliar (= 0). The continuous covariate is the average steps taken for each item. The continuous covariate was standardized with mean 0 and variance 1 to be used in the item explanatory models.

Four models were fit in the dataset: the JRT-DINA-LLTM, the JRT-DINA model, the DINA-LLTM model and the Lognormal-LLTM, and DINA model and the Lognormal model. The prior specifications and Bayesian MCMC estimation setup for these models remain the same as stated in the model parameter estimation section. Several aspects of the data analysis are reported. First, the model fit indices are used to indicate the performance of the proposed and alternative models. In addition, the parameter estimates of the regression coefficients are presented to understand the effects of the item covariates. Lastly, the consistency of the item parameter estimates and the attribute classifications among the data fitting models is examined.

Chapter 4: Simulation Study Results

The simulation study was conducted to address the following research questions: 1) to examine the absolute model fit and relative model fit of the proposed model and competing models; 2) to evaluate the discrepancy of the parameter recovery of the proposed model and competing models; 3) to investigate the impact of the manipulated factors on the parameter recovery of the proposed model. The proposed model and the competing models are summarized in Table 9. The competing models are different from the proposed model (i.e., JRT-DINA-LLTM) in terms of one or more of the following aspects: 1) lack of the second level model; 2) lack of one of the covariates (covariate misspecification); 3) two-step estimation of the regression coefficients of the covariates.

Table 9. *Overview of Model Specifications of the Data-Fitting Model in the Simulation Study.*

Model	Model Characteristics		
	Second-level Model	Covariate Misspecification	Two-step Estimation
JRT-DINA-LLTM	√		
JRT-DINA	√		√
DINA + Lognormal			√
DINA-LLTM + Lognormal-LLTM			
JRT-DINA-LLTM-C	√	√	
JRT-DINA-LLTM-D	√	√	

Note.√ represents presence.

The four manipulated factors in the simulation study include sample size (500, 1000), test length (20, 40), correlation between ability and speed (0.2, 0.5, 0.8), explanatory power of the covariates (small, large), resulting in 24 simulation conditions. Thirty replications were conducted for each condition, resulting in 720 replications. The models were estimated using the Bayesian MCMC method using R2jags software. The iterations run were summarized in Table 2 in the section 3.2. Two MCMC chains were run with a thinning of 2. As a result, all models converged with $\hat{R} < 1.1$ in all replications, which indicates that the chains converged and mixed

well. In addition to convergence, the current study evaluated the precision of the parameter estimation in terms of the effective sample size (ESS). Following Zitzmann and Hecht (2019), ESS larger than 400 was considered to be sufficient in terms of estimation precision for all parameters. The percentages of focal parameters (i.e., person ability, person speed, item intercept, item interaction, item intensity, regression coefficients) with ESS > 400 across replications for all fitted models for each simulation condition are shown in Table 10. It is observed that almost all focal parameters (≥ 95%) achieved satisfactory estimation precision in the Bayesian MCMC estimation. For those parameter estimates with $\hat{R} < 1.1$ but ESS < 400, the chains explored the parameter space rather slowly and more iterations were needed to ensure precise estimation of the posterior means (Zitzmann & Hecht, 2019). Specifically, focal parameter estimates with ESS < 400 mainly included person ability parameter estimates from models JRT-DINA-LLTM, JRT-DINA, DINA, DINA-LLTM, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D especially under conditions where sample size was small, and item intensity parameter estimates and person speed parameter estimates from the Lognormal model under conditions where test length was long. Given that all parameters had $\hat{R} < 1.1$ and the majority of the focal parameters had ESS > 400, the parameter estimation in the simulation study was considered to be satisfactory and all replications were used in the subsequent analysis.

The simulation results are summarized based on the types of parameters presented in Table 11 with respect to the outcome measure (e.g., correct classification rates, bias, SE, and RMSE). Inferential statistics (i.e., ANOVA) are provided for person ability, person speed, item intercept, item interaction, and item intensity parameters. Descriptive statistics (e.g., mean, min, and max) are provided for the other parameters. Detailed information on the outcome measures for each parameter is presented in Appendix A. The remaining part of the simulation results

section is presented in four parts: 1) performance of model fit indices; 2) parameter recovery of person parameters; 3) parameter recovery of item parameters; 4) parameter recovery of higher-order structural parameters.

Table 10. *Average Percentages of Focal Parameters with Effective Sample Sizes > 400 Across Replications.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.999	1.000	1.000	1.000	.999	.995	.998	.999
			Large	1.000	.999	1.000	1.000	.999	1.000	.999	1.000
		0.5	Small	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000
			Large	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	Small	.982	.978	1.000	1.000	1.000	.999	.961	.972
			Large	.988	.982	1.000	1.000	1.000	1.000	.983	.971
	40	0.2	Small	1.000	1.000	.988	1.000	1.000	.998	1.000	1.000
			Large	1.000	1.000	.986	1.000	1.000	1.000	1.000	1.000
		0.5	Small	1.000	1.000	.991	1.000	.999	.998	1.000	1.000
			Large	1.000	1.000	.991	1.000	1.000	1.000	1.000	1.000
		0.8	Small	.996	.999	.998	1.000	.999	.998	.999	.999
			Large	.997	.999	.991	1.000	.999	1.000	.998	.990
1000	20	0.2	Small	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000
			Large	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.5	Small	1.000	1.000	1.000	1.000	.999	.999	1.000	1.000
			Large	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	Small	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000
			Large	.995	.996	1.000	1.000	.999	1.000	.996	.984
	40	0.2	Small	.999	1.000	.965	1.000	.997	.998	.999	.999
			Large	1.000	1.000	.948	1.000	.997	.999	1.000	.999
		0.5	Small	.999	1.000	.989	1.000	.998	.997	.999	.999
			Large	1.000	1.000	1.000	1.000	.999	.999	1.000	1.000
		0.8	Small	.994	.995	1.000	1.000	.998	.998	.995	.996
			Large	1.000	1.000	1.000	1.000	.999	.999	1.000	1.000

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table 11. *Overview of the Model Parameters Evaluated in the Simulation Study.*

Parameter type	Parameter	Description	The presence of parameter in the models					
			JRT-DINA-LLTM	JRT-DINA (two-step)	DINA+ Lognormal (two-step)	DINA+ Lognormal-LLTM	JRT-DINA-LLTM-C	JRT-DINA-LLTM-D
Person parameter	α_{jk}	Attribute mastery status	√	√	√	√	√	√
	θ_j	Ability	√	√	√	√	√	√
	τ_j	Speed	√	√	√	√	√	√
	σ_τ^2	Variance of the speed	√	√	√	√	√	√
	$\sigma_{\theta\tau}$	Covariance between ability and speed	√	√	×	×	√	√
Item parameter	β_i	Item intercept	√	√	√	√	√	√
	δ_i	Item interaction	√	√	√	√	√	√
	ζ_i	Item intensity	√	√	√	√	√	√
	A_0	Regression intercepts	√	√	√	√	√	√
	A_1	Regression slopes for the continuous covariate	√	√	√	√	√	×
	A_2	Regression slopes for the dichotomous covariate	√	√	√	√	×	√
Higher-order structural parameter	γ_k	Attribute slope	√	√	√	√	√	√
	λ_k	Attribute intercept	√	√	√	√	√	√

Note. √ indicates the presence of the parameter; × indicates the absence of the parameter

4.1 Model-data Fit Evaluations

The model-data fit of the proposed model and competing models are evaluated in terms of absolute model fit indicated by the PPP-values obtained from the PPMC method and relative model fit indicated by AIC, BIC, and DIC. PPP-value is the probability that the replicated data could be more extreme than the observed data, which is quantified by the sum of squared residuals for the response model and RT model, respectively. Specifically, PPP-values in the range of 0.05 to 0.95 indicate adequate absolute model fit, while smaller relative model data fit indices indicate better model-data fit compared to competing models.

The descriptive statistics of the distribution of PPP-values of the fitted models across 30 replications are presented in Tables 12 to 17. For all models, the PPP-values fall in the range of 0.3 to 0.7 for the response model and fall in the range of 0.4 to 0.6 for the RT model. Therefore, all fitted models had a satisfactory absolute model fit from the sum of squares of residuals perspective. In the simulation study, the JRT-DINA-LLTM was the true data generating model. Therefore, it is expected that the replicated data is not extreme compared to the observed data if the true data generating model is JRT-DINA-LLTM. However, all competing models had a similar range of PPP-values for both the response model and RT model. This indicates that the replicated data using competing models possess similar characteristics as the observed data generated by the proposed model.

Table 12. *Summary of the Posterior Predictive P-Values of the JRT-DINA-LLTM under the 24 Simulated Conditions*

Condition No.	J	I	$\rho_{\theta\tau}$	γ	PPP value for response model				PPP value for RT model			
					Min	Median	Mean	Max	Min	Median	Mean	Max
1	500	20	0.2	Small	0.34	0.41	0.42	0.52	0.50	0.51	0.51	0.52
2				Large	0.31	0.38	0.38	0.45	0.51	0.51	0.51	0.52
3			0.5	Small	0.36	0.43	0.44	0.53	0.51	0.51	0.51	0.52
4				Large	0.36	0.39	0.39	0.46	0.50	0.51	0.51	0.52
5			0.8	Small	0.36	0.42	0.43	0.57	0.51	0.51	0.51	0.52
6				Large	0.34	0.38	0.39	0.47	0.51	0.51	0.51	0.52
7		40	0.2	Small	0.34	0.48	0.48	0.57	0.50	0.51	0.51	0.52
8				Large	0.38	0.45	0.45	0.61	0.50	0.51	0.51	0.52
9			0.5	Small	0.38	0.47	0.47	0.56	0.50	0.51	0.51	0.52
10				Large	0.38	0.46	0.46	0.62	0.50	0.51	0.51	0.52
11			0.8	Small	0.35	0.48	0.47	0.58	0.50	0.51	0.51	0.52
12				Large	0.35	0.47	0.46	0.55	0.50	0.51	0.51	0.52
13	1000	20	0.2	Small	0.39	0.43	0.44	0.54	0.50	0.51	0.51	0.52
14				Large	0.39	0.41	0.42	0.47	0.50	0.51	0.51	0.52
15			0.5	Small	0.40	0.45	0.45	0.52	0.50	0.51	0.51	0.52
16				Large	0.37	0.41	0.41	0.44	0.50	0.51	0.51	0.52
17			0.8	Small	0.40	0.44	0.44	0.50	0.50	0.51	0.51	0.52
18				Large	0.38	0.40	0.41	0.44	0.50	0.51	0.51	0.52
19		40	0.2	Small	0.40	0.53	0.52	0.65	0.50	0.51	0.51	0.52
20				Large	0.41	0.46	0.46	0.60	0.50	0.51	0.51	0.52
21			0.5	Small	0.44	0.52	0.52	0.67	0.49	0.51	0.51	0.52
22				Large	0.40	0.46	0.46	0.59	0.50	0.51	0.51	0.52
23			0.8	Small	0.44	0.52	0.52	0.61	0.50	0.51	0.51	0.52
24				Large	0.39	0.46	0.46	0.66	0.49	0.51	0.51	0.52

Note. PPP = Posterior Predictive P-Values.

Table 13. *Summary of the Posterior Predictive P-Values of the JRT-DINA under the 24 Simulated Conditions*

Condition No.	J	I	$\rho_{\theta\tau}$	γ	PPP value for response model				PPP value for RT model			
					Min	Median	Mean	Max	Min	Median	Mean	Max
1	500	20	0.2	Small	0.40	0.48	0.49	0.60	0.51	0.51	0.51	0.52
2				Large	0.42	0.47	0.47	0.51	0.51	0.51	0.51	0.52
3			0.5	Small	0.42	0.48	0.49	0.61	0.50	0.51	0.51	0.52
4				Large	0.45	0.48	0.49	0.55	0.50	0.51	0.51	0.52
5			0.8	Small	0.41	0.49	0.50	0.64	0.50	0.51	0.51	0.52
6				Large	0.43	0.48	0.48	0.55	0.51	0.51	0.51	0.52
7		40	0.2	Small	0.35	0.50	0.50	0.59	0.50	0.51	0.51	0.51
8				Large	0.46	0.53	0.54	0.61	0.50	0.51	0.51	0.52
9			0.5	Small	0.39	0.50	0.49	0.59	0.50	0.51	0.51	0.52
10				Large	0.48	0.54	0.54	0.69	0.50	0.51	0.51	0.52
11			0.8	Small	0.38	0.51	0.50	0.59	0.50	0.51	0.51	0.52
12				Large	0.44	0.55	0.54	0.61	0.50	0.51	0.51	0.51
13	1000	20	0.2	Small	0.44	0.48	0.49	0.58	0.50	0.51	0.51	0.52
14				Large	0.46	0.49	0.49	0.54	0.50	0.51	0.51	0.51
15			0.5	Small	0.44	0.50	0.50	0.56	0.50	0.51	0.51	0.52
16				Large	0.44	0.49	0.49	0.51	0.50	0.51	0.51	0.52
17			0.8	Small	0.44	0.49	0.49	0.56	0.50	0.51	0.51	0.52
18				Large	0.45	0.48	0.48	0.52	0.50	0.51	0.51	0.52
19		40	0.2	Small	0.43	0.56	0.55	0.68	0.50	0.51	0.51	0.51
20				Large	0.48	0.52	0.53	0.61	0.49	0.51	0.51	0.51
21			0.5	Small	0.46	0.54	0.55	0.67	0.49	0.51	0.51	0.52
22				Large	0.49	0.52	0.53	0.61	0.50	0.51	0.51	0.52
23			0.8	Small	0.46	0.55	0.55	0.63	0.49	0.50	0.50	0.51
24				Large	0.47	0.52	0.52	0.65	0.50	0.50	0.51	0.51

Note. PPP = Posterior Predictive P-Values.

Table 14. *Summary of the Posterior Predictive P-Values of the DINA+Lognormal under the 24 Simulated Conditions*

Condition No.	J	I	$\rho_{\theta\tau}$	γ	PPP value for response model				PPP value for RT model			
					Min	Median	Mean	Max	Min	Median	Mean	Max
1	500	20	0.2	Small	0.38	0.45	0.46	0.58	0.52	0.52	0.52	0.52
2				Large	0.42	0.47	0.47	0.51	0.52	0.52	0.52	0.52
3			0.5	Small	0.40	0.46	0.47	0.58	0.52	0.52	0.52	0.52
4				Large	0.45	0.48	0.48	0.54	0.52	0.52	0.52	0.52
5			0.8	Small	0.39	0.47	0.48	0.63	0.52	0.52	0.52	0.52
6				Large	0.43	0.47	0.48	0.55	0.52	0.52	0.52	0.52
7		40	0.2	Small	0.37	0.49	0.50	0.60	0.50	0.50	0.50	0.50
8				Large	0.45	0.49	0.50	0.56	0.50	0.50	0.50	0.50
9			0.5	Small	0.37	0.50	0.49	0.58	0.50	0.50	0.50	0.50
10				Large	0.45	0.50	0.50	0.66	0.50	0.50	0.50	0.50
11			0.8	Small	0.36	0.50	0.49	0.59	0.50	0.50	0.50	0.50
12				Large	0.43	0.50	0.50	0.56	0.50	0.50	0.50	0.50
13	1000	20	0.2	Small	0.42	0.46	0.46	0.57	0.51	0.51	0.51	0.52
14				Large	0.46	0.49	0.49	0.53	0.51	0.51	0.51	0.51
15			0.5	Small	0.40	0.47	0.48	0.56	0.51	0.51	0.51	0.51
16				Large	0.45	0.49	0.49	0.51	0.51	0.51	0.51	0.51
17			0.8	Small	0.43	0.46	0.47	0.53	0.51	0.51	0.51	0.52
18				Large	0.46	0.47	0.48	0.52	0.52	0.52	0.52	0.52
19		40	0.2	Small	0.44	0.55	0.54	0.69	0.50	0.50	0.50	0.51
20				Large	0.44	0.50	0.49	0.57	0.50	0.51	0.51	0.51
21			0.5	Small	0.44	0.53	0.54	0.66	0.50	0.50	0.50	0.51
22				Large	0.45	0.49	0.49	0.56	0.50	0.50	0.50	0.50
23			0.8	Small	0.47	0.54	0.54	0.62	0.50	0.50	0.50	0.50
24				Large	0.46	0.49	0.50	0.58	0.50	0.50	0.50	0.50

Note. PPP = Posterior Predictive P-Values.

Table 15. *Summary of the Posterior Predictive P-Values of the DINA-LLTM+Lognormal-LLTM under the 24 Simulated Conditions*

Condition No.	J	I	$\rho_{\theta\tau}$	γ	PPP value for response model				PPP value for RT model			
					Min	Median	Mean	Max	Min	Median	Mean	Max
1	500	20	0.2	Small	0.35	0.42	0.42	0.53	0.50	0.51	0.51	0.52
2				Large	0.32	0.38	0.38	0.44	0.50	0.51	0.51	0.52
3			0.5	Small	0.37	0.42	0.43	0.52	0.51	0.51	0.51	0.52
4				Large	0.36	0.39	0.39	0.46	0.50	0.52	0.52	0.53
5			0.8	Small	0.35	0.42	0.43	0.56	0.50	0.51	0.51	0.53
6				Large	0.34	0.39	0.39	0.48	0.50	0.51	0.51	0.53
7		40	0.2	Small	0.34	0.49	0.48	0.56	0.50	0.51	0.51	0.51
8				Large	0.36	0.45	0.45	0.60	0.50	0.51	0.51	0.51
9			0.5	Small	0.38	0.47	0.46	0.55	0.50	0.51	0.51	0.52
10				Large	0.38	0.45	0.46	0.62	0.50	0.51	0.51	0.52
11			0.8	Small	0.36	0.47	0.47	0.57	0.50	0.50	0.51	0.51
12				Large	0.34	0.46	0.46	0.54	0.50	0.51	0.51	0.51
13	1000	20	0.2	Small	0.39	0.43	0.44	0.54	0.50	0.50	0.50	0.51
14				Large	0.38	0.41	0.41	0.48	0.50	0.50	0.50	0.51
15			0.5	Small	0.40	0.45	0.45	0.52	0.50	0.50	0.50	0.52
16				Large	0.37	0.41	0.41	0.44	0.49	0.50	0.50	0.51
17			0.8	Small	0.41	0.44	0.45	0.51	0.49	0.50	0.51	0.51
18				Large	0.38	0.40	0.40	0.44	0.49	0.50	0.50	0.51
19		40	0.2	Small	0.40	0.53	0.52	0.67	0.50	0.50	0.51	0.52
20				Large	0.40	0.46	0.46	0.60	0.49	0.51	0.51	0.51
21			0.5	Small	0.44	0.52	0.52	0.66	0.49	0.50	0.50	0.52
22				Large	0.39	0.45	0.46	0.59	0.49	0.50	0.50	0.51
23			0.8	Small	0.43	0.52	0.52	0.60	0.50	0.50	0.50	0.51
24				Large	0.39	0.46	0.46	0.64	0.50	0.50	0.50	0.51

Note. PPP = Posterior Predictive P-Values.

Table 16. *Summary of the Posterior Predictive P-Values of the JRT-DINA-LLTM-C under the 24 Simulated Conditions*

Condition No.	J	I	$\rho_{\theta\tau}$	γ	PPP value for response model				PPP value for RT model			
					Min	Median	Mean	Max	Min	Median	Mean	Max
1	500	20	0.2	Small	0.37	0.44	0.44	0.55	0.51	0.51	0.51	0.52
2				Large	0.33	0.40	0.40	0.45	0.51	0.51	0.51	0.52
3			0.5	Small	0.38	0.44	0.45	0.55	0.51	0.51	0.51	0.52
4				Large	0.38	0.41	0.41	0.48	0.51	0.51	0.51	0.52
5			0.8	Small	0.38	0.44	0.45	0.59	0.50	0.51	0.51	0.52
6				Large	0.36	0.40	0.40	0.49	0.51	0.51	0.51	0.52
7		40	0.2	Small	0.34	0.49	0.48	0.57	0.50	0.51	0.51	0.52
8				Large	0.40	0.47	0.48	0.60	0.50	0.51	0.51	0.52
9			0.5	Small	0.38	0.47	0.47	0.56	0.50	0.51	0.51	0.52
10				Large	0.41	0.49	0.49	0.62	0.50	0.51	0.51	0.52
11			0.8	Small	0.35	0.47	0.47	0.57	0.50	0.51	0.51	0.52
12				Large	0.38	0.50	0.49	0.58	0.50	0.51	0.51	0.52
13	1000	20	0.2	Small	0.40	0.44	0.45	0.56	0.50	0.51	0.51	0.52
14				Large	0.40	0.43	0.43	0.48	0.50	0.51	0.51	0.52
15			0.5	Small	0.41	0.45	0.46	0.51	0.50	0.51	0.51	0.52
16				Large	0.39	0.43	0.43	0.46	0.51	0.51	0.51	0.52
17			0.8	Small	0.41	0.45	0.45	0.51	0.50	0.51	0.51	0.52
18				Large	0.39	0.42	0.42	0.46	0.50	0.51	0.51	0.52
19		40	0.2	Small	0.41	0.53	0.52	0.65	0.50	0.51	0.51	0.52
20				Large	0.44	0.48	0.49	0.60	0.49	0.51	0.50	0.51
21			0.5	Small	0.45	0.51	0.52	0.64	0.50	0.51	0.51	0.52
22				Large	0.42	0.49	0.49	0.59	0.50	0.51	0.51	0.51
23			0.8	Small	0.44	0.52	0.52	0.61	0.50	0.51	0.51	0.52
24				Large	0.42	0.48	0.49	0.64	0.50	0.51	0.51	0.52

Note. PPP = Posterior Predictive P-Values.

Table 17. *Summary of the Posterior Predictive P-Values of the JRT-DINA-LLTM-D under the 24 Simulated Conditions*

Condition No.	J	I	$\rho_{\theta\tau}$	γ	PPP value for response model				PPP value for RT model			
					Min	Median	Mean	Max	Min	Median	Mean	Max
1	500	20	0.2	Small	0.38	0.45	0.45	0.56	0.51	0.51	0.51	0.52
2				Large	0.38	0.43	0.43	0.48	0.50	0.51	0.51	0.52
3			0.5	Small	0.39	0.44	0.46	0.56	0.51	0.51	0.51	0.52
4				Large	0.39	0.44	0.43	0.48	0.50	0.51	0.51	0.53
5			0.8	Small	0.38	0.45	0.46	0.60	0.51	0.51	0.51	0.52
6				Large	0.37	0.43	0.43	0.51	0.51	0.51	0.51	0.52
7		40	0.2	Small	0.35	0.49	0.49	0.58	0.50	0.51	0.51	0.52
8				Large	0.39	0.45	0.46	0.57	0.50	0.51	0.51	0.52
9			0.5	Small	0.38	0.48	0.48	0.58	0.50	0.51	0.51	0.52
10				Large	0.39	0.45	0.46	0.57	0.50	0.51	0.51	0.52
11			0.8	Small	0.36	0.49	0.49	0.60	0.50	0.51	0.51	0.52
12				Large	0.38	0.46	0.46	0.56	0.50	0.51	0.51	0.52
13	1000	20	0.2	Small	0.41	0.44	0.45	0.55	0.50	0.51	0.51	0.51
14				Large	0.42	0.45	0.45	0.50	0.50	0.51	0.51	0.52
15			0.5	Small	0.41	0.45	0.46	0.52	0.50	0.51	0.51	0.52
16				Large	0.41	0.45	0.45	0.49	0.50	0.51	0.51	0.52
17			0.8	Small	0.41	0.45	0.45	0.52	0.50	0.51	0.51	0.52
18				Large	0.43	0.45	0.45	0.49	0.50	0.51	0.51	0.52
19		40	0.2	Small	0.42	0.54	0.53	0.67	0.50	0.51	0.51	0.52
20				Large	0.43	0.46	0.47	0.58	0.50	0.51	0.51	0.52
21			0.5	Small	0.45	0.52	0.53	0.66	0.50	0.51	0.51	0.51
22				Large	0.43	0.46	0.47	0.54	0.50	0.51	0.51	0.53
23			0.8	Small	0.44	0.53	0.53	0.61	0.50	0.51	0.51	0.52
24				Large	0.42	0.46	0.47	0.58	0.50	0.51	0.51	0.52

Note. PPP = Posterior Predictive P-Values.

In addition to the absolute model fit, the relative model fit indicated by AIC, BIC, and DIC was also examined to see whether these model fit indices can correctly identify the true data generating model (i.e., JRT-DINA-LLTM). The frequencies that each data-generating model was identified as the best data-fitting model in 30 replications in each condition are reported in Tables 18 to 20. Table 18 includes four joint models (i.e., JRT-DINA-LLTM, JRT-DINA, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D) fitted on both response and RT datasets. Table 19 includes DINA and DINA-LLTM which were fitted on the response datasets. Table 20 includes lognormal and lognormal-LLTM which were fitted on the RT datasets. In Table 18, AIC and BIC chose either JRT-DINA-LLTM-C or JRT-DINA-LLTM-D in almost all replications. A few exceptions that AIC and BIC chose JRT-DINA-LLTM are when the test length is long and/or the regression coefficients are large. A possible reason is that AIC and BIC penalize complex models and tend to choose simpler models. The results of DIC, on the other hand, were mixed. Among the 30 replications in each condition, DIC chose various different models but an observed pattern is that it tended to choose JRT-DINA-LLTM when test length is long and regression coefficients are large. In Table 19, AIC and BIC chose DINA-LLTM over DINA model in all 30 replications in all conditions, while DIC tended to choose DINA-LLTM in conditions where test length is long and regression coefficients are large. In Table 20, AIC, BIC and DIC chose the Lognormal-LLTM over the Lognormal model as the best-fitting model on RT datasets. In sum, all three relative model fit indices cannot correctly identify the true model, JRT-DINA-LLTM as the best fitting model in joint modeling. When the second level model was omitted, AIC and BIC can correctly identify models with covariates (i.e., DINA-LLTM and Lognormal-LLTM), while DIC can identify Lognormal-LLTM but only tended to choose DINA-LLTM when test length is long and/or regression coefficients are large.

Table 18. *The Number of Replications of Each Joint Model Identified as the Best-Fitting Model in the Simulation Study*

				AIC				BIC				DIC			
J	I	$\rho_{\theta\tau}$	γ	JRT-DINA-LLTM	JRT-DINA	JRT-DINA-LLTM-C	JRT-DINA-LLTM-D	JRT-DINA-LLTM	JRT-DINA	JRT-DINA-LLTM-C	JRT-DINA-LLTM-D	JRT-DINA-LLTM	JRT-DINA	JRT-DINA-LLTM-C	JRT-DINA-LLTM-D
500	20	0.2	Small	0	0	12	18	0	0	12	18	3	14	10	3
			Large	0	0	25	5	0	0	25	5	13	2	13	2
		0.5	Small	0	0	11	19	0	0	11	19	4	10	6	10
			Large	0	0	23	7	0	0	23	7	11	2	14	3
		0.8	Small	0	0	16	14	0	0	16	14	9	7	6	8
			Large	0	0	23	7	0	0	23	7	11	2	15	2
	40	0.2	Small	0	0	16	14	0	0	16	14	6	5	11	8
			Large	9	0	21	0	0	0	30	0	23	1	6	0
		0.5	Small	0	0	15	15	0	0	15	15	12	2	6	10
			Large	5	0	25	0	0	0	30	0	18	3	9	0
		0.8	Small	0	0	17	13	0	0	17	13	12	5	9	4
			Large	6	0	24	0	0	0	29	1	14	1	15	0
1000	20	0.2	Small	0	0	6	24	0	0	6	24	3	14	7	6
			Large	0	0	17	13	0	0	17	13	6	6	10	8
		0.5	Small	0	0	6	24	0	0	6	24	5	6	10	9
			Large	0	0	19	11	0	0	19	11	12	3	9	6
		0.8	Small	0	0	8	22	0	0	8	22	5	14	4	7
			Large	1	0	18	11	1	0	18	11	8	2	10	10
	40	0.2	Small	0	0	13	17	0	0	13	17	11	4	9	6
			Large	0	0	30	0	0	0	30	0	13	6	9	2
		0.5	Small	0	0	18	12	0	0	18	12	4	6	9	11
			Large	0	0	30	0	0	0	30	0	11	4	14	1
		0.8	Small	0	0	12	18	0	0	12	18	6	7	6	11
			Large	1	0	29	0	0	0	30	0	15	2	10	3

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients

Table 19. *The Number of Replications of Each Response Model Identified as the Best-Fitting Model in the Simulation Study*

J	I	$\rho_{\theta\tau}$	γ	AIC		BIC		DIC	
				DINA-LLTM	DINA	DINA-LLTM	DINA	DINA-LLTM	DINA
500	20	0.2	Small	30	0	30	0	3	27
			Large	30	0	30	0	27	3
		0.5	Small	30	0	30	0	7	23
			Large	30	0	30	0	21	9
		0.8	Small	30	0	30	0	3	27
			Large	30	0	30	0	25	5
	40	0.2	Small	30	0	30	0	23	7
			Large	30	0	30	0	29	1
		0.5	Small	30	0	30	0	26	4
			Large	30	0	30	0	30	0
		0.8	Small	30	0	30	0	24	6
			Large	30	0	30	0	27	3
1000	20	0.2	Small	30	0	30	0	12	18
			Large	30	0	30	0	13	17
		0.5	Small	30	0	30	0	7	23
			Large	30	0	30	0	20	10
		0.8	Small	30	0	30	0	11	19
			Large	30	0	30	0	19	11
	40	0.2	Small	30	0	30	0	15	15
			Large	30	0	30	0	15	15
		0.5	Small	30	0	30	0	22	8
			Large	30	0	30	0	24	6
		0.8	Small	30	0	30	0	17	13
			Large	30	0	30	0	23	7

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table 20. *The Number of Replications of Each Response Time Model Identified as the Best-Fitting Model in the Simulation Study*

J	I	$\rho_{\theta\tau}$	γ	AIC		BIC		DIC	
				Lognormal-LLTM	Lognormal	Lognormal-LLTM	Lognormal	Lognormal-LLTM	Lognormal
500	20	0.2	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.5	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.8	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
	40	0.2	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.5	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.8	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
1000	20	0.2	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.5	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.8	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
	40	0.2	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.5	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0
		0.8	Small	30	0	30	0	30	0
			Large	30	0	30	0	30	0

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Although it is convenient to compare the magnitude of relative model fit indices, it is also necessary to examine whether the discrepancy of model fit indices is significantly large. The evidence ratio (Anderson, 2008) was used to determine whether there is strong evidence that the best fitting model had better model fit than other models. Specifically, an evidence ratio larger than 55 suggested that the two models are evidently different (Anderson, 2008).

Table 21 presents the counts of evidence ratio > 55 for each joint model. The reference model in each replication was the best data-fitting model identified by the model fit indices as shown in Table 18. Several conclusions are made based on the findings shown in Table 21. First, evidence ratios calculated based on AIC suggested that there is strong evidence that JRT-DINA-LLTM-C and JRT-DINA-LLTM-D had better model fit than JRT-DINA but not JRT-DINA-LLTM. Second, evidence ratios calculated based on BIC suggested that there is strong evidence that JRT-DINA-LLTM-C and JRT-DINA-LLTM-D had better model fit than both JRT-DINA and JRT-DINA-LLTM. A possible reason is that BIC has more penalty on the complexity of the model than AIC does. Third, DIC still showed a mixed result and did not show any evidence on which model had a better model fit.

Table 22 presents the counts of evidence ratio > 55 when there is strong evidence that DINA/ DINA-LLTM had a better model fit. When model fit indices were AIC and BIC, the reference model was the DINA-LLTM, while the reference model was the one identified by DIC in each replication when the model fit index was DIC, as shown in Table 20. As a result, AIC and BIC consistently indicated that DINA-LLTM had a better model fit than DINA in all replications, while DIC cannot differentiate the two models and tended to choose DINA-LLTM when test length was long and/or regression coefficients were large. Table 23 presents the counts of evidence ratio > 55 when there is strong evidence that Lognormal-LLTM had a better model

fit than the Lognormal model. The reference model was Lognormal-LLTM for all replications. There is strong evidence that all three model fit indices showed better model fit for the Lognormal-LLTM.

In conclusion, AIC and BIC cannot correctly identify the true data generating model (i.e., JRT-DINA-LLTM), although AIC did not show a significant different model fit between the JRT-DINA-LLTM and the best data-fitting model it chose. For models omitting the second level structure, however, AIC and BIC correctly identified models with covariates (i.e., DINA-LLTM and Lognormal-LLTM). DIC, on the other hand, cannot correctly identify JRT-DINA-LLTM and DINA-LLTM but can correctly identify Lognormal-LLTM.

Table 21. *The Number of Replications of the Evidence Ratio of the Joint Models Being Greater than 55.*

				AIC				BIC				DIC			
J	I	$\rho_{\theta\tau}$	γ	JRT-DINA-LLTM	JRT-DINA	JRT-DINA-LLTM-C	JRT-DINA-LLTM-D	JRT-DINA-LLTM	JRT-DINA	JRT-DINA-LLTM-C	JRT-DINA-LLTM-D	JRT-DINA-LLTM	JRT-DINA	JRT-DINA-LLTM-C	JRT-DINA-LLTM-D
500	20	0.2	Small	0	30	0	0	30	30	0	0	25	9	17	23
			Large	0	30	0	0	30	30	0	0	10	27	7	22
		0.5	Small	0	30	0	0	30	30	0	0	21	13	18	15
			Large	0	30	0	0	30	30	0	0	10	23	9	19
		0.8	Small	0	30	0	0	30	30	0	0	16	13	20	17
			Large	0	30	0	0	30	30	0	0	12	25	7	19
	40	0.2	Small	0	30	0	0	30	30	0	0	17	19	13	17
			Large	0	30	0	5	30	30	0	3	6	23	12	29
		0.5	Small	0	30	0	0	30	30	0	0	8	23	20	11
			Large	0	30	0	6	30	30	0	4	4	25	17	29
		0.8	Small	0	30	0	0	30	30	0	0	10	17	13	17
			Large	1	30	1	8	29	30	1	7	11	26	9	30
1000	20	0.2	Small	0	30	0	0	30	30	0	0	25	14	21	19
			Large	1	30	0	0	30	30	0	0	23	22	18	20
		0.5	Small	0	30	0	0	30	30	0	0	25	21	17	17
			Large	1	30	0	0	30	30	0	0	17	23	15	22
		0.8	Small	0	30	0	0	30	30	0	0	24	15	25	22
			Large	0	30	1	1	29	30	1	1	21	26	19	18
	40	0.2	Small	0	30	0	0	30	30	0	0	16	22	18	23
			Large	0	30	0	1	30	30	0	1	12	21	19	24
		0.5	Small	0	30	0	0	30	30	0	0	24	19	15	19
			Large	2	30	0	6	30	30	0	6	15	23	12	25
		0.8	Small	0	30	0	0	30	30	0	0	23	21	19	17
			Large	0	30	0	5	30	30	0	5	13	25	14	27

Note. The number in each cell indicates the times each model was considered as significantly worse than the best fitting model in each condition. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients. The reference model was the one identified by each model fit index in each replication as shown in Table 18.

Table 22. *The Number of Replications of the Evidence Ratio of the Response Models Being Greater than 55.*

J	I	$\rho_{\theta\tau}$	γ	AIC		BIC		DIC	
				DINA-LLTM	DINA	DINA-LLTM	DINA	DINA-LLTM	DINA
500	20	0.2	Small	0	30	0	30	25	2
			Large	0	30	0	30	1	20
		0.5	Small	0	30	0	30	11	5
			Large	0	30	0	30	3	13
		0.8	Small	0	30	0	30	16	1
			Large	0	30	0	30	2	16
	40	0.2	Small	0	30	0	30	0	5
			Large	0	30	0	30	0	21
		0.5	Small	0	30	0	30	1	7
			Large	0	30	0	30	0	23
		0.8	Small	0	30	0	30	0	4
			Large	0	30	0	30	0	24
1000	20	0.2	Small	0	30	0	30	14	5
			Large	0	30	0	30	8	11
		0.5	Small	0	30	0	30	16	4
			Large	0	30	0	30	7	13
		0.8	Small	0	30	0	30	19	4
			Large	0	30	0	30	7	14
	40	0.2	Small	0	30	0	30	0	5
			Large	0	30	0	30	3	7
		0.5	Small	0	30	0	30	3	12
			Large	0	30	0	30	2	20
		0.8	Small	0	30	0	30	7	5
			Large	0	30	0	30	2	17

Note. The number in each cell indicates the times each model was considered as significantly worse than the best fitting model in each condition. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients. The reference model was DINA-LLTM for AIC and BIC; the reference model was the one identified by DIC in each replication for DIC as shown in Table 19.

Table 23. *The Number of Replications of the Evidence Ratio of the Response Time Models Being Greater than 55.*

J	I	$\rho_{\theta\tau}$	γ	AIC		BIC		DIC	
				Lognormal-LLTM	Lognormal	Lognormal-LLTM	Lognormal	Lognormal-LLTM	Lognormal
500	20	0.2	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.5	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.8	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
	40	0.2	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.5	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.8	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
1000	20	0.2	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.5	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.8	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
	40	0.2	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.5	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30
		0.8	Small	0	30	0	30	0	30
			Large	0	30	0	30	0	30

Note. The number in each cell indicates the times each model was considered as significantly worse than the best fitting model in each condition. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients. The reference model was Lognormal-LLTM for all three model fit indices.

4.2 Recovery of the Person Parameters

The person parameters investigated in the simulation study include four latent attributes (α), latent ability parameters (θ), latent speed parameters (τ), the variance of latent speed (σ_τ^2) and correlation between ability and speed ($\rho_{\theta\tau}$). Specifically, the four latent attributes are all binary variables with 1 indicating mastery and 0 otherwise. Latent ability and latent speed are continuous parameters which indicate the proficiency and work speed levels of the respondents. The correlation between ability and speed is a parameter only existing in the four joint models (i.e., JRT-DINA-LLTM, JRT-DINA, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D), while the variance of latent speed is a parameter in the RT models. Among these parameters, the estimation errors in the latent ability and latent speed parameters were analyzed using the mixed-effect ANOVA to examine the impact of manipulated factors including the models fitted to the data on the model parameter recovery. The mixed-effect ANOVA was conducted separately for conditions where sample size was 500 and sample size was 1000 because mixed-effect ANOVA results from groups with equal sample sizes are robust to the violation of the homogeneous residual variance assumption. In addition, a four-way ANOVA was conducted to examine the impact of manipulated factors on the parameter recovery of latent ability and latent speed in the proposed model JRT-DINA-LLTM.

4.2.1 Attribute mastery status

Latent attributes are discrete parameters. Therefore, the attribute correct classification rates (ACCR) and the attribute pattern correct classification rates (PCCR) were calculated across replications for each simulated condition. To examine the impact of manipulated factors on ACCRs and PCCRs, the marginal means of the ACCRs and PCCRs of each model with respect to each manipulated factor are presented in Figures 4 and 5, respectively. Among the six models,

the two models omitting the second level structure, i.e., DINA and DINA-LLTM, had slightly lower ACCRs than the four joint models especially when test length is short (= 20), sample size is small (= 500), explanatory power of the covariates is small, and the correlation between ability and speed is large (= 0.8). For all models, longer test and larger explanatory power of the item covariates led to higher ACCRs. The increase of sample size led to small increase of the ACCRs for all models. The increase of correlation between ability and speed, especially from 0.2 to 0.5, led to some increase of the ACCRs of the joint models. The ACCRs of each attribute for each simulated condition are presented in Tables A1 to A4 in Appendix A.

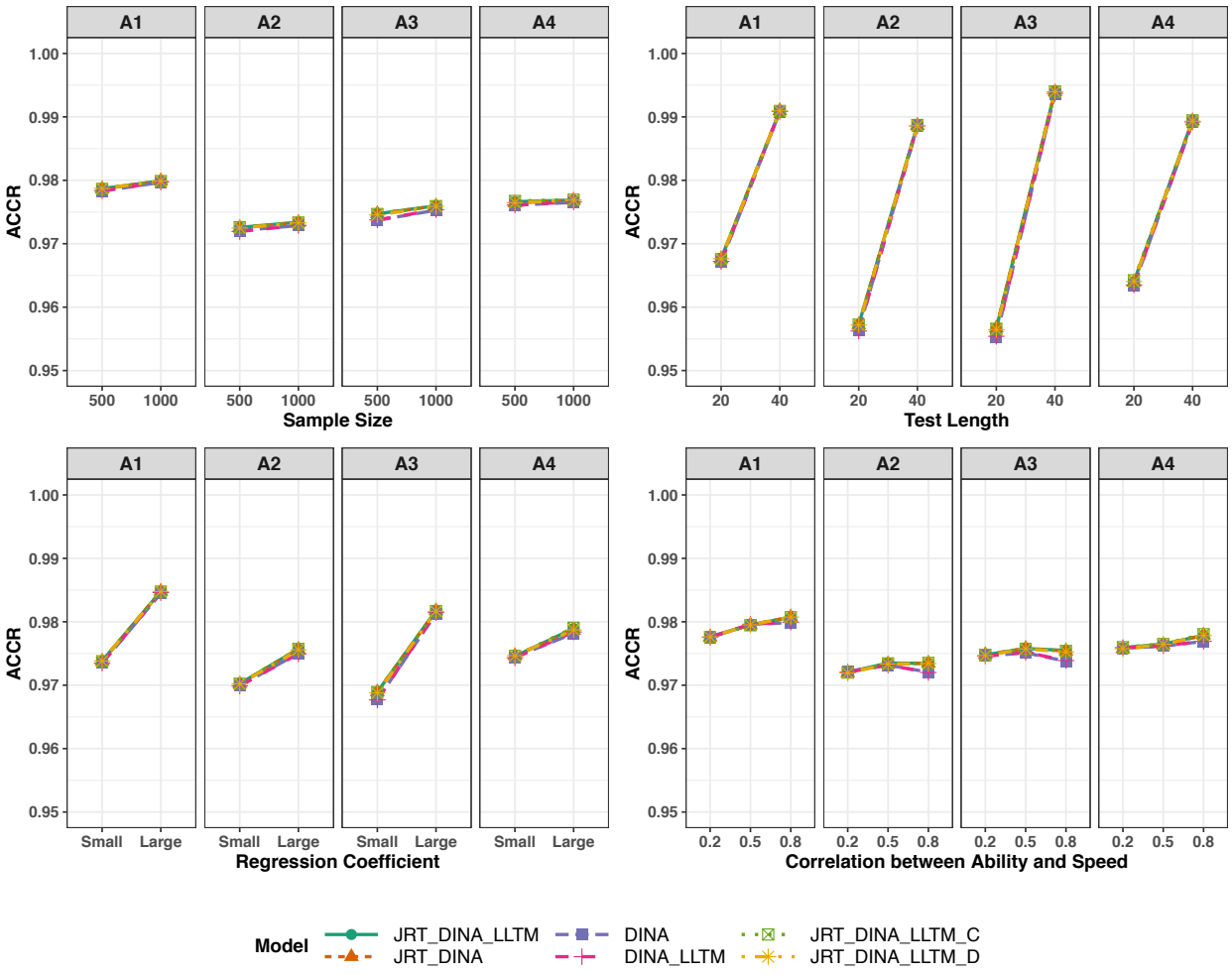


Figure 4. Marginal mean attribute correct classification rates (ACCRs) at each level of the manipulated factors. A1 to A4 indicate Attribute 1 to Attribute 4.

Similarly, the two models omitting the second level structure (i.e., DINA and DINA-LLTM) had slightly lower PCCRs than the four joint models (i.e., JRT-DINA-LLTM, JRT-DINA, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D) especially when the test is long and correlation between ability and speed is large as shown in Figure 5. In addition, long test and large explanatory power of item covariates led to larger PCCRs for all models.

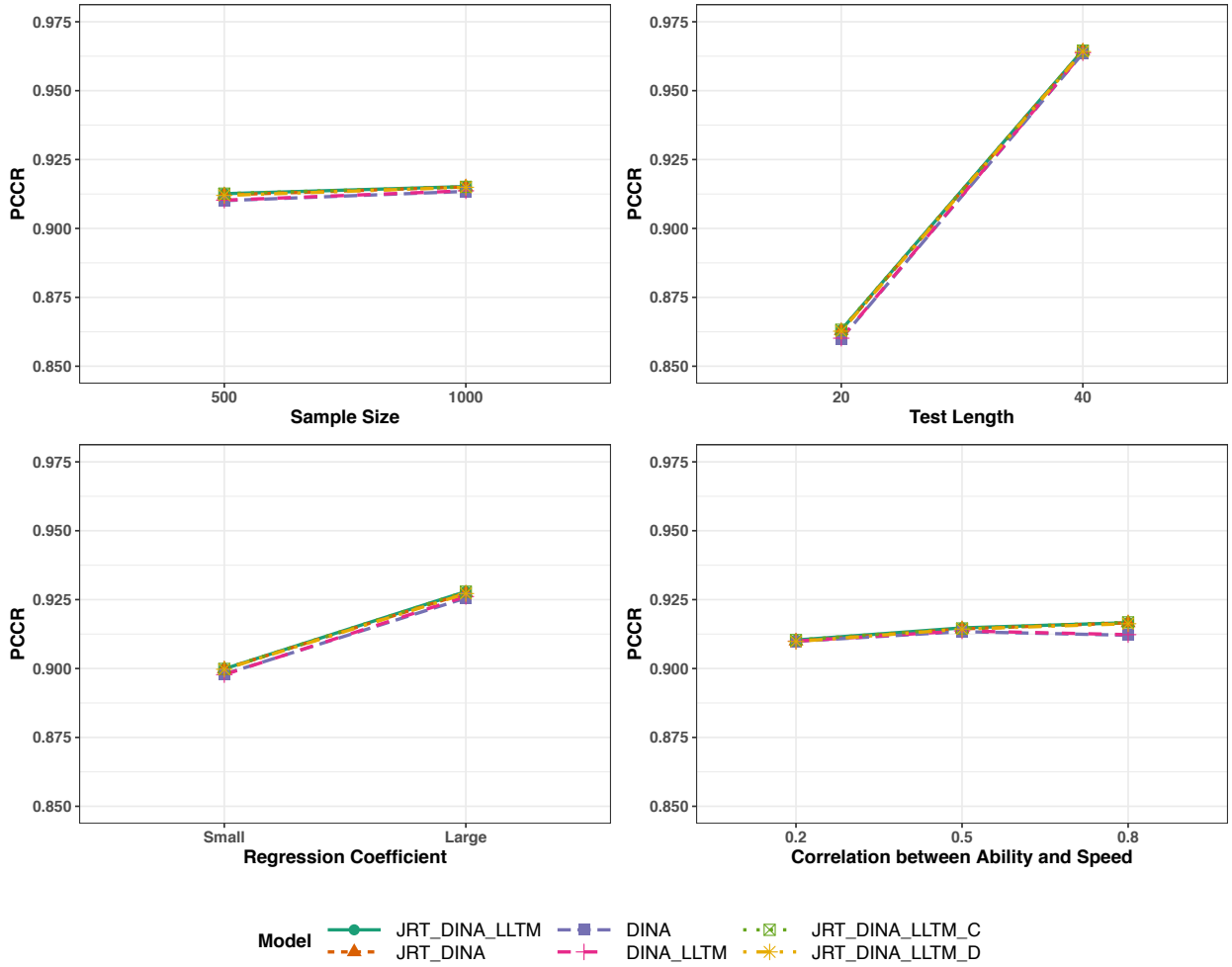


Figure 5. Marginal mean attribute pattern correct classification rates (PCCRs) at each level of the manipulated factors.

The PCCRs for all models in all 24 simulated conditions are presented in Table 24. In each simulated condition, the PCCRs for all models were similarly satisfactory. This is expected because the item covariates misspecifications in competing models mainly influence item

parameters in the models rather than person parameters. PCCRs were over 0.8 for all models in all simulated conditions, which indicates that over 80% of the attribute patterns were recovered by all models.

In summary, both ACCRs and PCCRs were similarly satisfactory for all models except for separate models (i.e., DINA and DINA-LLTM), which showed slightly smaller marginal means especially when the correlation between ability and speed became larger than 0.5. This is consistent with previous studies that joint models of RA and RT provided better attribute classification accuracy than separate models (e.g., Zhan et al., 2018). In addition, longer test and larger regression coefficients of the item covariates led to larger ACCRs and PCCRs.

Table 24. *Attribute Profile Correct Classification Rate*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	0.822	0.821	0.822	0.821	0.822	0.822
			Large	0.884	0.882	0.884	0.884	0.885	0.882
		0.5	Small	0.835	0.834	0.833	0.832	0.835	0.835
			Large	0.893	0.892	0.891	0.891	0.892	0.892
		0.8	Small	0.832	0.833	0.825	0.825	0.832	0.831
			Large	0.895	0.894	0.884	0.885	0.895	0.894
	40	0.2	Small	0.957	0.957	0.957	0.956	0.957	0.957
			Large	0.963	0.963	0.963	0.963	0.963	0.962
		0.5	Small	0.966	0.966	0.965	0.965	0.966	0.966
			Large	0.968	0.967	0.967	0.967	0.967	0.966
		0.8	Small	0.970	0.970	0.967	0.967	0.971	0.970
			Large	0.967	0.967	0.965	0.966	0.967	0.967
1000	20	0.2	Small	0.841	0.842	0.840	0.840	0.841	0.841
			Large	0.888	0.887	0.887	0.888	0.888	0.888
		0.5	Small	0.843	0.843	0.843	0.843	0.842	0.843
			Large	0.887	0.887	0.884	0.884	0.887	0.887
		0.8	Small	0.842	0.843	0.837	0.836	0.843	0.842
			Large	0.897	0.897	0.892	0.892	0.897	0.896
	40	0.2	Small	0.961	0.961	0.961	0.961	0.961	0.961
			Large	0.966	0.966	0.966	0.966	0.966	0.966
		0.5	Small	0.964	0.963	0.963	0.963	0.964	0.963
			Large	0.963	0.963	0.963	0.963	0.963	0.963
		0.8	Small	0.965	0.965	0.964	0.964	0.965	0.965
			Large	0.965	0.964	0.963	0.964	0.965	0.964

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

4.2.2 Person ability parameter

Person ability parameters were included in six models: JRT-DINA-LLTM, JRT-DINA, DINA, DINA-LLTM, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D. For scale identification, the mean and variance of the person ability parameter were constrained to be 0 and 1, respectively. Therefore, only SE and RMSE were examined as the outcome measure of the person ability parameters. Specifically, dependent variables were the SE and RMSE of person ability parameters in each simulation cell, between-subject factors were the three manipulated factors other than sample size (i.e., test length, correlation between ability and speed, and explanatory power of the item covariates), within-subject factors were the six data-fitting models. As mentioned in section 3.4, the criterion of the magnitude of the effect size is: partial $0.01 \leq \eta^2 < 0.06$ for a small effect, $0.06 \leq \eta^2 < 0.14$ for a medium effect, $\eta^2 \geq 0.14$ for a large effect (Cohen, 1988). In the result section, only effects that are both statistically significant (i.e., $p < 0.05$) and have at least a small effect size ($\eta^2 \geq 0.01$) are reported.

In addition, the visualization and elaboration focus on the highest-order interaction effects if existed. Mean and standard deviation of the outcome measures of person ability parameter estimates from each model in each simulation condition are shown in Appendix A.

Table 25 presents the significant effects of studied factors on the SE and RMSE of the person ability parameter estimates when sample size was 500 and 1000, respectively. Specifically, two three-way interaction effects were found to be significant on SE when sample size was 500: one is among data-fitting model type, explanatory power of the item covariates, and correlation between ability and speed; the other is among data-fitting model type, test length, and correlation between ability and speed, as shown in Figure 6. Specifically, the two interaction effects had a large effect size (Partial $\eta^2 = 0.113$) and a medium effect size (Partial $\eta^2 = 0.048$),

respectively. According to Figure 6, models omitting the second-level structure (i.e., DINA and DINA-LLTM) had larger SE than the four joint models. The discrepancy was larger when the test was short (= 20), correlation between ability and speed was large (= 0.8) and explanatory power of the item covariates was large. A two-way interaction effect between data-fitting model type and explanatory power of the item covariates was significant with a medium effect size (Partial η^2 = 0.075) on RMSE of person ability parameter estimates when sample size was 500, as presented in Figure 7. In general, models omitting the second-level structure (i.e., DINA and DINA-LLTM) had larger RMSE than the four joint models. In addition, larger correlation between ability and speed led to smaller RMSE of the joint models but had little impact on DINA and DINA-LLTM. This finding is consistent with previous studies that the precision of ability estimates increases when the correlation between two latent traits is larger (e.g., Klein Entink, 2009; Patton, 2015).

Table 25. *Significant Effects on the Estimation Errors of the Person Ability Estimates from the Mixed-Effect ANOVA*

J	Source	SE			RMSE		
		F Statistics	p-value	Partial η^2	F Statistics	p-value	Partial η^2
500	Model	1098.112	< 0.001	0.155	633.574	< 0.001	0.096
	Model*G	324.521	< 0.001	0.051			
	Model*R	611.866	< 0.001	0.170	242.268	< 0.001	0.075
	Model*I	449.858	< 0.001	0.070			
	Model*G*R	381.441	< 0.001	0.113			
	Model*R*I	150.752	< 0.001	0.048			
	R	510.291	< 0.001	0.146	34.207	< 0.001	0.011
	I	4394.937	< 0.001	0.423			
	G*R	187.692	< 0.001	0.059			
	R*I	85.901	< 0.001	0.028			
1000	Model	433.106	< 0.001	0.035	1058.856	< 0.001	0.081
	Model*R	273.173	< 0.001	0.044	404.271	< 0.001	0.063
	Model*I	1630.058	< 0.001	0.120			
	Model*G*I	251.863	< 0.001	0.021			
	Model*R*I	668.750	< 0.001	0.100			
	Model*G*R*I	433.992	< 0.001	0.068			
	R	913.700	< 0.001	0.132	98.448	< 0.001	0.016
	I	6223.986	< 0.001	0.342			
	G*R*I	204.591	< 0.001	0.033			

Note. G = explanatory power of item covariates; I = test length; R = correlation between ability and speed.

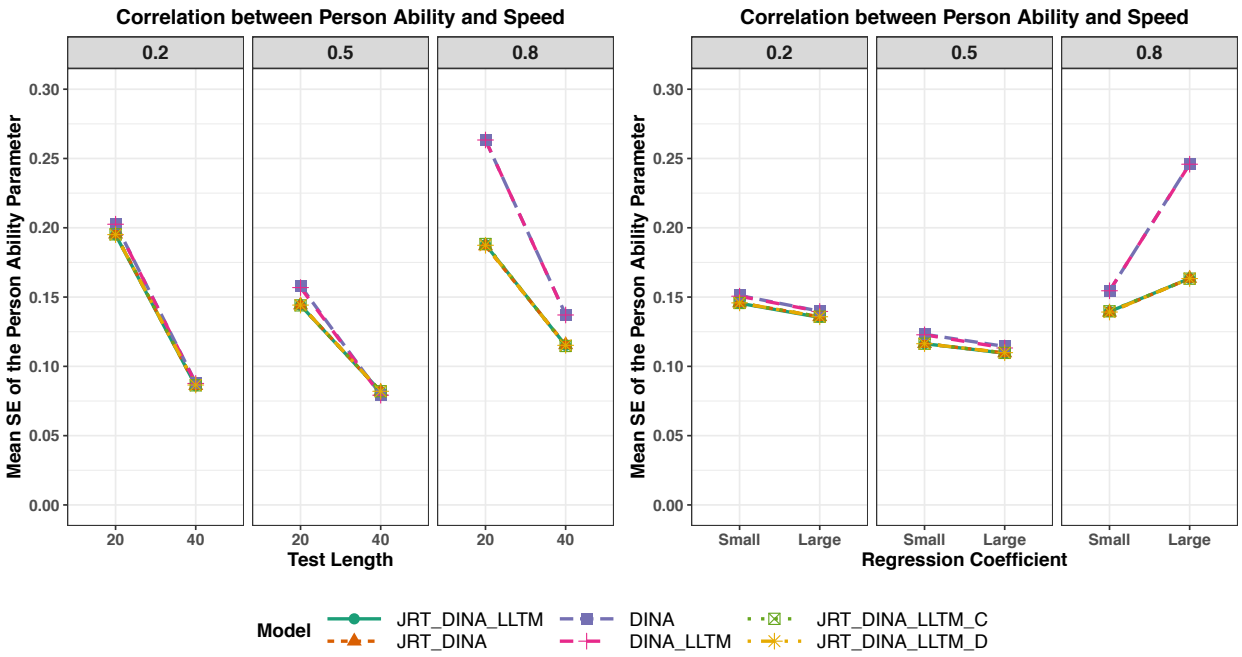


Figure 6. Three-way interaction effects on SE of Person Ability Parameter Estimates (J = 500).

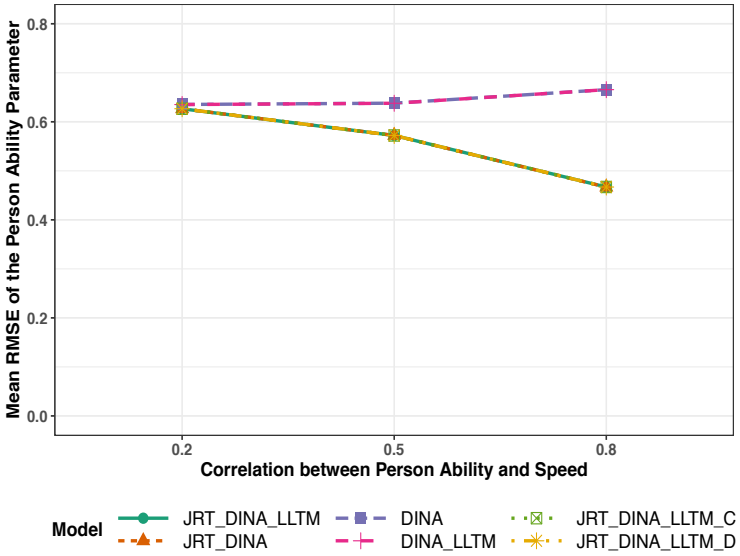


Figure 7. Two-way interaction effects on RMSE of Person Ability Parameter Estimates (J = 500).

Similar patterns were found when sample size was 1000. A four-way interaction among model type, explanatory power of the item covariates, test length and correlation between ability

and speed was significant with a medium effect size (Partial η^2 = 0.068). The four-way interaction is separated into two three-way interactions as shown in Figure 8. Models omitting the second-level (i.e., DINA and DINA-LLTM) had larger SE than the four joint models when test length was 20 but smaller SE then test length was 40. The discrepancy of the SE for models omitting the second-level and joint models increased when correlation between ability and speed increased especially from 0.5 to 0.8 and when explanatory power of the item covariates was large. The two-way interaction between model type and correlation between ability and speed was found significant for the RMSE of person ability parameter estimates when sample size was 1000. Similar to that when sample size was 500, models omitting the second-level (i.e., DINA and DINA-LLTM) had larger RMSE than the four joint models. In addition, larger correlation between ability and speed led to smaller RMSE of the joint models.

A three-way ANOVA was conducted to examine the impact of the manipulated factors (test length, correlation between ability and speed, explanatory power of the item covariates) on the parameter recovery of person ability parameter estimates from the proposed model JRT-DINA-LLTM. The results are shown in Table 26. In general, short test length led to larger SE and RMSE of the person ability parameter estimates. Large correlation between ability and speed led to larger SE and smaller RMSE. The SE of person ability parameter estimates was smaller when the item covariates had larger explanatory power, which indicated larger correlations among item parameters.

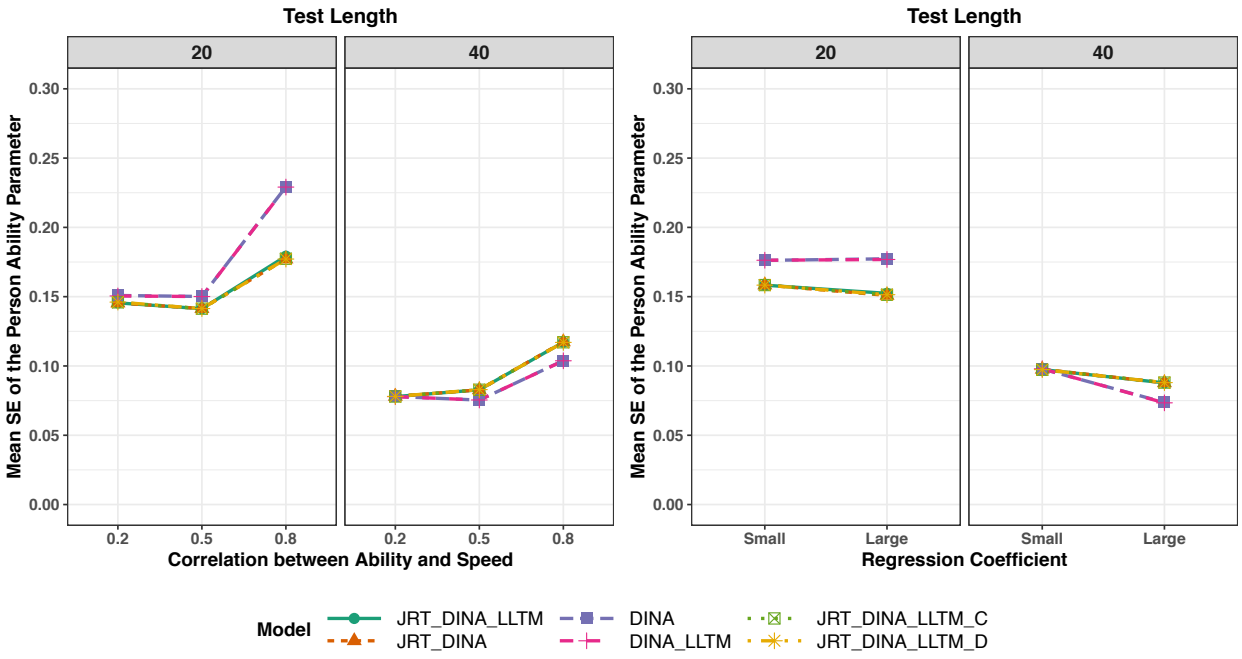


Figure 8. Three-way interaction effects on SE of Person Ability Parameter Estimates (J = 1000).

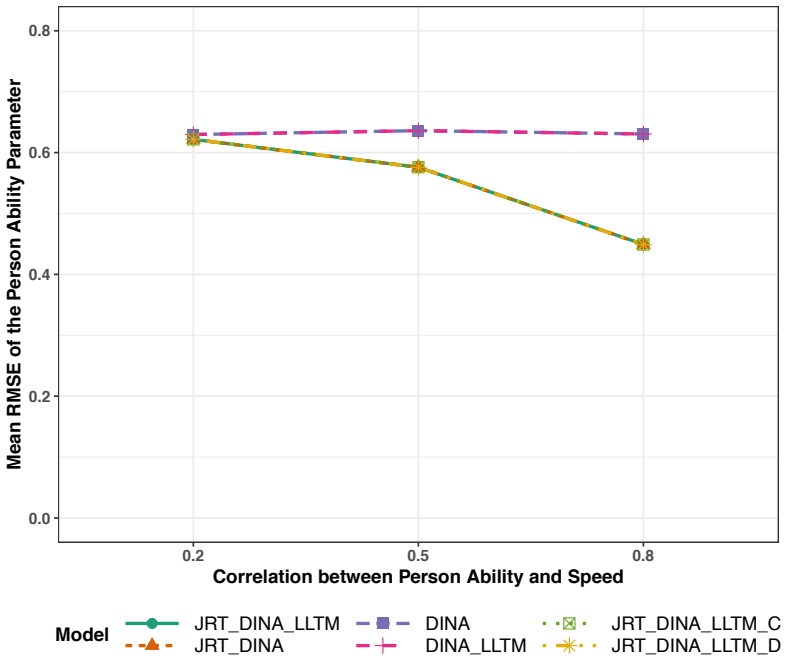


Figure 9. Two-way interaction effects on RMSE of Person Ability Parameter Estimates (J = 1000).

Table 26. *Significant Effects in the Three-way ANOVA of the Person Ability Estimates from the JRT-DINA-LLTM*

J	Source	SE			RMSE		
		F Statistics	p-value	Partial η^2	F Statistics	p-value	Partial η^2
500	R				93.818	< 0.001	0.030
	G*R	83.633	< 0.001	0.027			
	I*R	141.838	< 0.001	0.045			
1000	R				217.323	< 0.001	0.035
	G*R*I	114.506	< 0.001	0.019			

Note. G = explanatory power of item covariates; I = test length; R = correlation between ability and speed.

4.2.3 Person speed parameter and speed variance

Person speed parameters indicate respondents' work speed during the assessments. The mean of the person speed parameter was fixed to be 0 for identifiability purposes. Thus, the recovery of speed parameter focuses on SE and RMSE. The ANOVA design was the same as that for person ability parameters. In addition, the impact of model type and manipulated factors on the recovery of the variance of the person speed parameter was also examined by plotting the marginal means of its bias, SE, RMSE in each level of the manipulated factors.

4.2.3.1 Person speed parameter

Table 27 presents the significant effects on SE and RMSE when sample size was 500 and 1000, respectively. The visualizations for the interaction effects with the highest level are shown in Figures 10-15. The ANOVA results and visualizations when sample size was 1000 were very similar to those when sample size equaled 500. Therefore, only results when sample size equaled 500 were elaborated. A three-way interaction effect was significant with a small effect size (Partial η^2 = 0.050) for SE when sample size was 500. Specifically, short test length led to larger SE of person speed parameters. Models omitting the second-level structure (DINA, DINA-LLTM) had larger SE than the four joint models especially when test length was short and correlation between ability and speed was large (= 0.8). In addition, a two-way interaction effect

was significant with a small effect size (Partial η^2 = 0.018) and a significant main effect with a large effect size (Partial η^2 = 0.606) for RMSE when sample size was 500. Specifically, short test length led to larger RMSE of person speed parameter estimates. In addition, larger correlation between speed and ability led to smaller RMSE of person speed parameter estimates from the joint models, thus, resulting in larger discrepancy between models without second-level structure and joint models.

Table 27. *Significant Effects in the Mixed-Effect ANOVA Results of the SE and RMSE of the Person Speed Estimates*

J	Effect	SE			RMSE		
		F Statistics	p-value	Partial η^2	F Statistics	p-value	Partial η^2
500	Model	2229.859	< 0.001	0.271	119.853	< 0.001	0.020
	Model*R	846.068	< 0.001	0.220	53.489	< 0.001	0.018
	Model*R*I	158.857	< 0.001	0.050			
	I	8723.593	< 0.001	0.593	9227.535	< 0.001	0.606
1000	Model	3404.833	< 0.001	0.221	276.499	< 0.001	0.023
	Model*R	1189.325	< 0.001	0.166	89.235	< 0.001	0.015
	Model*I	733.856	< 0.001	0.058			
	Model*R*I	239.186	< 0.001	0.038			
	I	17237.871	< 0.001	0.590	18095.832	< 0.001	0.602

Note. I = test length; R = Correlation between ability and speed; J = Sample size.

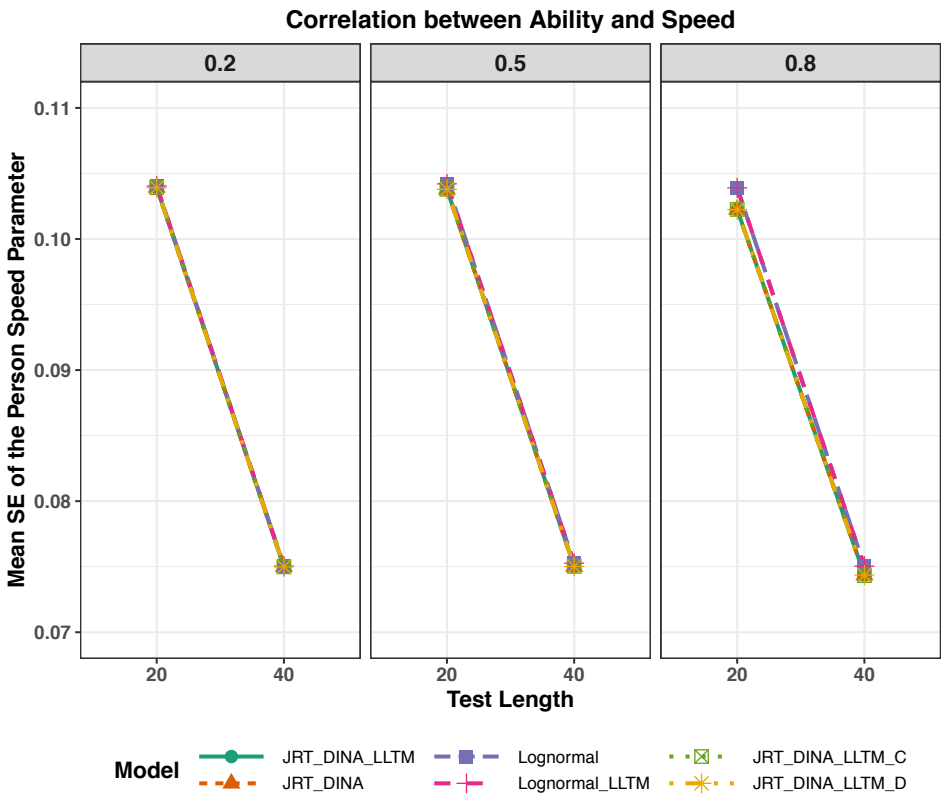


Figure 10. Three-way interaction effects on SE of person speed parameter estimates (J = 500).

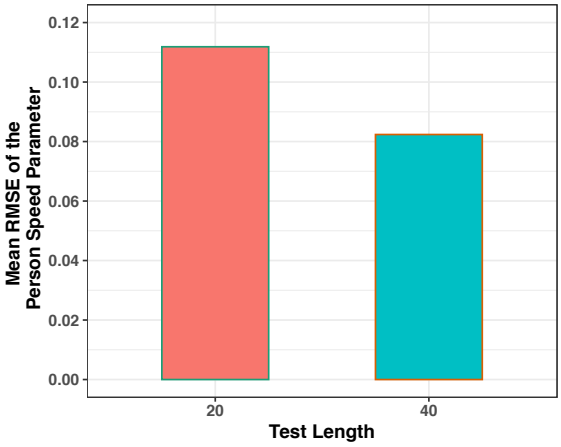


Figure 11. The main effect of test length on RMSE of person speed parameter (J = 500).

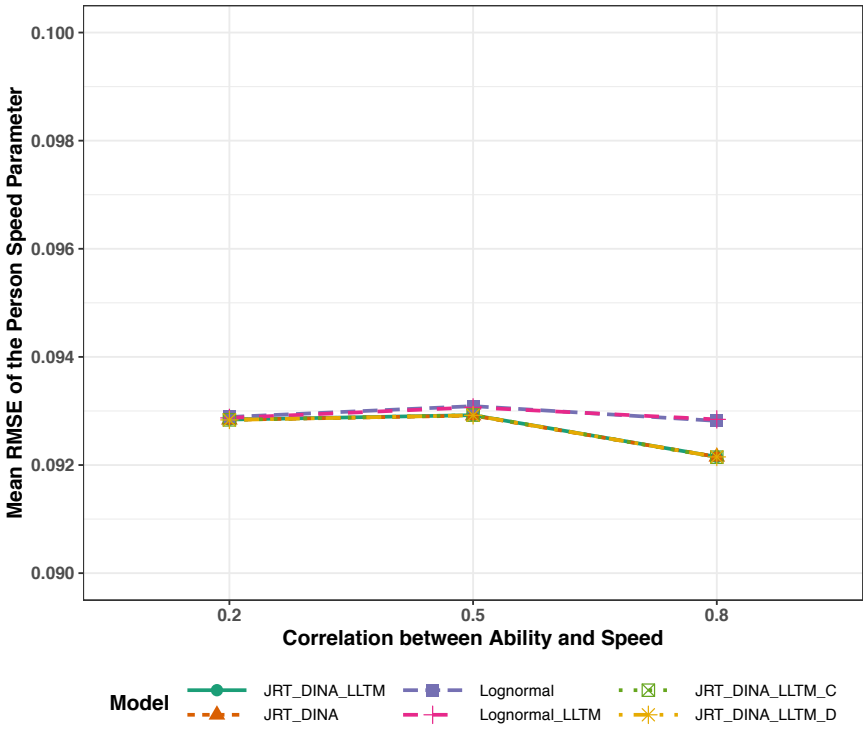


Figure 12. Two-way interaction effects on RMSE of person speed parameter estimates ($J = 500$).

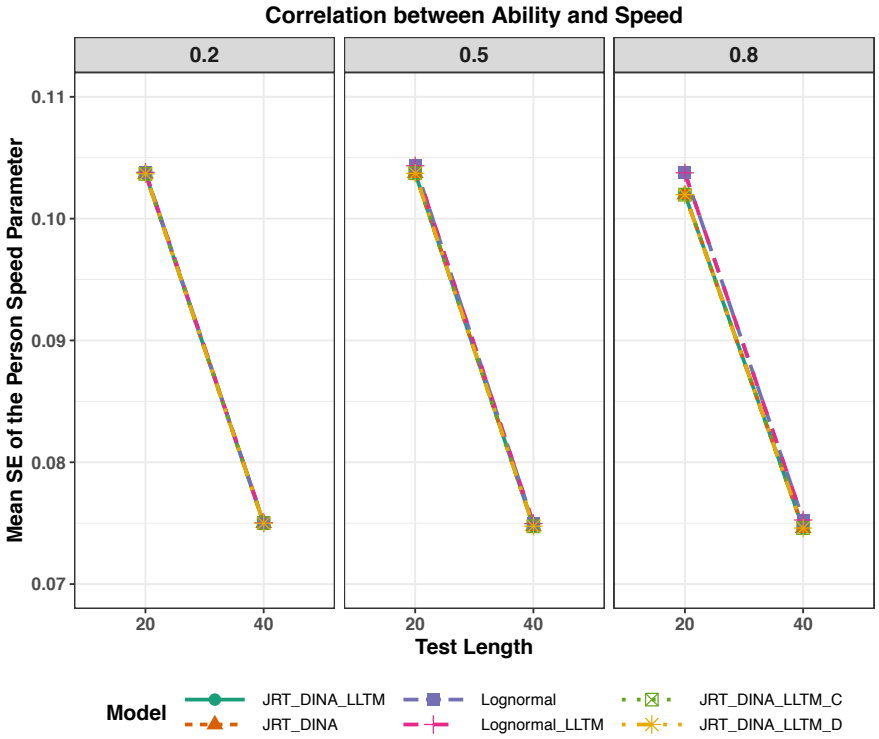


Figure 13. Three-way interaction effects on SE of person speed parameter estimates ($J = 1000$).

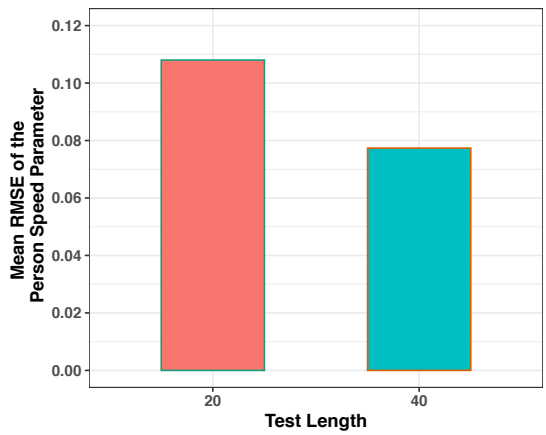


Figure 14. The main effect of test length on RMSE of person speed parameter (J = 1000).

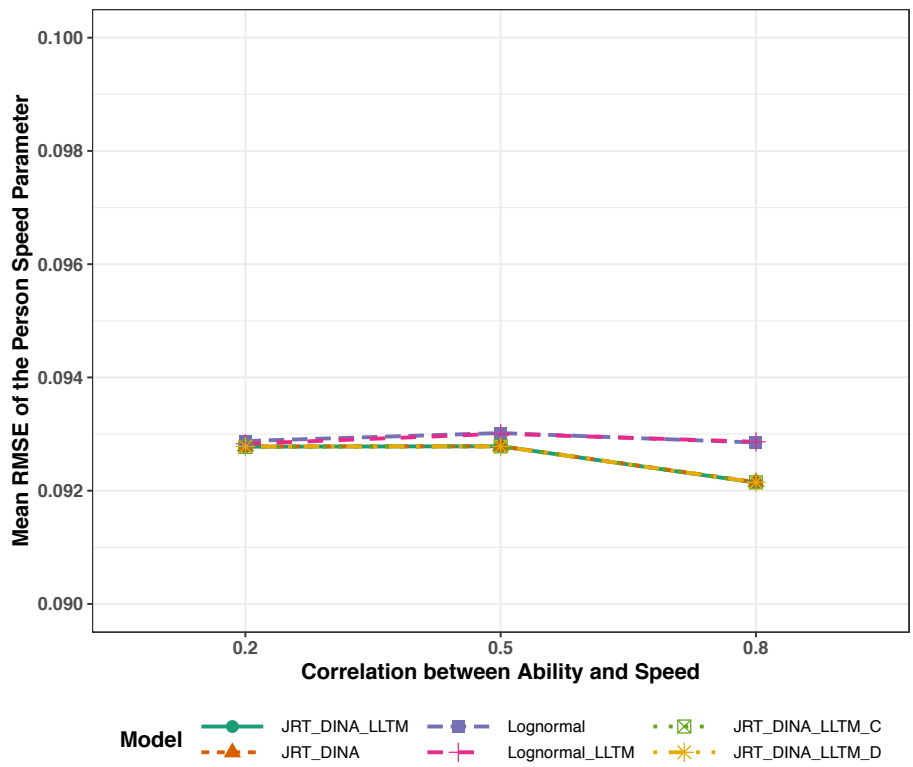


Figure 15. Two-way interaction effects on RMSE of person speed parameter estimates (J = 1000).

The three-way ANOVA was conducted to examine the impact of manipulated factors, including test length, correlation between ability and speed, explanatory power of the item covariates, on the parameter recovery of person speed parameter estimates from the proposed

model JRT-DINA-LLTM. The results are shown in Table 28. Test length had a significant main effect with large effect sizes. In general, short test length led to larger SE and RMSE of the person speed parameter estimates obtained from JRT-DINA-LLTM.

Table 28. *Significant Effects in the Three-way ANOVA of the Person Speed Estimates from the JRT-DINA-LLTM*

J	Source	SE			RMSE		
		F Statistics	p-value	Partial η^2	F Statistics	p-value	Partial η^2
500	I	8666.249	< 0.001	0.591	9086.330	< 0.001	0.603
1000	I	17079.591	< 0.001	0.588	17816.401	< 0.001	0.598

Note. I = test length; J = Sample size.

4.2.3.2 Person speed variance

The impact of model type and manipulated factors on the recovery of person speed variance was examined by plotting the mean Bias, SE, and RMSE of the person speed variance estimates from different data-fitting models at each level of the manipulated factors as shown in Figures 16-18. All models overestimated the speed variance parameter at all levels of manipulated factors given the positive mean Bias. Models without the second level structure (i.e., Lognormal and Lognormal-LLTM) had larger mean systematic error indicated by mean bias at all factor levels. This is expected because multidimensional models usually had higher measurement precision. The average random error quantified by mean SE was negligible for models at all factor levels. As expected, the average total error indicated by mean RMSE had a similar trend as mean bias that Lognormal and Lognormal-LLTM had larger total error than joint models. However, the discrepancy became negligible when sample size and correlation between ability and speed were large.

For the proposed model, sample size and correlation between ability and speed had noticeable impact on the recovery of the speed variance same as all other data-fitting models. Specifically, larger sample size led to smaller mean bias and RMSE of the speed variance

parameter estimates, while larger correlation between ability and speed led to larger mean bias and RMSE of the speed variance parameter estimates.

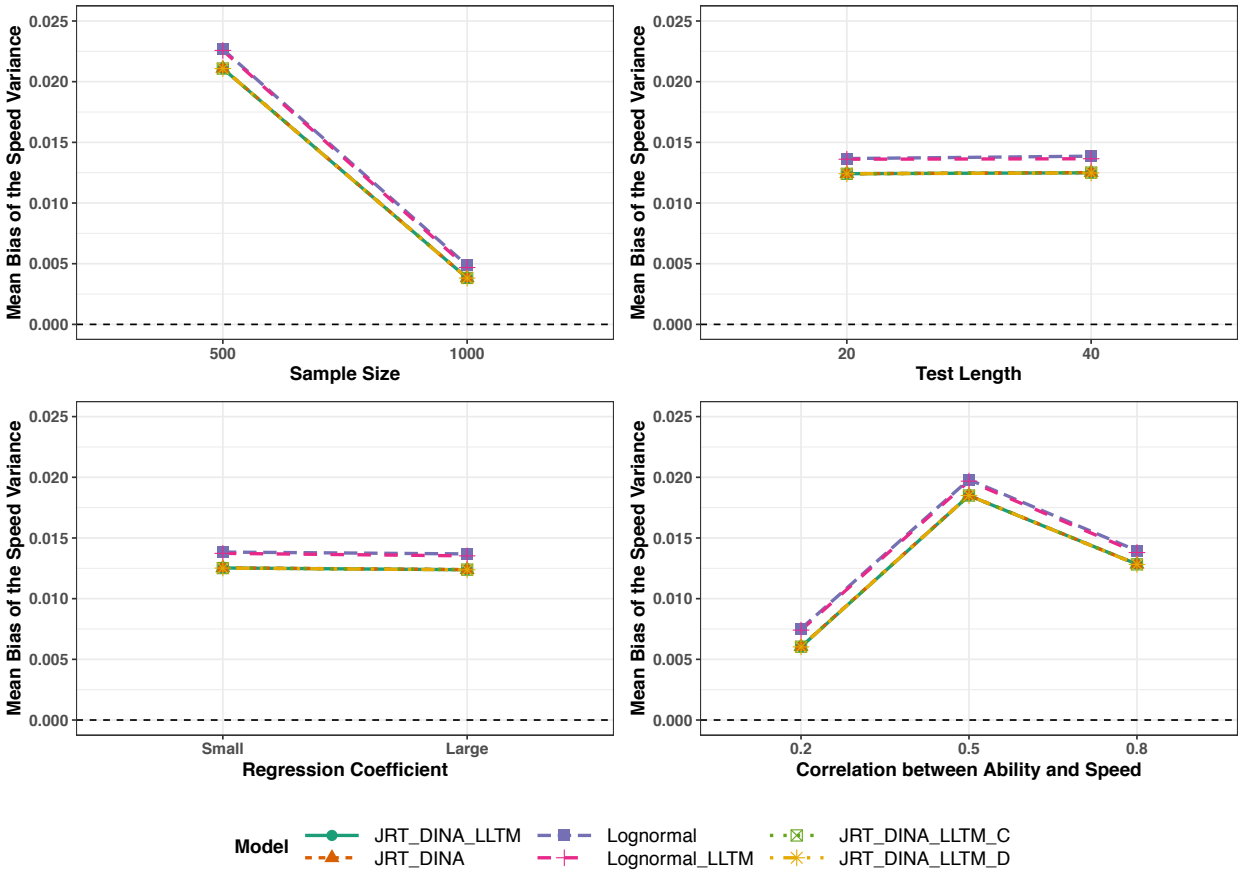


Figure 16. Marginal mean bias of the speed variance estimates at each level of the manipulated factors.

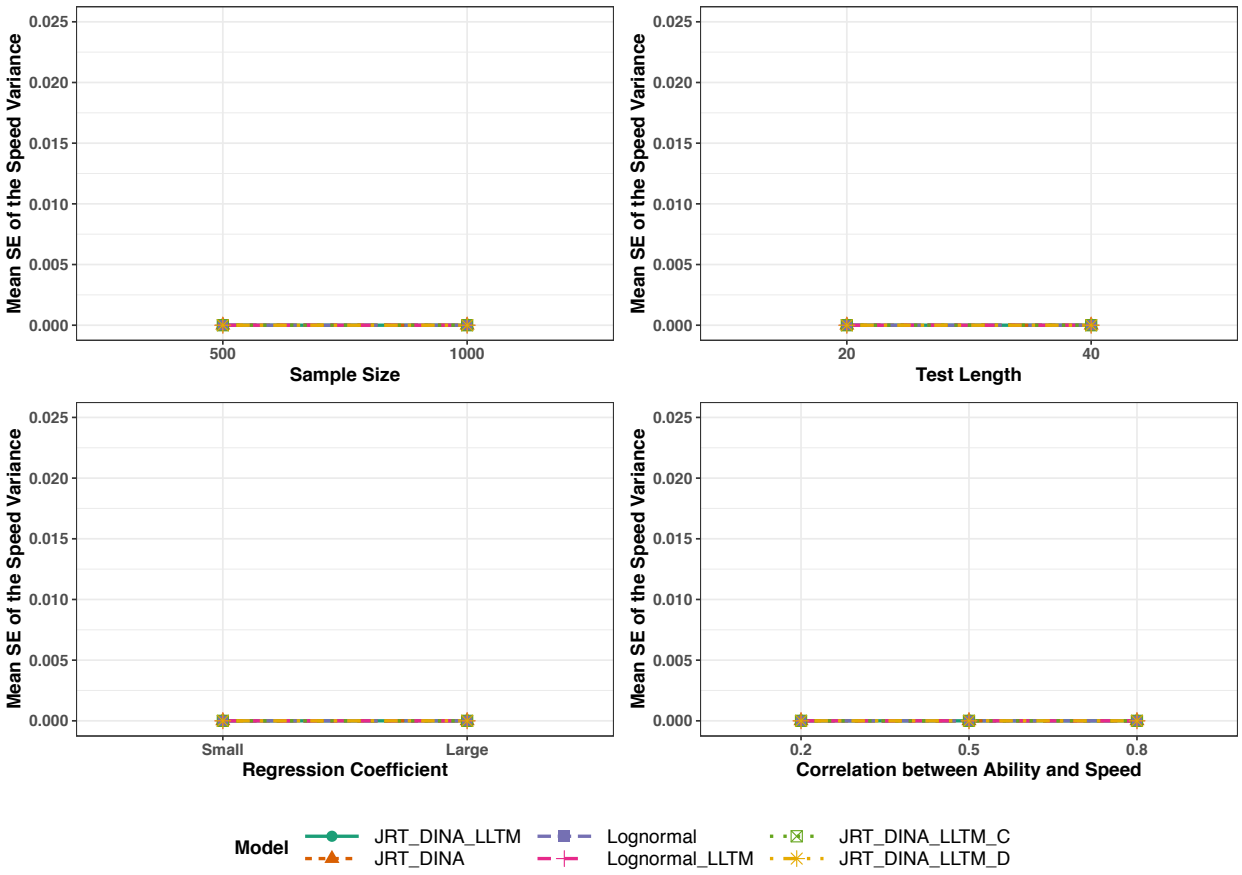


Figure 17. Marginal mean SE of the speed variance estimates at each level of the manipulated factors.

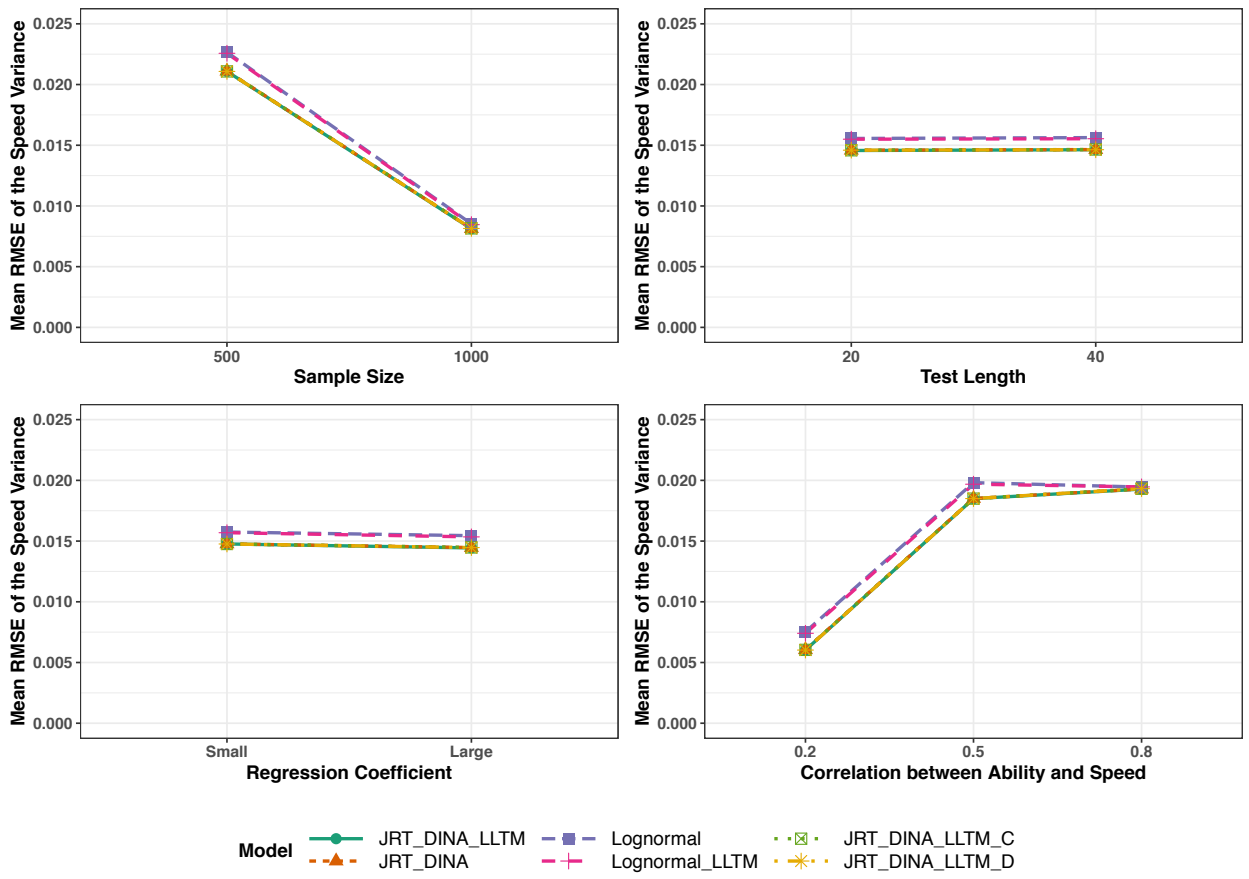


Figure 18. Marginal mean RMSE of the speed variance estimates at each level of the manipulated factors.

4.2.3 Correlation between Ability and Speed $\rho_{\theta\tau}$

On average, all models overestimated $\rho_{\theta\tau}$ given the positive mean bias at all levels of manipulated factors as shown in Figure 19. In addition, the marginal mean SE at all levels of manipulated factors for all models were zero, which indicates that the average random error was trivial. Further, the four joint models (i.e., JRT-DINA-LLTM, JRT-DINA, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D) had similar mean bias and RMSE at all levels of manipulated factors. This is expected because the four models are different in terms of the covariate specifications which may not affect the estimation of person parameters. Given that average SE was negligible, the mean RMSE represents the average absolute values of systematic errors. As shown in Figure 21,

larger sample size, longer test length, smaller explanatory power of item covariates and smaller $\rho_{\theta\tau}$ led to larger total error in the recovery of $\rho_{\theta\tau}$ for all data-fitting models.

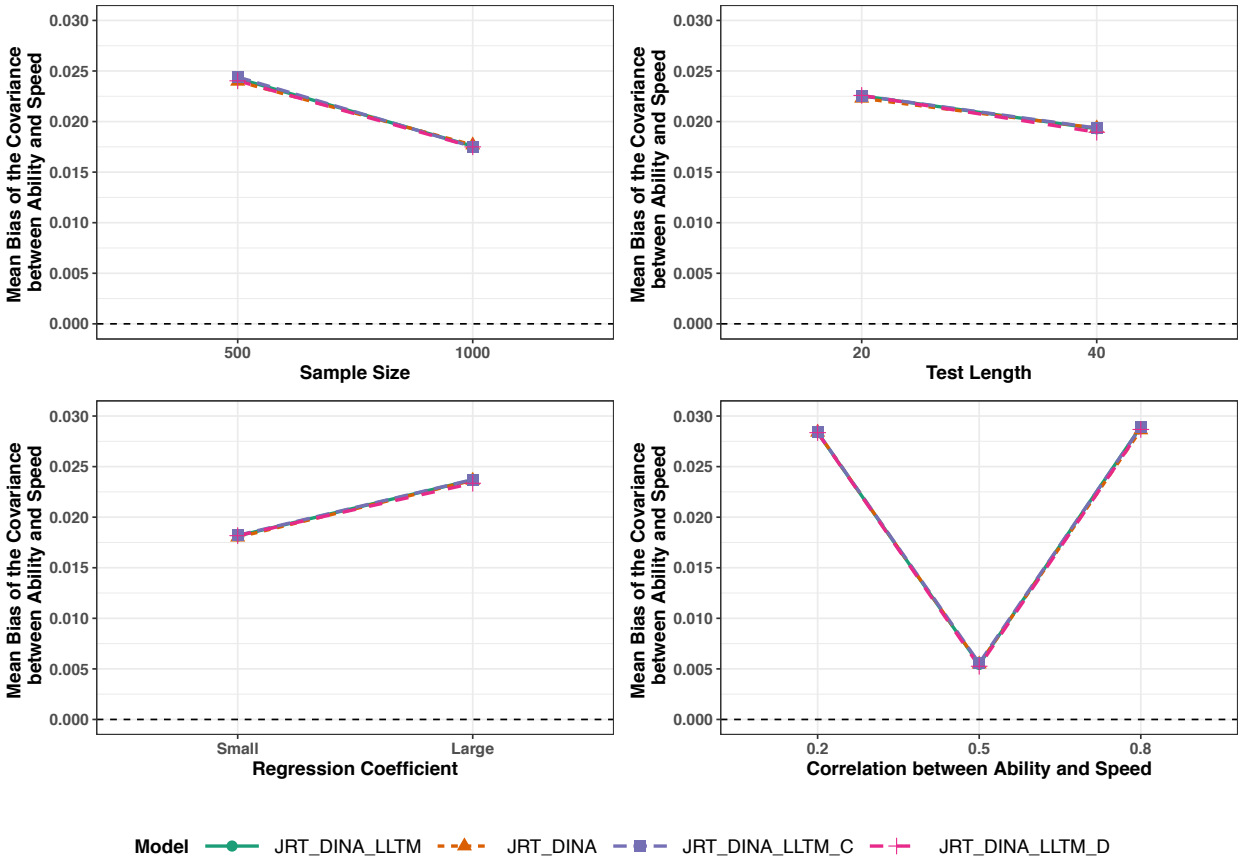


Figure 19. Marginal mean bias of the correlation between ability and speed estimates at each level of the manipulated factors.

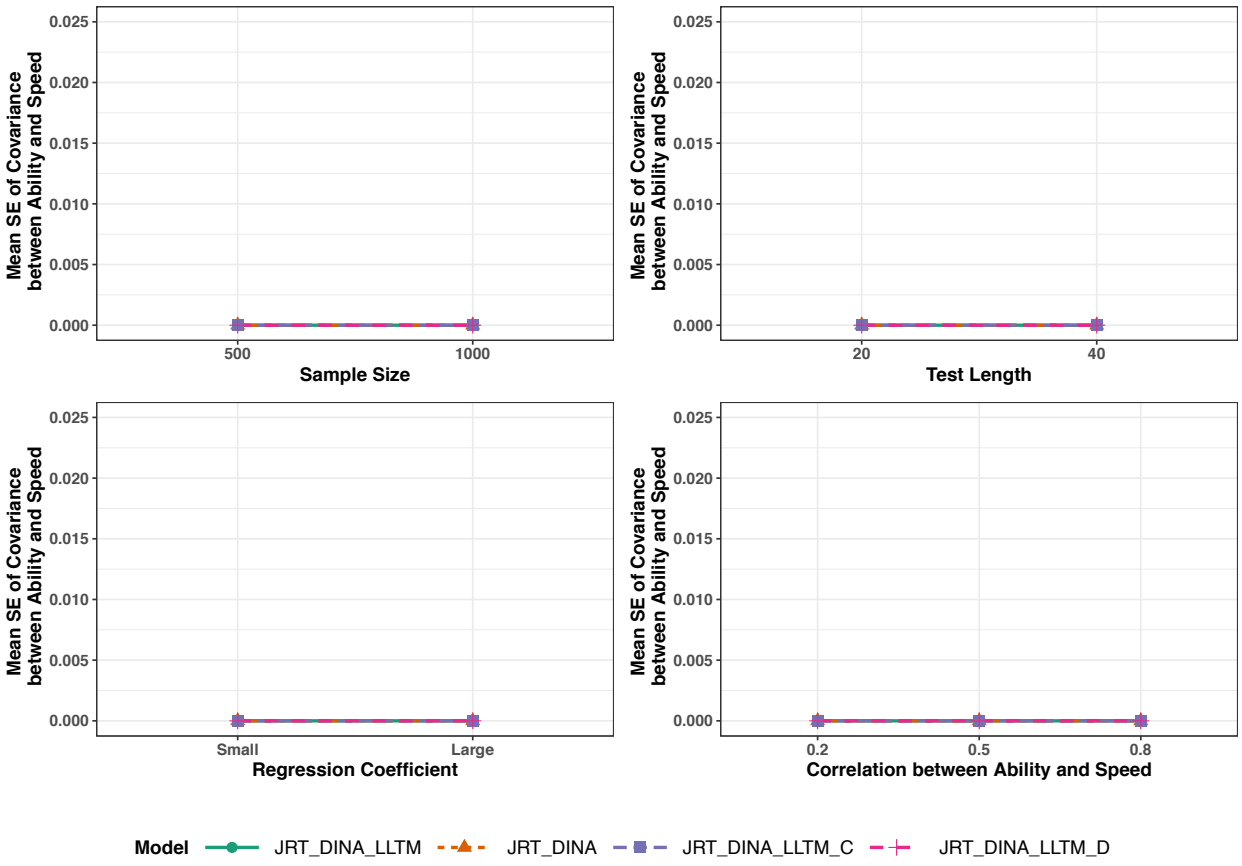


Figure 20. Marginal mean SE of the correlation between ability and speed at each level of the manipulated factors.

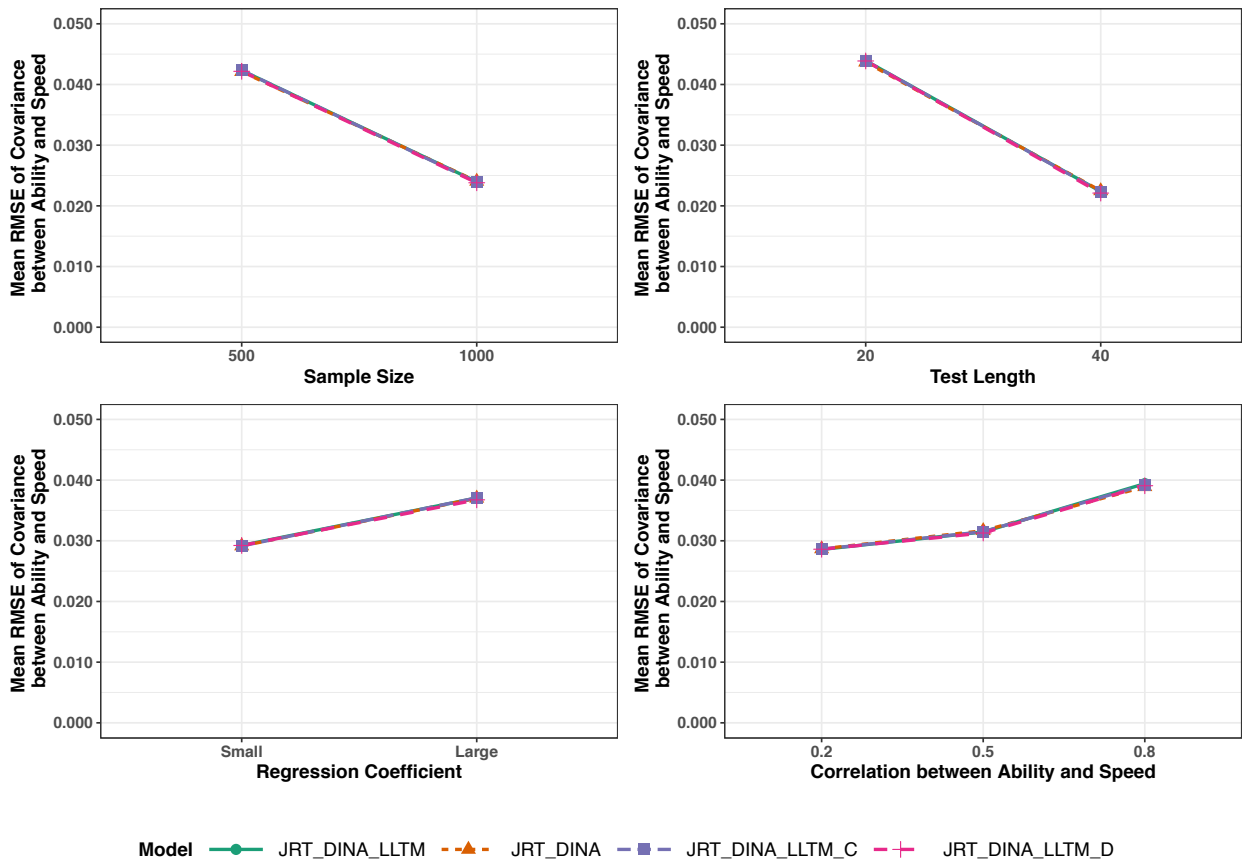


Figure 21. Marginal mean RMSE of the correlation between ability and speed at each level of the manipulated factors.

4.3 Recovery of the Item Parameters

The item parameters investigated in the simulation study include item intercept parameter (β), item interaction parameter (δ), item intensity parameter (ζ), and regression intercepts ($\mathbf{A}_0$), regression coefficients for the dichotomous covariate ($\mathbf{A}_1$), regression coefficients for the continuous covariate ($\mathbf{A}_2$). Specifically, item intercept parameter and item interaction parameter existed in the response models, while item intensity parameter existed in the RT model. Regression coefficients existed in all models. JRT-DINA-LLTM, DINA-LLTM, Lognormal-LLTM, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D estimated the regression coefficients in a one-step manner, while JRT-DINA, DINA, Lognormal model estimated the regression

coefficients in a two-step manner. Among these parameters, item intercept, item interaction and item intensity are item-specific and had enough sample sizes in each condition for the implementation of the mixed-effect ANOVA to examine the impact of manipulated factors and different data-fitting models. The mixed-effect ANOVA was conducted separately for conditions where test length was 20 and test length was 40 because mixed-effect ANOVA results from groups with equal sample sizes are robust to the violation of the homogeneous residual variance assumption. In addition, a three-way ANOVA was conducted to examine the impact of manipulated factors other than test length on the parameter recovery of item intercept, item interaction and item intensity in the proposed model JRT-DINA-LLTM. Detailed information on the mean bias, SE, RMSE of the item parameter estimates is presented in Appendix B.

4.3.1 Item Intercept Parameter

The significant effects obtained from the mixed-effect ANOVA results on the bias of item intercept parameters are presented in Table 29. When test length was 20, a significant main effect of model type found with a small effect size (Partial η^2 = 0.026), as shown in Figure 22. When test length was 40, a significant interaction effect between the explanatory power of the item covariates and model type was found with a small effect size (Partial η^2 = 0.016), as shown in Figure 23. According to Figure 22, the mean bias for each data fitting model was negative when test length was 20, indicating that the models tended to underestimate the intercept parameters on average. When test length was long (I = 40), mean biases of item intercept parameter estimates obtained from JRT-DINA-LLTM and DINA-LLTM canceled out the most and were the closest to 0. The other models overestimated the item intercept parameter estimates based on the positive mean bias, but the magnitude was negligible. As shown in Figure 23, when the true regression coefficients of item covariates were large, JRT-DINA-LLTM, DINA-LLTM,

and DINA underestimated the item intercept parameter on average given the negative mean bias, while the other models still overestimated the item intercept parameter. One noticeable observation is that JRT-DINA-LLTM-D (i.e., model including only the dichotomous covariate) had an inflated mean bias when the true regression coefficients became larger. A possible reason is that the estimation of the regression coefficient of dichotomous covariate or related regression intercept is less precise than that of the regression coefficient of continuous covariate. This can be seen that the mean bias of JRT-DINA-LLTM-C (i.e., model including only the continuous covariate) did not change as the true regression coefficients became larger.

Table 29. *Significant Effects in the Mixed-Effect ANOVA Results on Bias of the Item Intercept Parameter Estimates*

I	Effect	Bias		
		F Statistics	p-value	Partial η^2
20	Model	6.088	0.014	0.026
40	Model	10.761	< 0.001	0.022
	Model*G	5.912	0.001	0.012

Note. G = Explanatory power of regression coefficients.

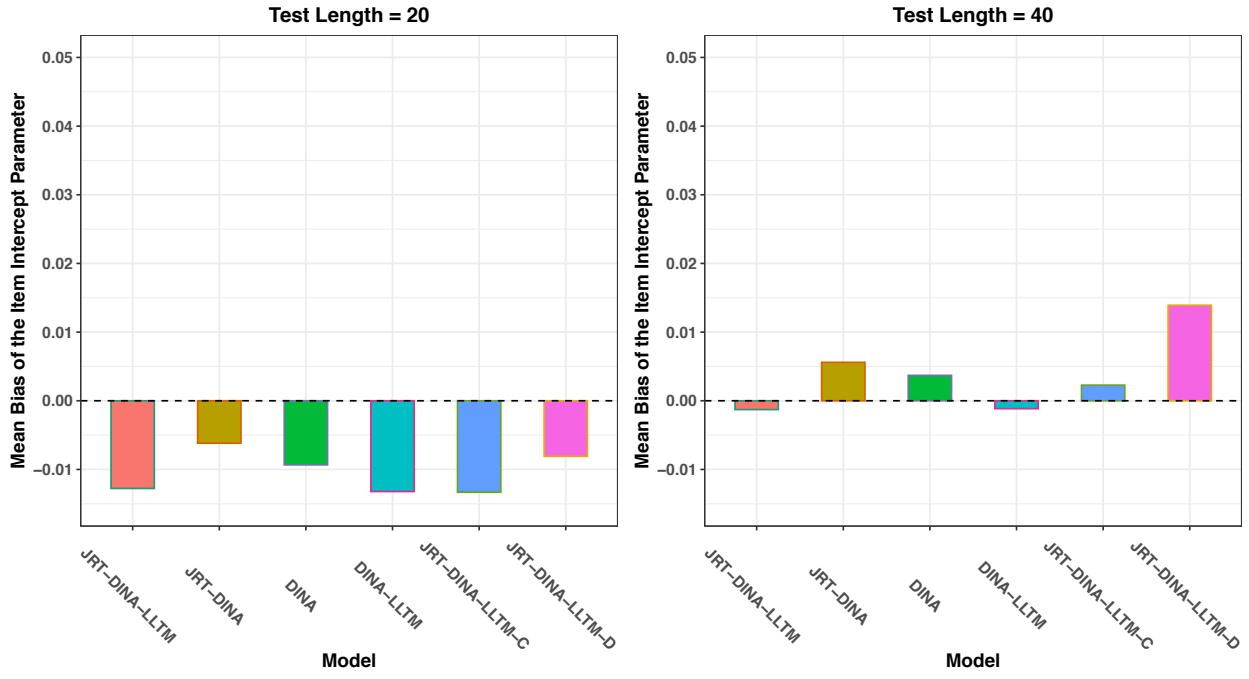


Figure 22. Marginal mean bias of the item intercept parameter estimates for each data fitting model.

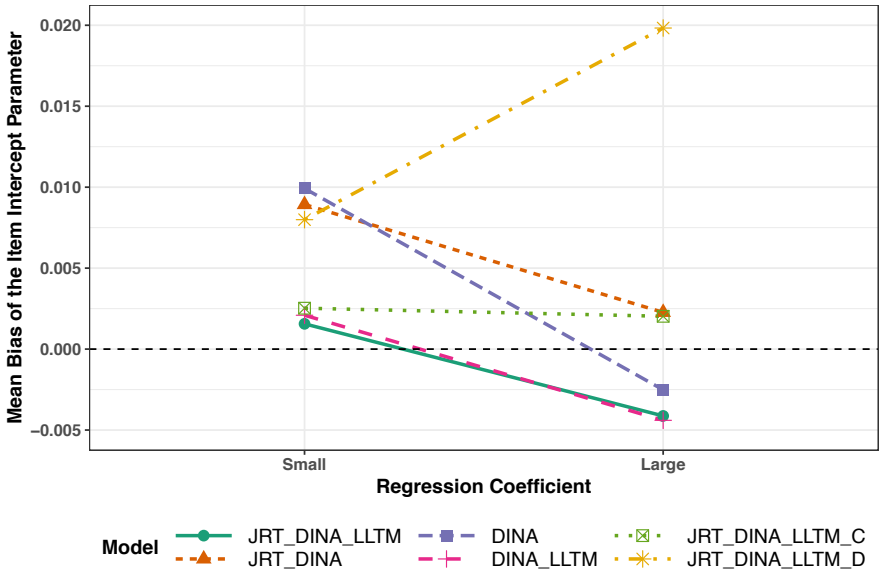


Figure 23. Marginal mean Bias of the item intercept parameter at each level of the manipulated factors (I = 40).

The significant effects of the mixed-effect ANOVA on SE and RMSE of the item intercept parameter estimates are presented in Table 30. When test length was 20, a significant three-way interaction effect among model type, sample size, and explanatory power of regression coefficients existed with a small effect size partial η^2 = 0.056 and partial η^2 = 0.024 on SE and RMSE, respectively. In general, models with one-step estimation of the regression coefficients (i.e., JRT-DINA-LLTM, DINA-LLTM, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D) had smaller mean SE and RMSE than models with two-step estimation (i.e., DINA and JRT-DINA) when the explanatory power of item covariates was large, but larger mean SE and RMSE when the explanatory power of item covariates was small. In addition, the discrepancy became smaller when the sample size increased. Among models with one-step estimation of the regression coefficients, JRT-DINA-LLTM, DINA-LLTM, JRT-DINA-LLTM-C had similar mean SE and RMSE at each level of the manipulated factors, while JRT-DINA-LLTM-D had noticeable larger

mean SE and RMSE when the explanatory power of item covariates was large. This is consistent with the finding on mean bias. For the two models with two-step estimation, JRT-DINA had smaller mean SE and RMSE than DINA especially when the explanatory power of the item covariates was large. A possible reason is that large explanatory power of the item covariates indicates large correlations among the item parameters, which may increase the measurement precision of the item parameters.

When test length was 40, two significant interaction effects: model and sample size, model and explanatory power of item covariates existed on SE and RMSE. In general, SE and RMSE of item intercept parameter estimates decreased when sample size increased and explanatory power of item covariates became large. The discrepancy among models in terms of SE or RMSE was large when explanatory power of item covariates was large but negligible when explanatory power of item covariates was small. Specifically, JRT-DINA-LLTM and DINA-LLTM had the smallest SE and RMSE at all levels of manipulated factors.

Table 30. *Significant Effects in the Mixed-Effect ANOVA Results on SE and RMSE of the Item Intercept Parameter Estimates*

		SE			RMSE		
I	Effect	F Statistics	*p*-value	Partial η^2	F Statistics	*p*-value	Partial η^2
20	Model	14.200	< 0.001	0.059	10.772	< 0.001	0.045
	Model*G	106.017	< 0.001	0.317	47.726	< 0.001	0.173
	Model*G*J	13.546	< 0.001	0.056	5.711	0.004	0.024
	J	41.145	< 0.001	0.153	44.771	< 0.001	0.164
40	Model	143.444	< 0.001	0.235	67.974	< 0.001	0.127
	Model*G	172.158	< 0.001	0.269	33.854	< 0.001	0.067
	Model*J	13.092	0.001	0.027	6.600	0.001	0.014
	G	20.815	< 0.001	0.043	15.249	< 0.001	0.032
	J	62.593	< 0.001	0.118	35.310	< 0.001	0.070

Note. *G* = Explanatory power of item covariates; *J* = Sample size.

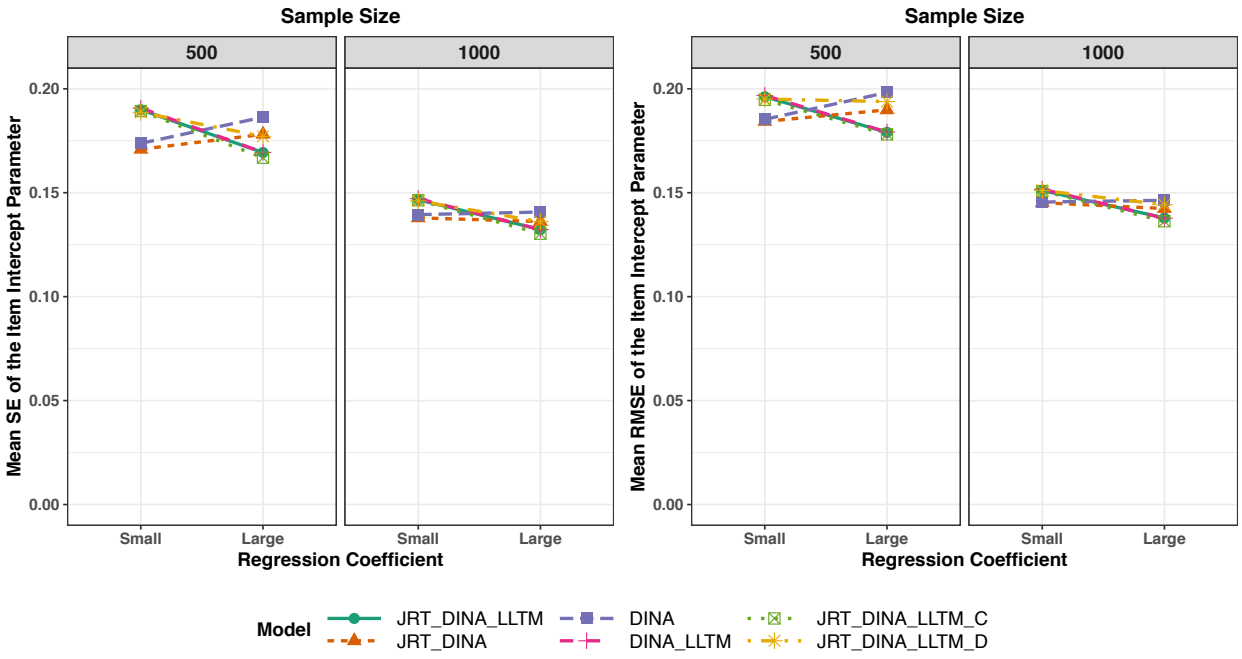


Figure 24. Marginal mean SE and RMSE of the item intercept parameter at each level of the manipulated factors ($I = 20$).

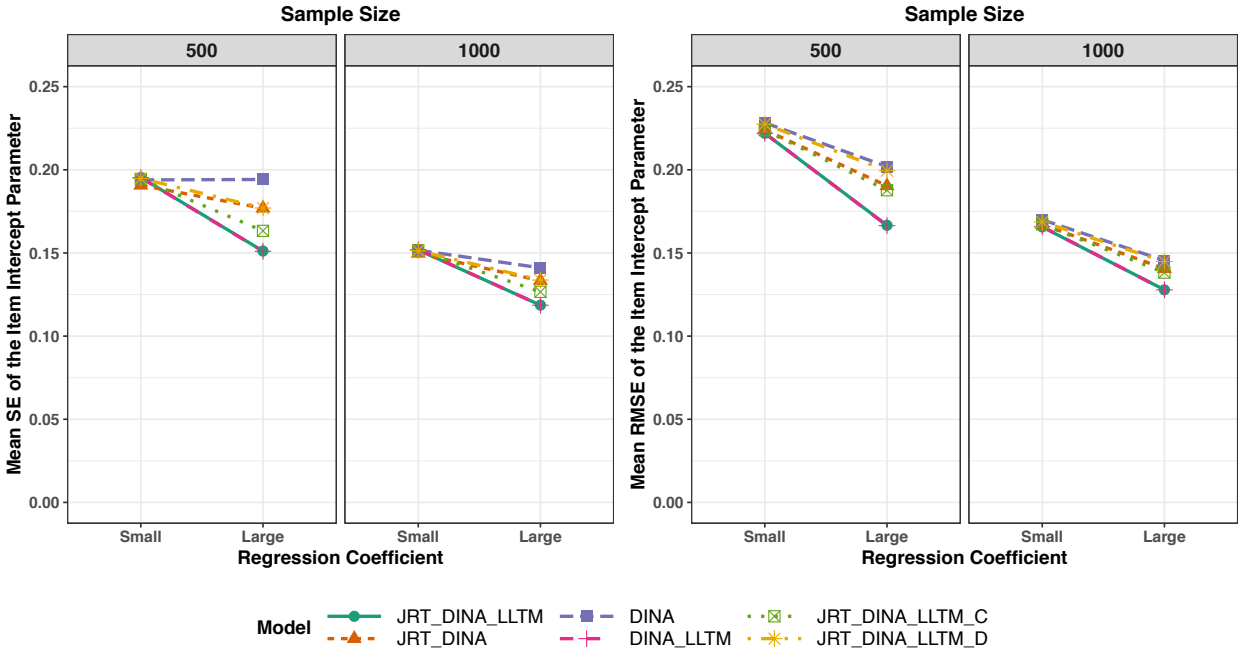


Figure 25. Marginal mean SE and RMSE of the item intercept parameter at each level of the manipulated factors ($I = 40$).

The three-way ANOVA was conducted to examine the impact of the manipulated factors on the parameter recovery of the item intercept parameter estimates from the JRT-DINA-LLTM. According to Table 31, the explanatory power of the item covariates and sample size had main effects on SE and RMSE under both test length scenarios. When test length was 20, the explanatory power of the item covariates had a small effect size on SE (Partial η^2 = 0.031) and RMSE (Partial η^2 = 0.022), respectively. Sample size had a large effect size on SE (Partial η^2 = 0.147) and RMSE (Partial η^2 = 0.155), respectively. When test length was 40, the explanatory power of the item covariates had a large effect size on SE (Partial η^2 = 0.121) and a medium effect size on RMSE (Partial η^2 = 0.071), respectively. Sample size had a large effect size on SE (Partial η^2 = 0.117) and a medium effect size on RMSE (Partial η^2 = 0.073), respectively. An inspection on the marginal means of the SE and RMSE shows that large explanatory power of the item covariates and large sample size led to smaller SE and RMSE of the item intercept parameter estimates obtained from JRT-DINA-LLTM.

Table 31. *Significant Effects in the Three-way ANOVA of the Item Intercept Parameter Estimates from the JRT-DINA-LLTM*

I	Source	SE			RMSE		
		F Statistics	*p*-value	Partial η^2	F Statistics	*p*-value	Partial η^2
20	G	7.406	0.007	0.031	5.156	0.024	0.022
	J	39.146	< 0.001	0.147	41.669	< 0.001	0.155
40	G	64.406	< 0.001	0.121	35.526	< 0.001	0.071
	J	62.129	< 0.001	0.117	37.039	< 0.001	0.073

Note. G = Explanatory power of item covariates; J = Sample size; I = Test length.

4.3.2 Item Interaction Parameter

The significant effects obtained from the mixed-effect ANOVA results on the bias of item interaction parameters are presented in Table 32. Under both test length conditions, a significant main effect was found among different model types with small effect sizes. The main effect of model type is presented in Figure 26. When test length was 20, all models

overestimated the item interaction parameter on average. Specifically, JRT-DINA-LLTM and DINA-LLTM had larger mean bias, while the two models with two-step estimation (i.e., JRT-DINA and DINA) had smaller mean bias. When test length was 40, however, the biases for JRT-DINA-LLTM and DINA-LLTM canceled out the most. In addition, JRT-DINA and DINA had negative mean biases while models omitting one covariate (i.e., JRT-DINA-LLTM-C, JRT-DINA-LLTM-D) had positive mean biases. As shown in Table B4 in the Appendix B, JRT-DINA-LLTM and DINA-LLTM had smaller mean bias only when test length was long ($I = 40$). This is reasonable because longer test length may provide more precise regression coefficient estimates, thus, more precise item interaction parameter estimates.

Table 32. *Significant Effects in the Mixed-Effect ANOVA Results on Bias of the Item Interaction Parameter Estimates*

I	Effect	Bias		
		F Statistics	p-value	Partial η^2
20	Model	6.389	0.002	0.027
40	Model	4.573	0.004	0.010

Note. I = Test length.

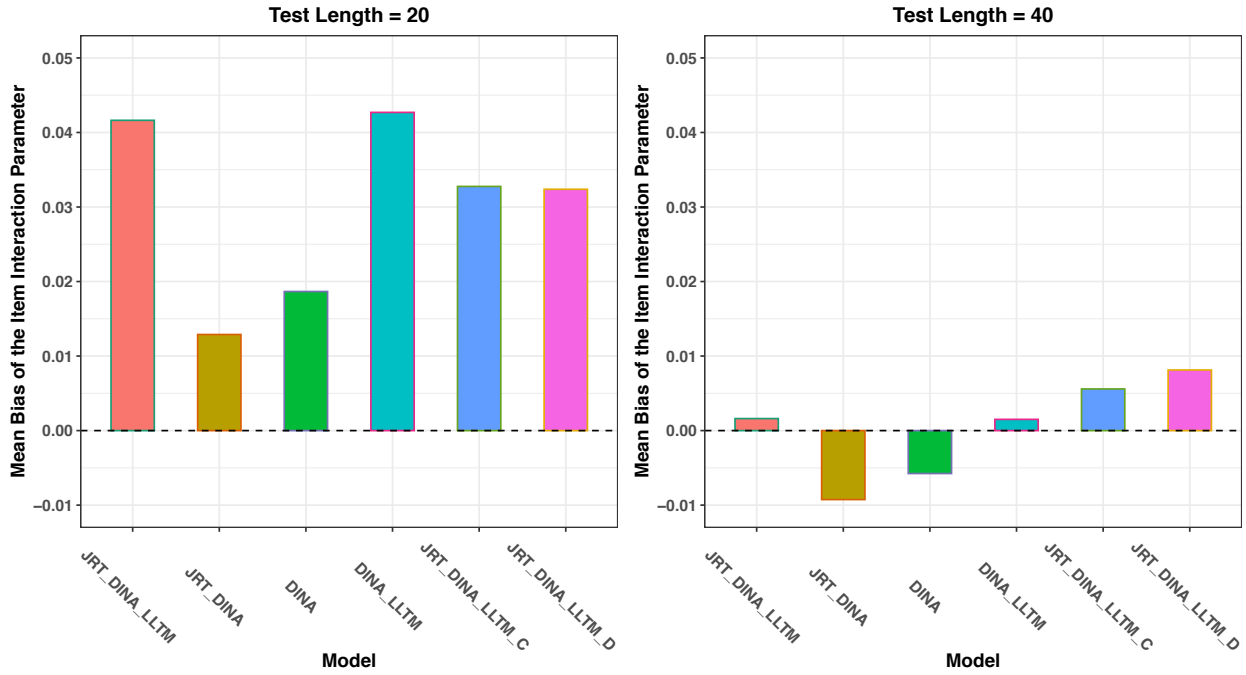


Figure 26. Marginal mean Bias of the item interaction parameter at each level of the test length.

The significant effects on SE and RMSE of item interaction parameter estimates with at least a small effect size are presented in Table 33. When test length was 20, a three-way interaction among model type, sample size and explanatory power of item covariates was found to be significant with small effect sizes for SE and RMSE. In general, based on Figure 27, all models yielded smaller SE and RMSE as sample size increased and explanatory power of item covariates increased. Among all models, models with one-step estimation (i.e., JRT-DINA-LLTM, DINA-LLTM, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D) had larger SE and RMSE than models with two-step estimation (i.e., JRT-DINA, DINA) when explanatory power of item covariates was small. However, when explanatory power of item covariates became large, mean SE and RMSE of JRT-DINA-LLTM, DINA-LLTM and JRT-DINA-LLTM-C became smaller than that of models with two-step estimation. JRT-DINA-LLTM-D, however, yielded the largest SE and RMSE when explanatory power of item covariates was large. This indicates that model with only the dichotomous covariate yielded the largest random error and total error on average.

Similar to the findings for the study conditions with test length of 20, a significant three-way interaction among model type, sample size, and explanatory power of item covariates was found for SE with a small effect size (partial η^2 = 0.018) when test length was 40. In addition, an interaction effect between model and explanatory power of item covariates and an interaction effect between model and sample size on RMSE were found to be significant with a medium effect size (partial η^2 = 0.067) and a small effective size (partial η^2 = 0.014), respectively. According to Figures 28 and 29, the advantages of models with two-step estimation when the explanatory power was small diminished when test length was long (I = 40). That is, all models showed comparable mean SE and RMSE when the explanatory power was small. In addition, larger sample size led to smaller SE and RMSE for all models just as that when test length was

20. The discrepancy among the models also remained the same, except that the JRT-DINA-LLTM-C had larger SE and RMSE than JRT-DINA-LLTM and DINA-LLTM while they were similar when test length was 20. JRT-DINA-LLTM-D still yielded the largest random error and total error on average when test length was 40.

Table 33. *Significant Effects in the Mixed-Effect ANOVA Results on SE and RMSE of the Item Interaction Parameter Estimates*

I	Effect	SE			RMSE		
		F Statistics	p-value	Partial η^2	F Statistics	p-value	Partial η^2
20	Model	8.798	< 0.001	0.037	10.772	< 0.001	0.045
	Model*G	22.619	< 0.001	0.090	47.726	< 0.001	0.173
	Model*J*G	3.762	0.025	0.016	5.711	0.004	0.024
	G	39.411	< 0.001	0.147			
	J	82.663	< 0.001	0.266	44.771	< 0.001	0.164
40	Model	113.586	< 0.001	0.195	67.947	< 0.001	0.127
	Model*G	89.672	< 0.001	0.161	33.854	< 0.001	0.067
	Model*J	11.874	< 0.001	0.025	6.600	0.001	0.014
	Model*J*G	8.439	< 0.001	0.018			
	G	34.463	< 0.001	0.069	15.249	< 0.001	0.032
	J	61.858	< 0.001	0.117	35.310	< 0.001	0.070

Note. G = Explanatory power of item covariates; J = Sample size; I = Test length.

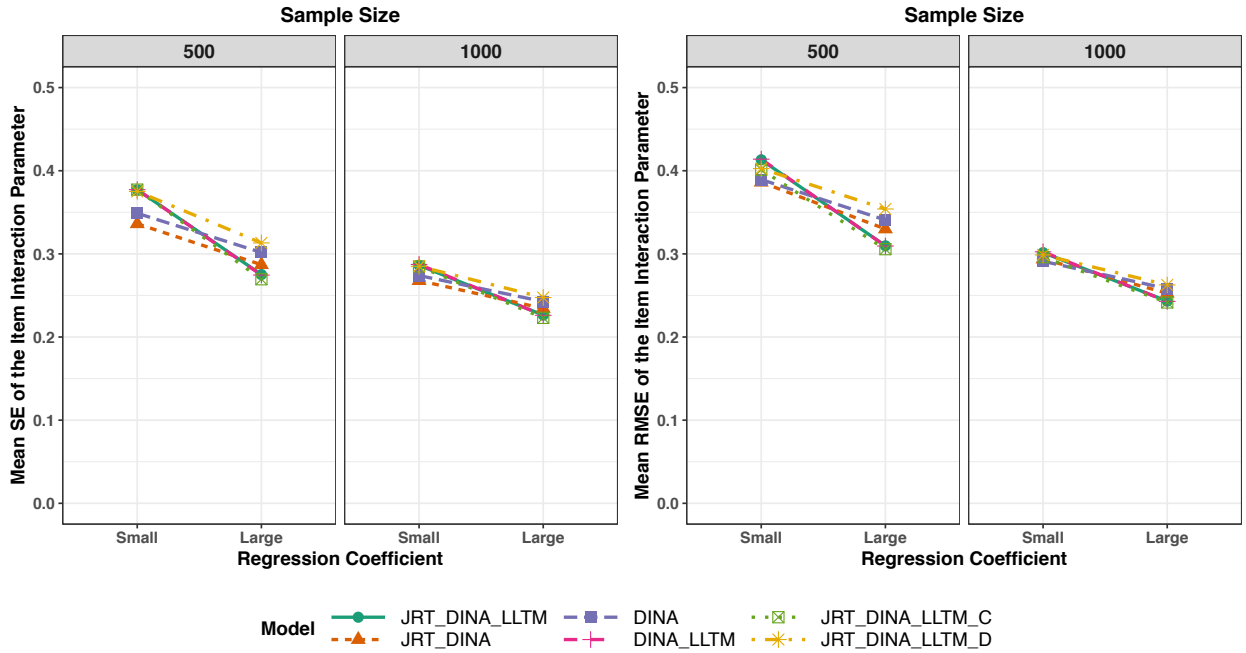


Figure 27. Marginal mean SE and RMSE of the item interaction parameter at each level of the manipulated factors (I = 20).

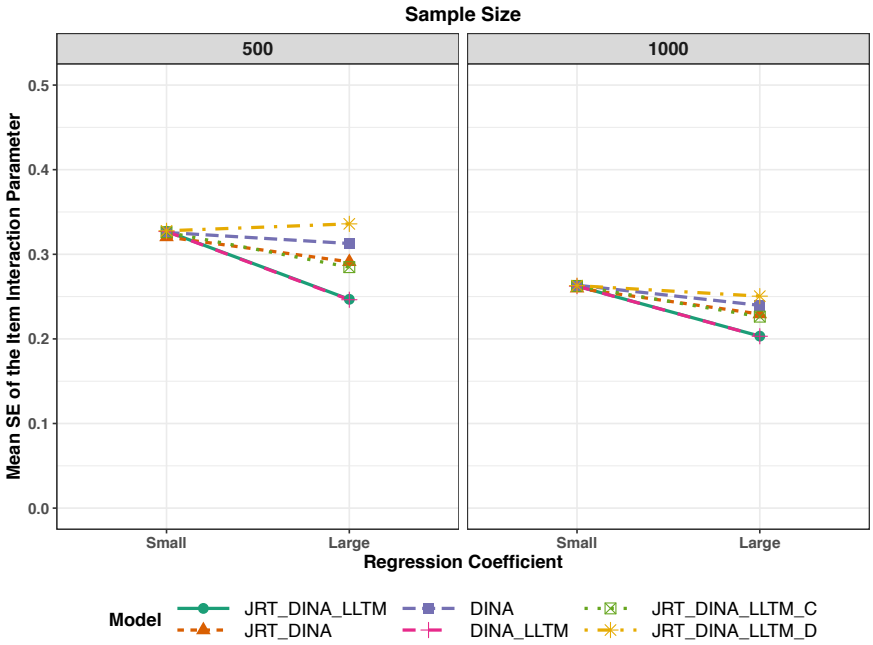


Figure 28. Marginal mean SE of the item interaction parameter at each level of the manipulated factors ($I = 40$).

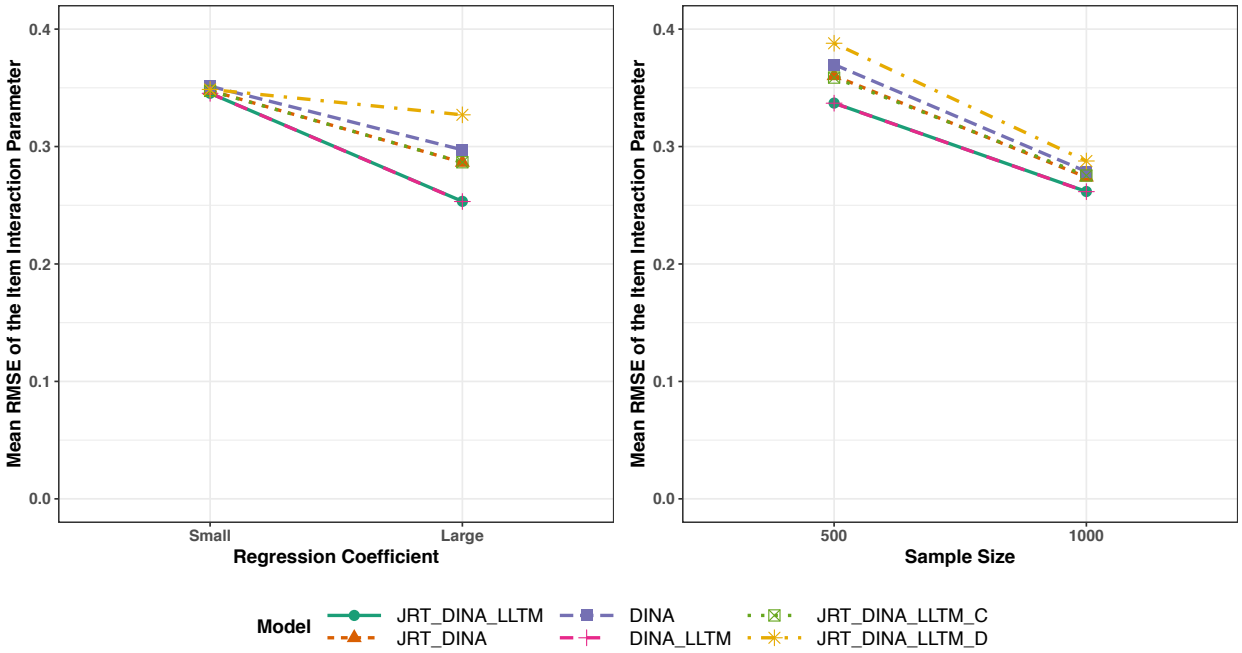


Figure 29. Marginal mean RMSE of the item interaction parameter at each level of the manipulated factors ($I = 40$).

The three-way ANOVA was conducted to examine the impact of the manipulated factors on the parameter recovery of the item interaction parameter estimates from the JRT-DINA-

LLTM. As presented in Table 34, the explanatory power of the item covariates and sample size had an interaction effect on SE when test length was 20 with a small effect size (Partial η^2 = 0.024). The explanatory power of the item covariates had a small effect size on RMSE (Partial η^2 = 0.022). Sample size had a large effect size on RMSE (Partial η^2 = 0.155). When test length was 40, the explanatory power of the item covariates had a large effect size on SE (Partial η^2 = 0.275) and a medium effect size on RMSE (Partial η^2 = 0.071), respectively. Sample size had a large effect size on SE (Partial η^2 = 0.186) and a medium effect size on RMSE (Partial η^2 = 0.073), respectively. An inspection on the marginal means of the SE and RMSE shows that large explanatory power of the item covariates and large sample size led to smaller SE and RMSE of the item interaction parameter estimates obtained from JRT-DINA-LLTM.

Table 34. *Significant Effects in the Three-way ANOVA Results on Bias, SE and RMSE of the Item Interaction Parameter Estimates in JRT-DINA-LLTM*

I	Source	SE			RMSE		
		F Statistics	*p*-value	Partial η^2	F Statistics	*p*-value	Partial η^2
20	G				5.156	0.024	0.022
	J				41.669	< 0.001	0.155
	G*J	5.656	0.018	0.024			
40	G	177.298	< 0.001	0.275	35.526	< 0.001	0.071
	J	106.835	< 0.001	0.186	37.039	< 0.001	0.073

Note. G = Explanatory power of item covariates; I = Test length; R = Correlation between ability and speed.

4.3.2 Item Intensity Parameter

The significant effects obtained from the mixed-effect ANOVA results on the bias of item intensity parameter estimates are presented in Table 35. Under both test length conditions, a significant main effect was found among different model types with small effect sizes. The main effect of model type is presented in Figure 30. Under both test length conditions, the lognormal model had slightly larger mean biases, but the mean biases of the item intensity parameter

estimate for each data fitting model were trivial. In addition, all the data fitting models tended to underestimate the item intensity parameter estimates given the negative mean biases.

Table 35. *Significant Effects in the Mixed-Effect ANOVA Results on Bias of the Item Intensity Parameter Estimates*

I	Effect	Bias		
		F Statistics	*p*-value	Partial η^2
20	Model	84.238	< 0.001	0.027
40	Model	27.045	< 0.001	0.055

Note. I = Test length.

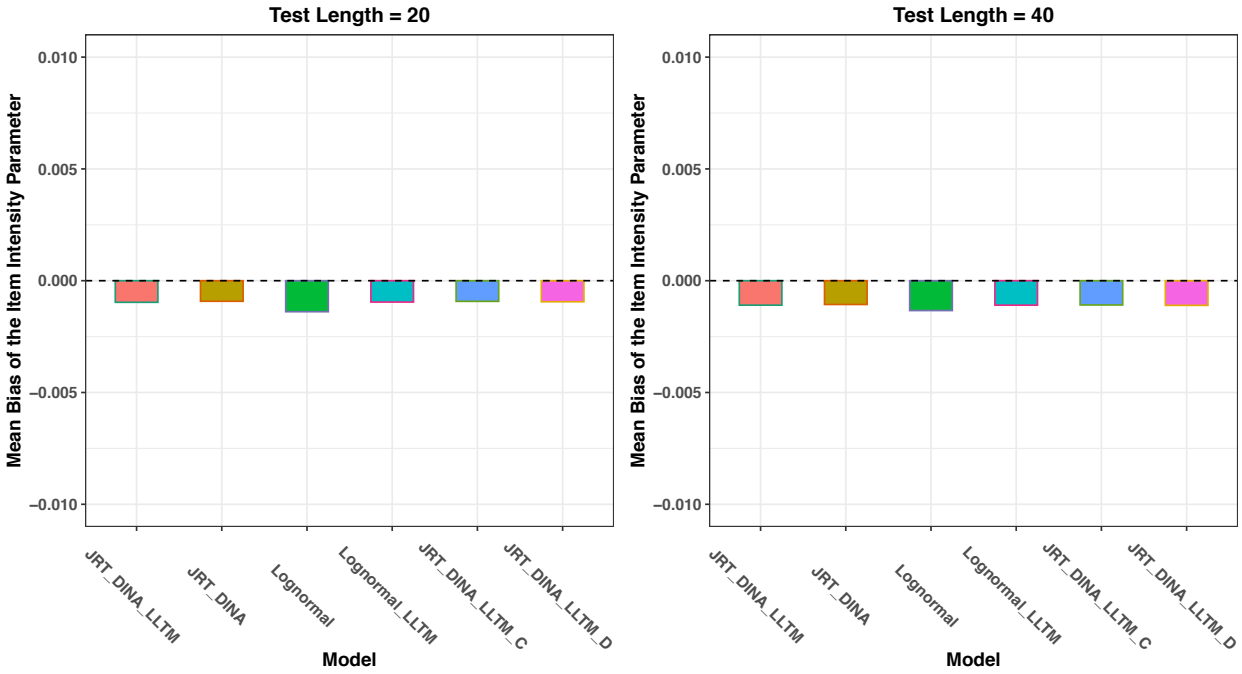


Figure 30. Marginal mean bias of the item intensity parameter at each level of the manipulated factors.

The significant effects obtained from the mixed-effect ANOVA results on the SE and RMSE of item intensity parameter estimates are presented in Table 36. Under both test length conditions, sample size was found to have a large effect size on both SE and RMSE of the item intensity parameter estimates. As shown in Figures 31 to 32, when sample size was 1000, both the mean SE and RMSE of the item intensity parameter estimates were smaller than that when

sample size was 500. This indicates that large sample size tended to decrease the magnitude of the random error and total error of the item intensity parameter estimates.

Table 36. *Significant Effects in the Mixed-Effect ANOVA Results on SE and RMSE of the Item Intensity Parameter Estimates*

		SE			RMSE		
I	Effect	F Statistics	*p*-value	Partial η^2	F Statistics	*p*-value	Partial η^2
20	J	334.846	< 0.001	0.595	267.136	< 0.001	0.540
40	J	885.721	< 0.001	0.654	686.493	< 0.001	0.595

Note. I = Test length; J = Sample size.

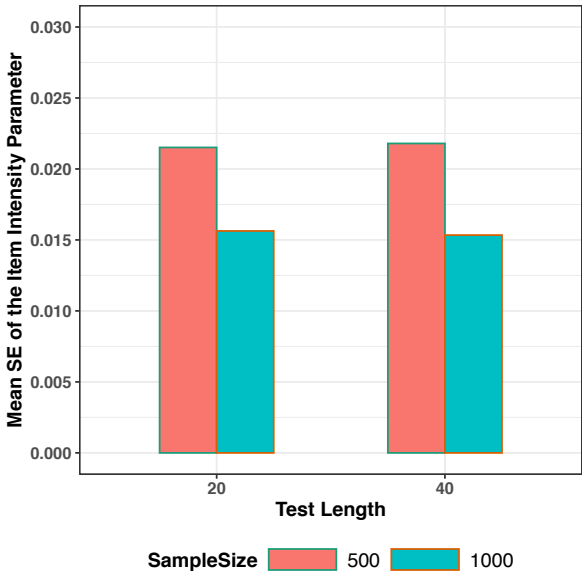


Figure 31 Marginal mean SE of the item intensity parameter at each level of sample sizes.

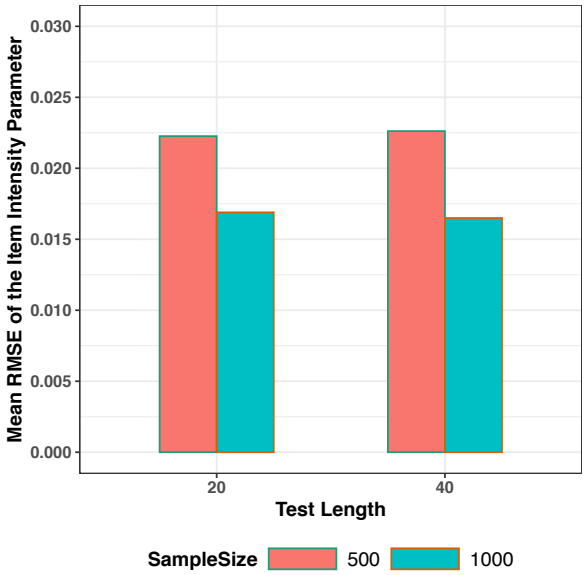


Figure 32. Marginal mean RMSE of the item intensity parameter at each level of sample sizes.

The impact of manipulated factors on the parameter recovery of the item intensity parameter estimates obtained from JRT-DINA-LLTM was examined using the three-way ANOVA. The results are shown in Table 37. Specifically, sample size was found to have a large effect size on bias, SE, and RMSE of the item intensity parameter estimates obtained from JRT-DINA-LLTM when test length was 20 and 40. In addition, the explanatory power of item covariates had large effect sizes on bias and SE and a small effect size on RMSE when test length was 40. An inspection of the marginal means indicates that large sample sizes led to smaller absolute values of the mean bias, SE, and RMSE. In addition, large explanatory power of item covariates also led to smaller absolute values of the mean bias, SE, and RMSE when test length was 40.

Table 37. *Significant Effects in the Three-way ANOVA Results on Bias, SE and RMSE of the Item Intensity Parameter Estimates in JRT-DINA-LLTM*

		Bias			SE			RMSE		
I	Source	F Statistics	*p*-value	Partial η^2	F Statistics	*p*-value	Partial η^2	F Statistics	*p*-value	Partial η^2
20	J	35.226	< 0.001	.134	336.766	< 0.001	0.596	266.964	< 0.001	0.539

40	G	354.604	< 0.001	.602	177.298	< 0.001	0.275	9.175	0.003	0.019
	J	56.065	< 0.001	.107	106.835	< 0.001	0.186	686.715	< 0.001	0.595

Note. G = Explanatory power of item covariates; I = Test length; J = Sample size.

4.3.3 Regression coefficients for the item parameters

In the data generating model (i.e., JRT-DINA-LLTM), three item parameters, including item intercept parameter, item interaction parameter, and item intensity parameter, were decomposed into a linear combination of regression coefficients and item covariates. In total, there were two item covariates (one dichotomous and one continuous) and nine regression parameters in the true data generating model. In addition to JRT-DINA-LLTM, five other competing models can be categorized as models omitting one item covariate (i.e., JRT-DINA-LLTM-C and JRT-DINA-LLTM-D) and models using two-step estimation of the regression coefficients (i.e., JRT-DINA and DINA + Lognormal) and model omitting the second-level model (i.e., DINA-LLTM + Lognormal-LLTM and DINA + Lognormal). A summary of estimated regression parameters in each data-fitting model is provided in Table 38. The recovery of the regression coefficients is examined by plotting the marginal means of bias, SE and RMSE of the parameter estimates from data-fitting models with the presence of the parameter in following paragraphs.

Table 38. *Summary of Regression Parameters in Each Data-fitting Model*

Model	Regression Intercept			Regression Coefficients for Continuous Covariate			Regression Coefficients for Dichotomous Covariate		
	$\mathbf{A}_{0(\beta)}$	$\mathbf{A}_{0(\delta)}$	$\mathbf{A}_{0(\zeta)}$	$\mathbf{A}_{c(\beta)}$	$\mathrm{A}_{c(\delta)}$	$\mathrm{A}_{c(\zeta)}$	$\mathrm{A}_{d(\beta)}$	$\mathrm{A}_{d(\delta)}$	$\mathrm{A}_{d(\zeta)}$
JRT-DINA-LLTM	√	√	√	√	√	√	√	√	√
JRT-DINA	√	√	√	√	√	√	√	√	√
DINA + Lognormal	√	√	√	√	√	√	√	√	√

DINA-LLTM + Lognormal-LLTM	√	√	√	√	√	√	√	√	√
JRT-DINA-LLTM-C	√	√	√	√	√	√	×	×	×
JRT-DINA-LLTM-D	√	√	√	×	×	×	√	√	√

Note. √ indicates the presence of the parameter; × indicates the absence of the parameter.

4.3.3.1 Regression Parameters for the Item Intercept Parameter

The regression parameters for the item intercept parameter include the regression coefficient for continuous covariate $\mathbf{A}_{c(\beta)}$, regression coefficient for dichotomous covariate $\mathrm{A}_{d(\beta)}$, and regression intercept $\mathbf{A}_{0(\beta)}$. The mean bias, SE and RMSE of $\mathbf{A}_{c(\beta)}$ are presented in Figures 33 to 35. Among the four manipulated factors, test length and explanatory power of item covariates affected the marginal means of biases, while mean biases did not change much at different levels of sample size and correlation between ability and speed. Specifically, when test was long and explanatory power of regression coefficients was large, mean biases were closer to 0. In addition, all models tended to overestimate $\mathbf{A}_{c(\beta)}$ given the positive mean biases at all levels of manipulated factors except when test length was 40. Among the data fitting models, JRT-DINA-LLTM and DINA-LLTM had smaller mean biases than the two models using two-step estimation (i.e., JRT-DINA, DINA), while JRT-DINA-LLTM-C omitting the dichotoumous covariate yielded slightly larger mean biases than JRT-DINA-LLTM and DINA-LLTM.

The mean SE of $\mathbf{A}_{c(\beta)}$, however, showed the opposite pattern. Models with covariates (i.e., JRT-DINA-LLTM, DINA-LLTM, and JRT-DINA-LLTM-C) had larger mean SE than the JRT-DINA and DINA model. This is expected because models with covariates are more complex. In terms of total error, JRT-DINA-LLTM-C yielded the smallest RMSE while JRT-DINA and DINA model yielded the largest RMSE at all levels of manipulated factors. In sum,

the models with two-step estimation had larger systematic error and total error in the estimation of the regression coefficient of the continuous covariate of item intercept parameter.

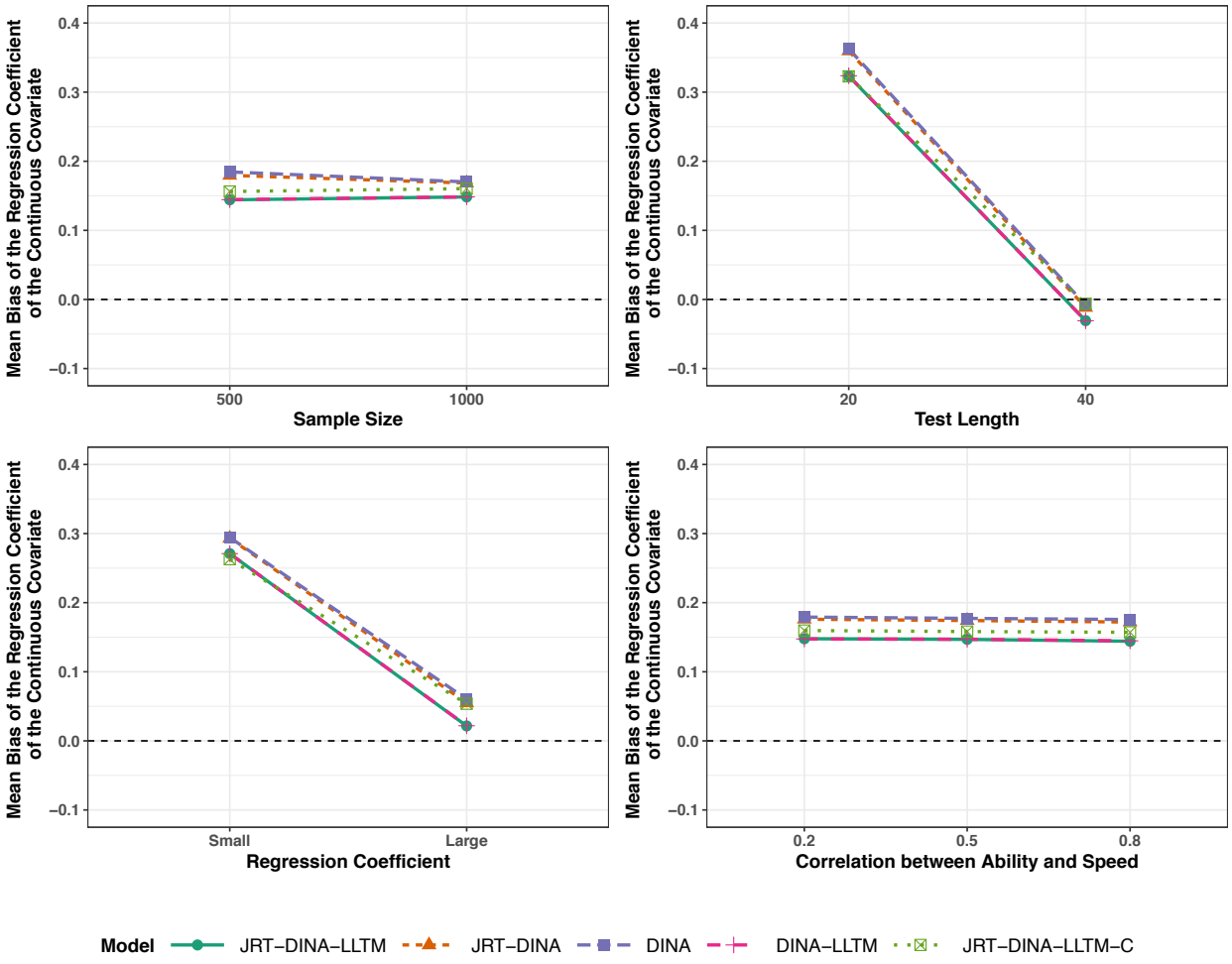


Figure 33. Mean bias of the regression coefficient of the continuous covariate of the item intercept parameter.

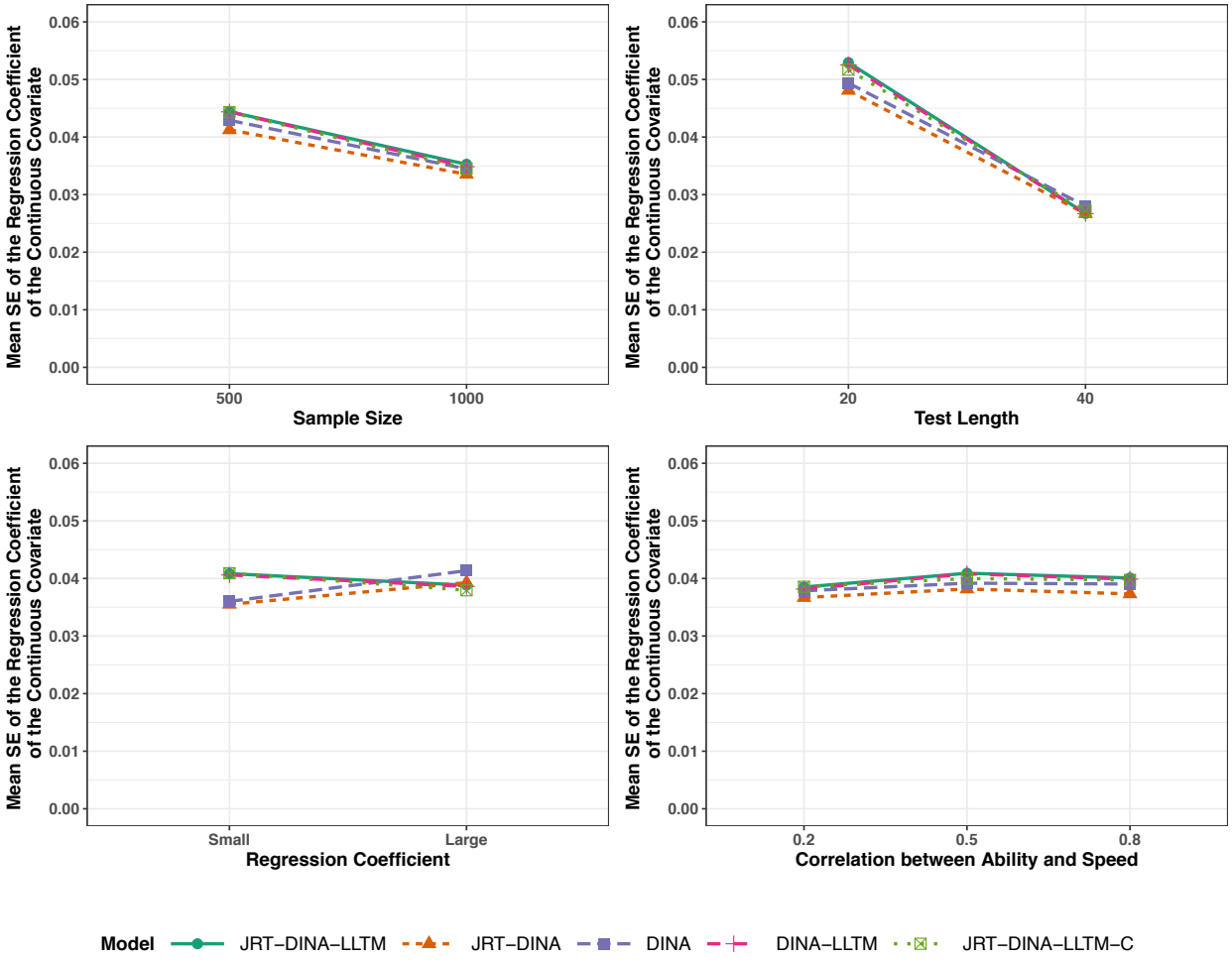


Figure 34. Mean SE of the regression coefficient of the continuous covariate of the item intercept parameter.

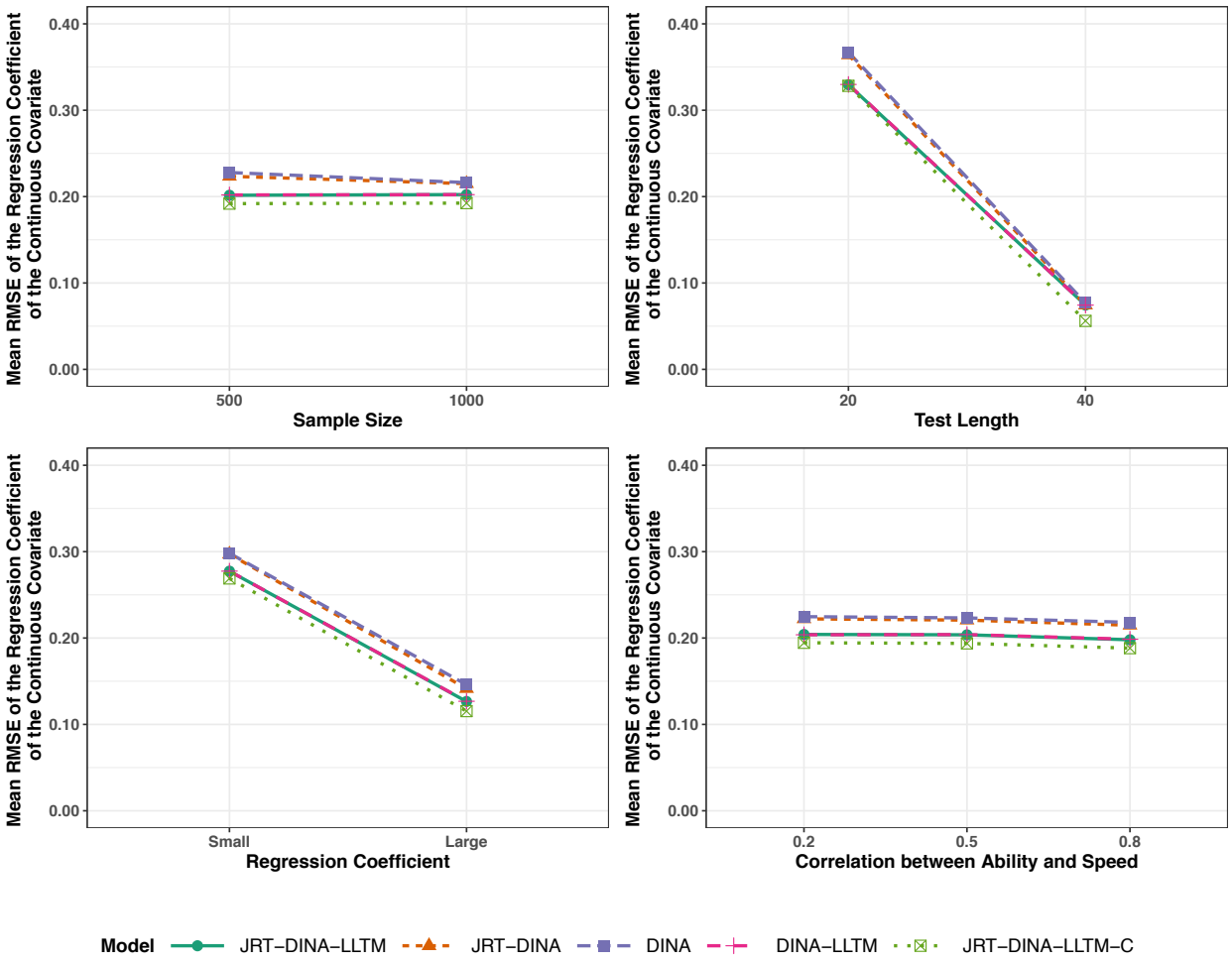


Figure 35. Mean RMSE of the regression coefficient of the continuous covariate of the item intercept parameter.

The mean bias, SE and RMSE for the regression coefficient estimates of the dichotomous covariate of the item intercept parameter (i.e., $A_{d(\beta)}$) are shown in Figures 36 to 38. The mean biases, SE and RMSE for all the models had similar patterns with those for the regression coefficient estimates of the continuous covariate. The only difference is that JRT-DINA-LLTM-D had the largest mean biases in all levels of manipulated factors except when the regression coefficients were small. In addition, JRT-DINA-LLTM-D had the largest mean RMSE in all

levels of manipulated factors. This indicates that omitting the continuous covariates did affect the estimation accuracy of the regression coefficient of dichotomous covariates.

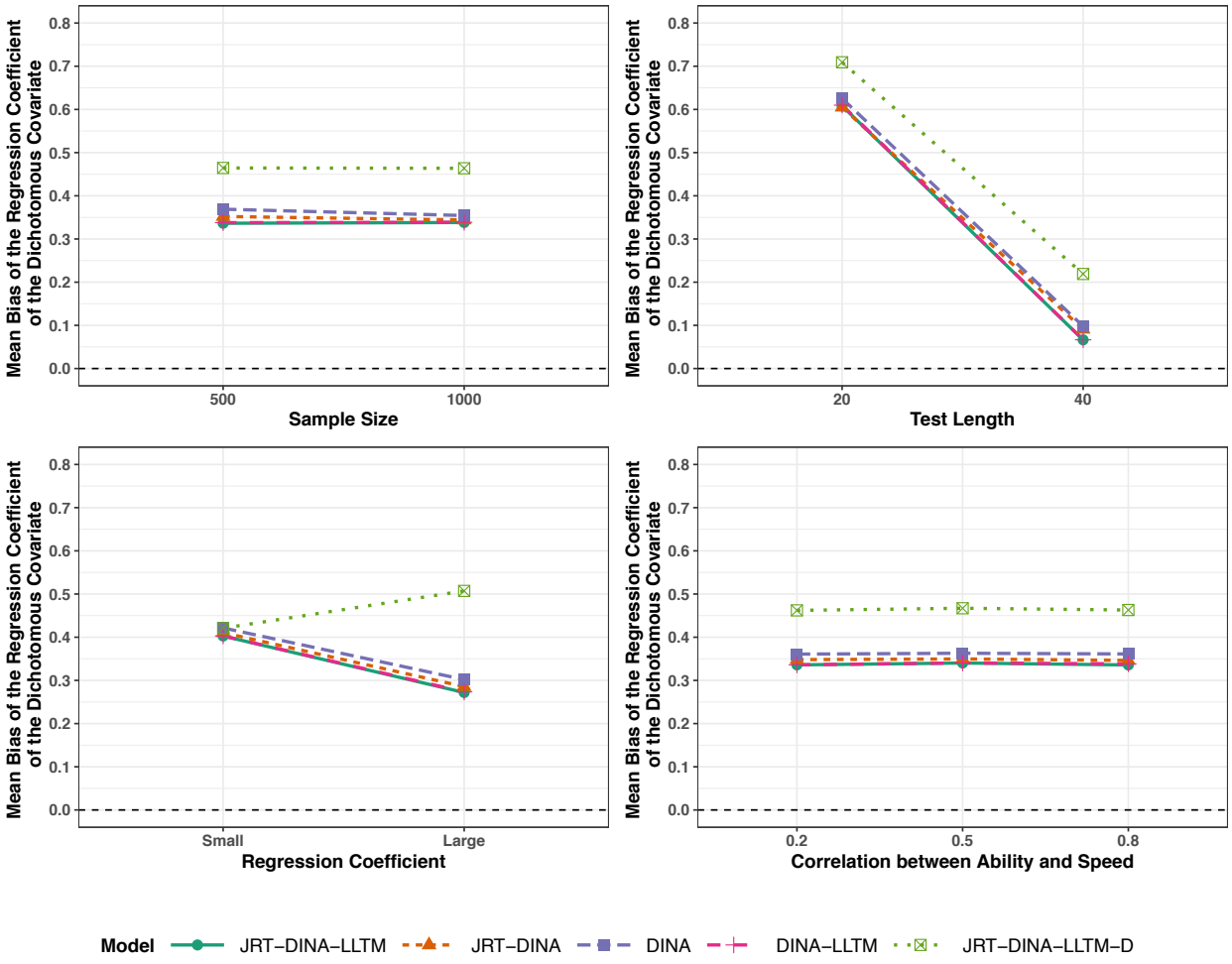


Figure 36. Mean Bias of the regression coefficient of the dichotomous covariate of the item intercept parameter.

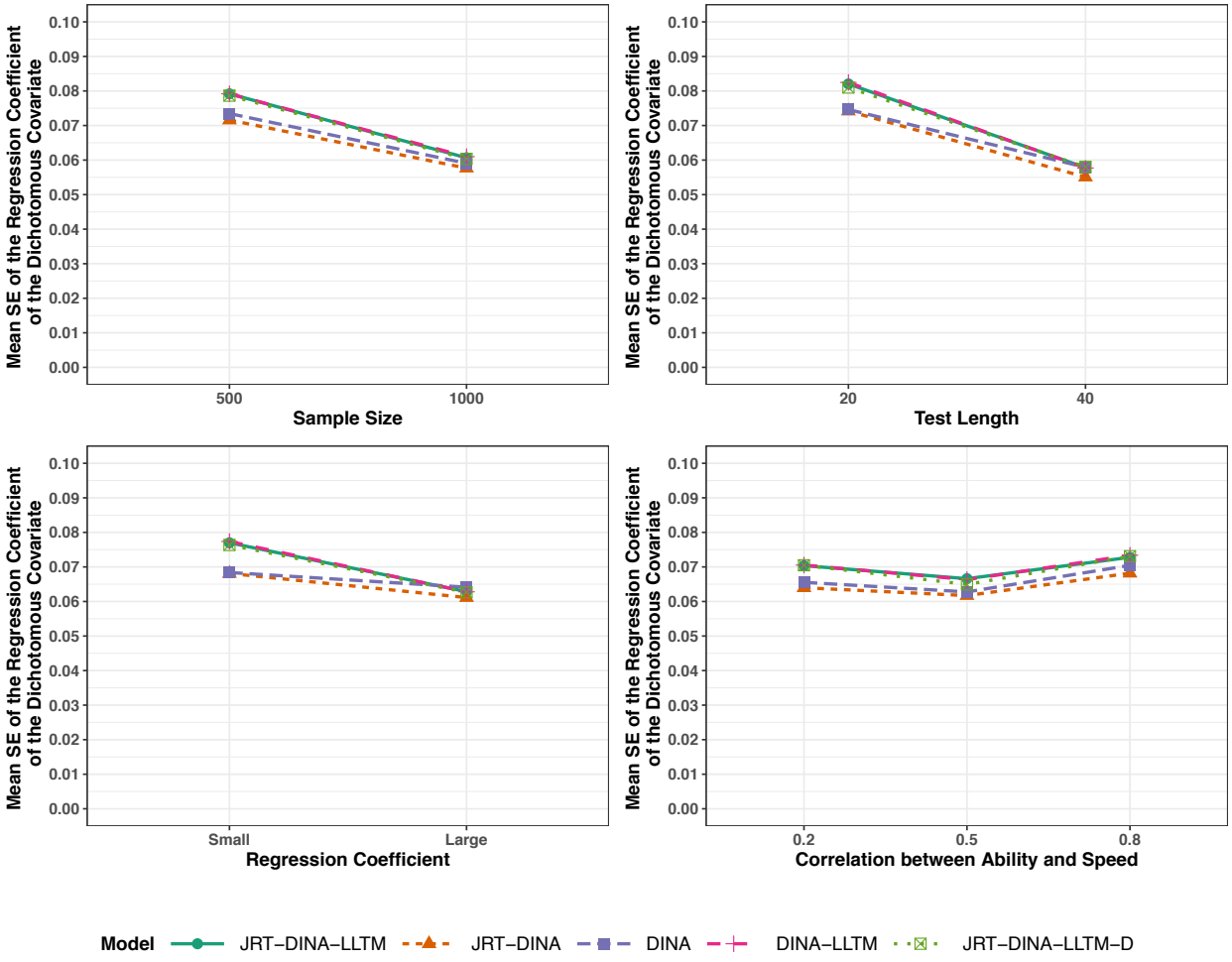


Figure 37. Mean SE of the regression coefficient of the dichotomous covariate of the item intercept parameter.

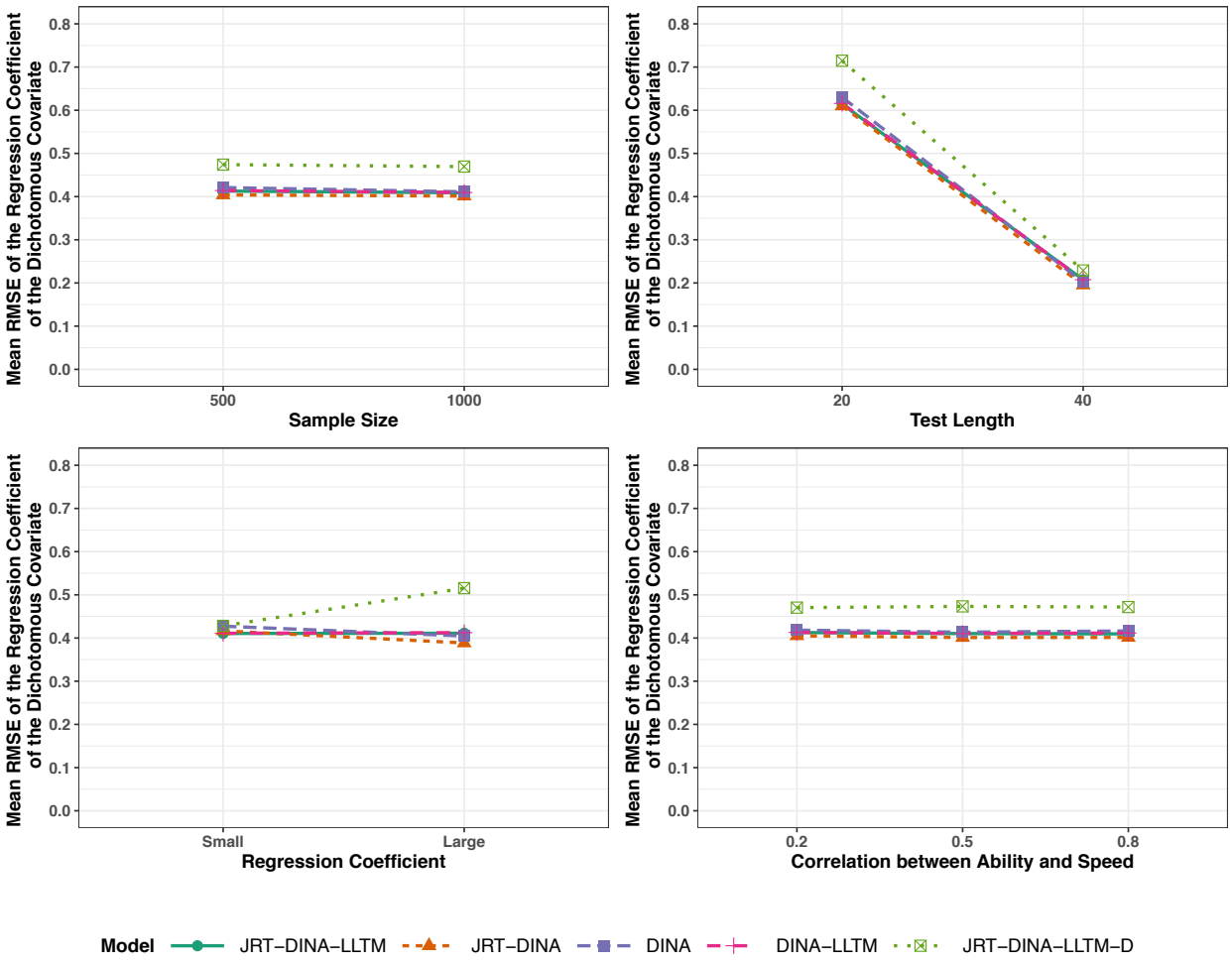


Figure 38. Mean RMSE of the regression coefficient of the dichotomous covariate of the item intercept parameter.

The mean bias, SE and RMSE of the regression intercept parameter estimates (i.e., $\mathbf{A}_{0(\beta)}$) at each level of the manipulated factor is shown in Figures 39 to 41. The most notable observation is that JRT-DINA-LLTM-C, the model omitting the dichotomous covariate, had the largest mean bias and RMSE of the regression intercept parameter estimates in all simulated conditions. Although the regression coefficient estimate of $\mathbf{A}_{c(\beta)}$ in JRT-DINA-LLTM-C was well recovered, the omitting of the dichotomous covariate affected the estimation accuracy of $\mathbf{A}_{0(\beta)}$. Yet, this is not critical given that statistical inference on the item covariate focuses on

$\mathbf{A}_{c(\beta)}$. In addition, all models understimated the regression intercept parameter estimates given the negative mean biases under all simulated conditions. This is expected because both regression coefficient estimates for the continuous and dichotomous covariates were over-estimated while the item parameters recovered well as shown in the previous section.

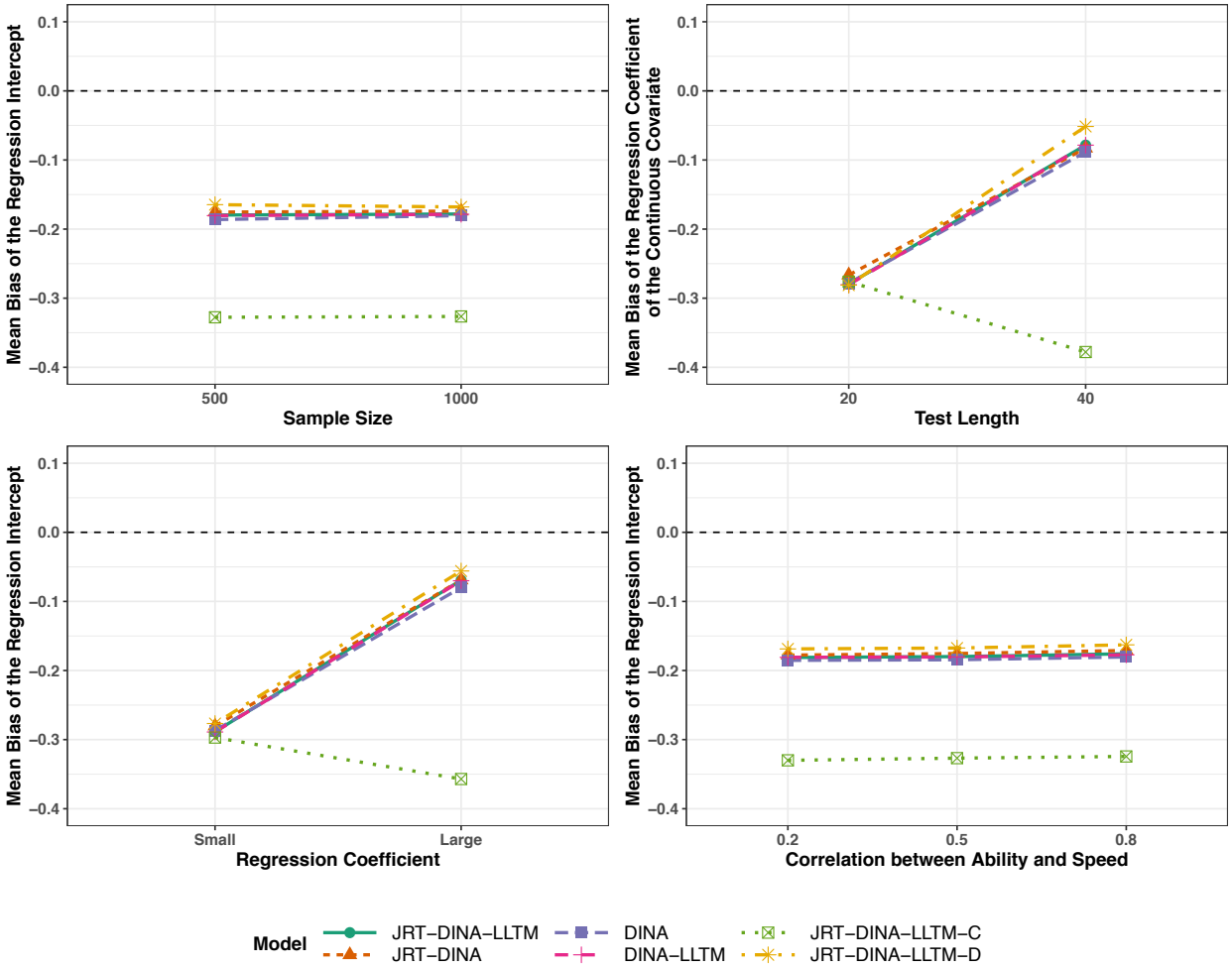


Figure 39. Mean Bias of the regression intercept of the item intercept parameter.

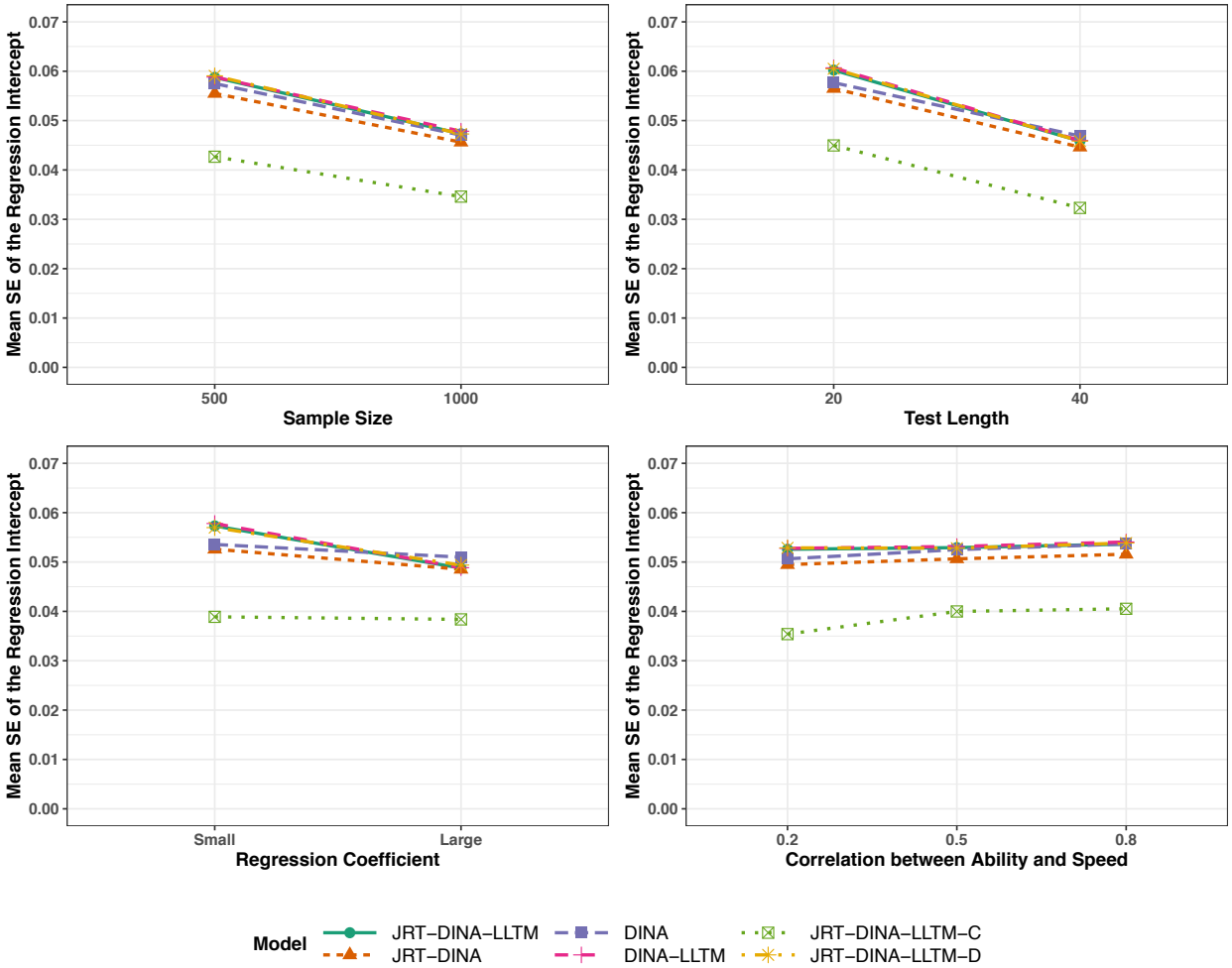


Figure 40. Mean SE of the regression intercept of the item intercept parameter.

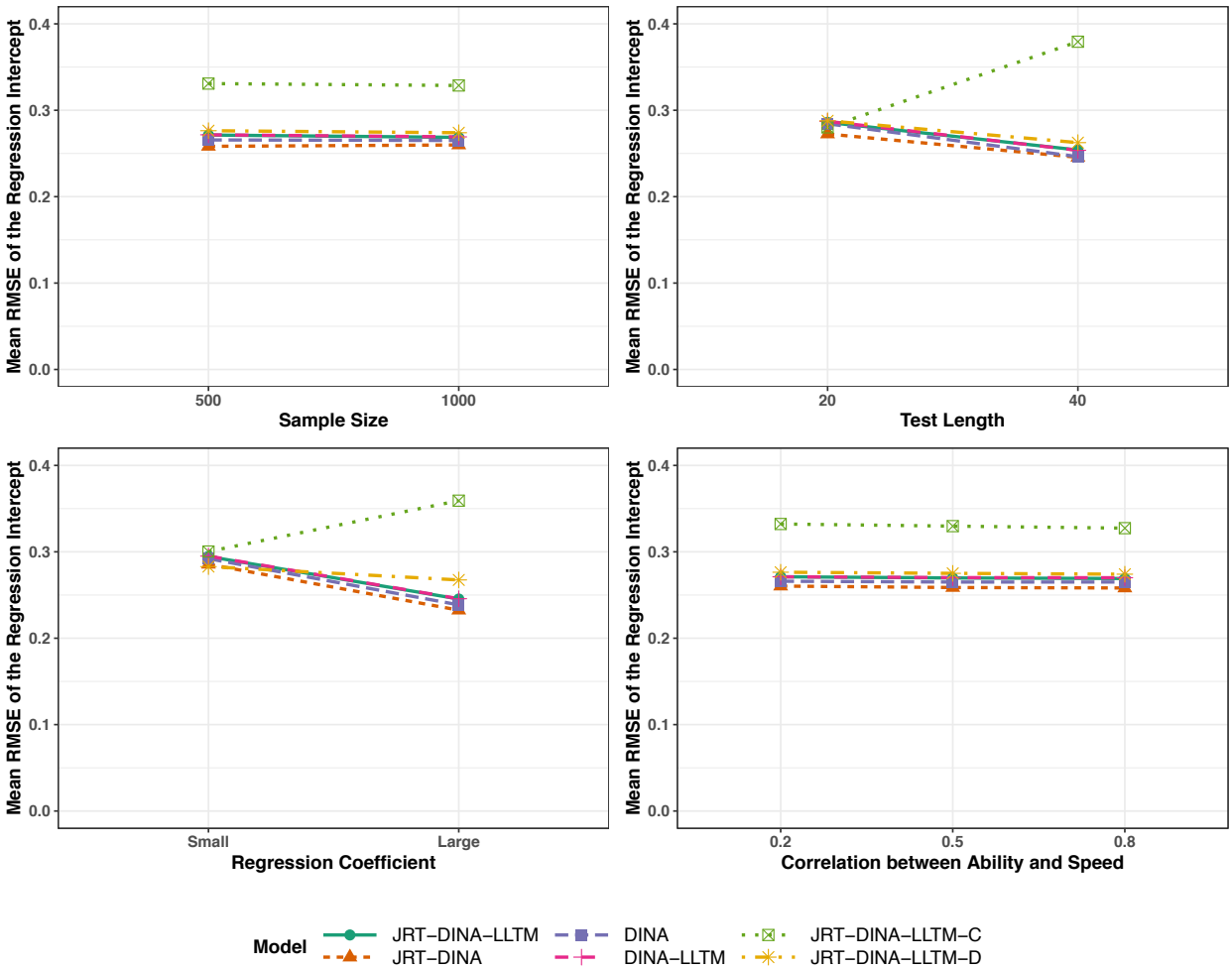


Figure 41. Mean RMSE of the regression intercept of the item intercept parameter.

4.3.3.2 Regression Coefficients for the Item Interaction Parameter

Regression parameters of the item interaction parameter include the regression coefficient for the concinuous covariate (i.e., $A_{c(\delta)}$), regression coefficient for the dichotomous covariate (i.e., $A_{d(\delta)}$), regression intercept (i.e., $\mathbf{A}_{0(\delta)}$). The mean biases, SE, RMSE of $A_{c(\delta)}$ are presented in Figures 42 to 44. Different from that of the item intercept parameter, mean biases for JRT-DINA-LLTM and DINA-LLTM were larger than models using two-step estimation and the model omitting one covariate except when test length was long and explanatory power of

item covariate was small. Similarly, the mean SE and RMSE for JRT-DINA-LLTM and DINA-LLTM were larger than those from the other models under all simulated conditions, which indicates that these two models had the largest random errors and systematic errors. It is expected that the random errors of these two models in the estimation of $A_{c(\delta)}$ would be larger due do the model complexity. However, larger systematic errors were not expected and could be due to the constraint of the item interaction parameter to be set as positive in the MCMC estimation.

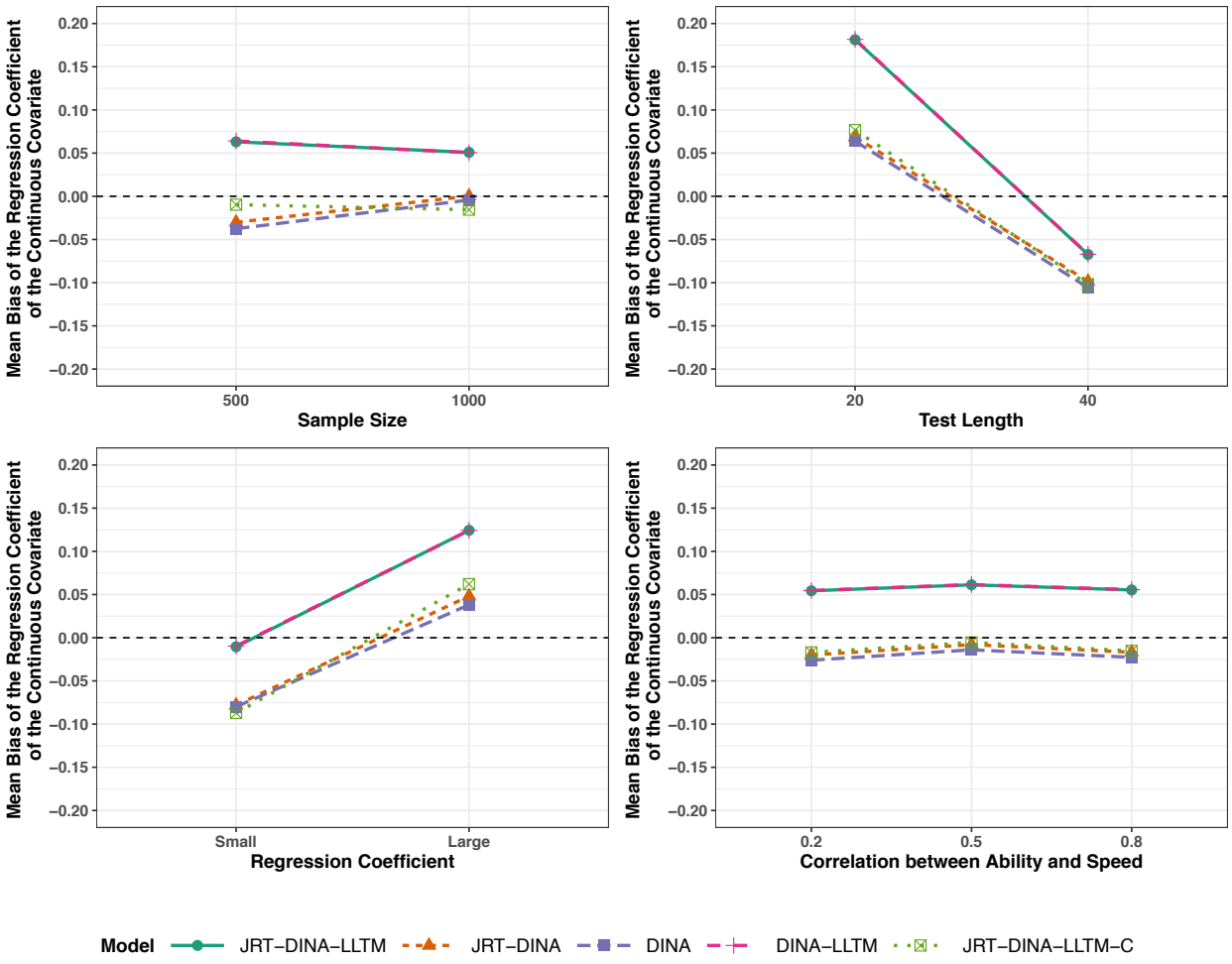


Figure 42. Mean bias of the regression coefficient of the continuous covariate of the item interaction parameter.

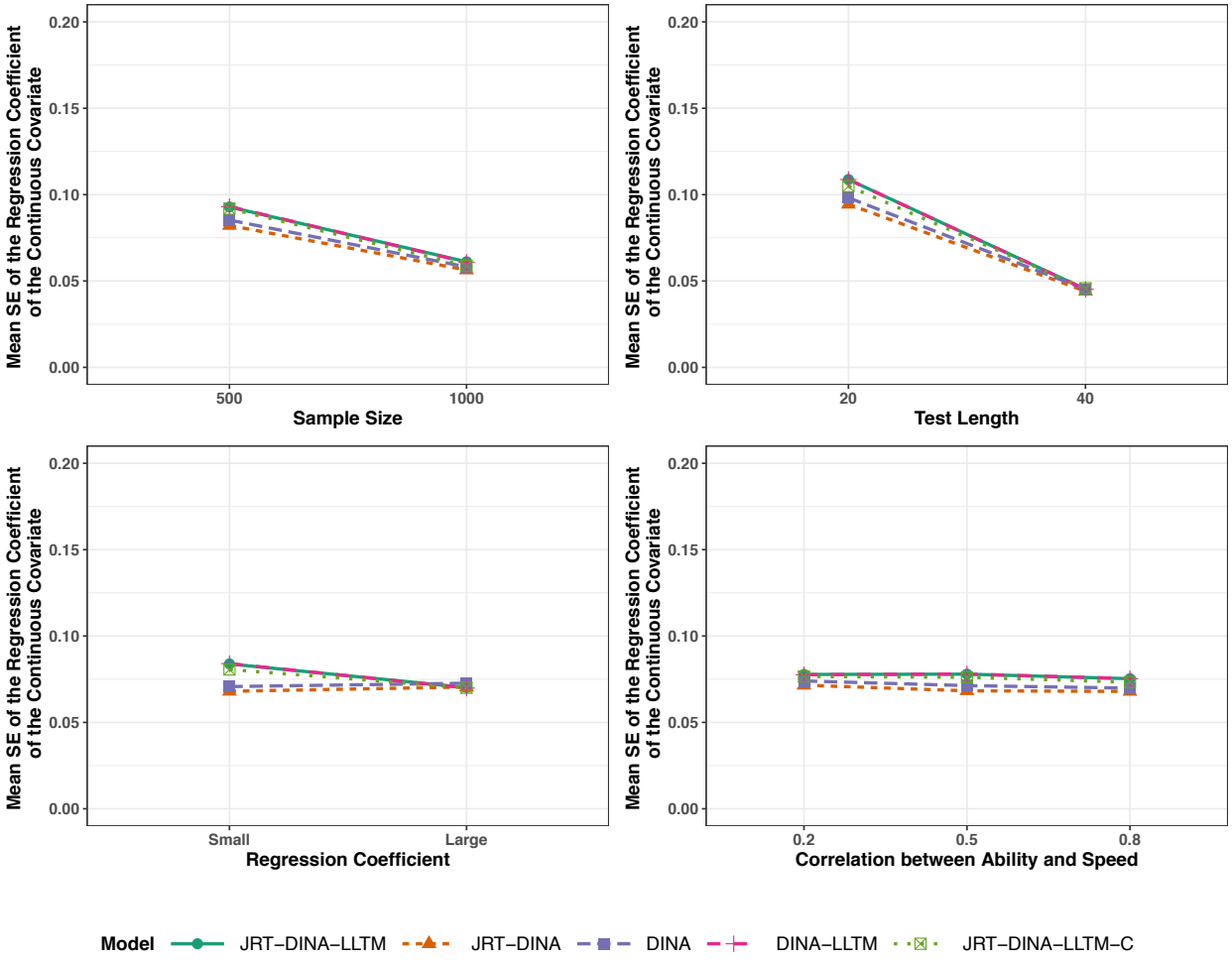


Figure 43. Mean SE of the regression coefficient of the continuous covariate of the item interaction parameter.

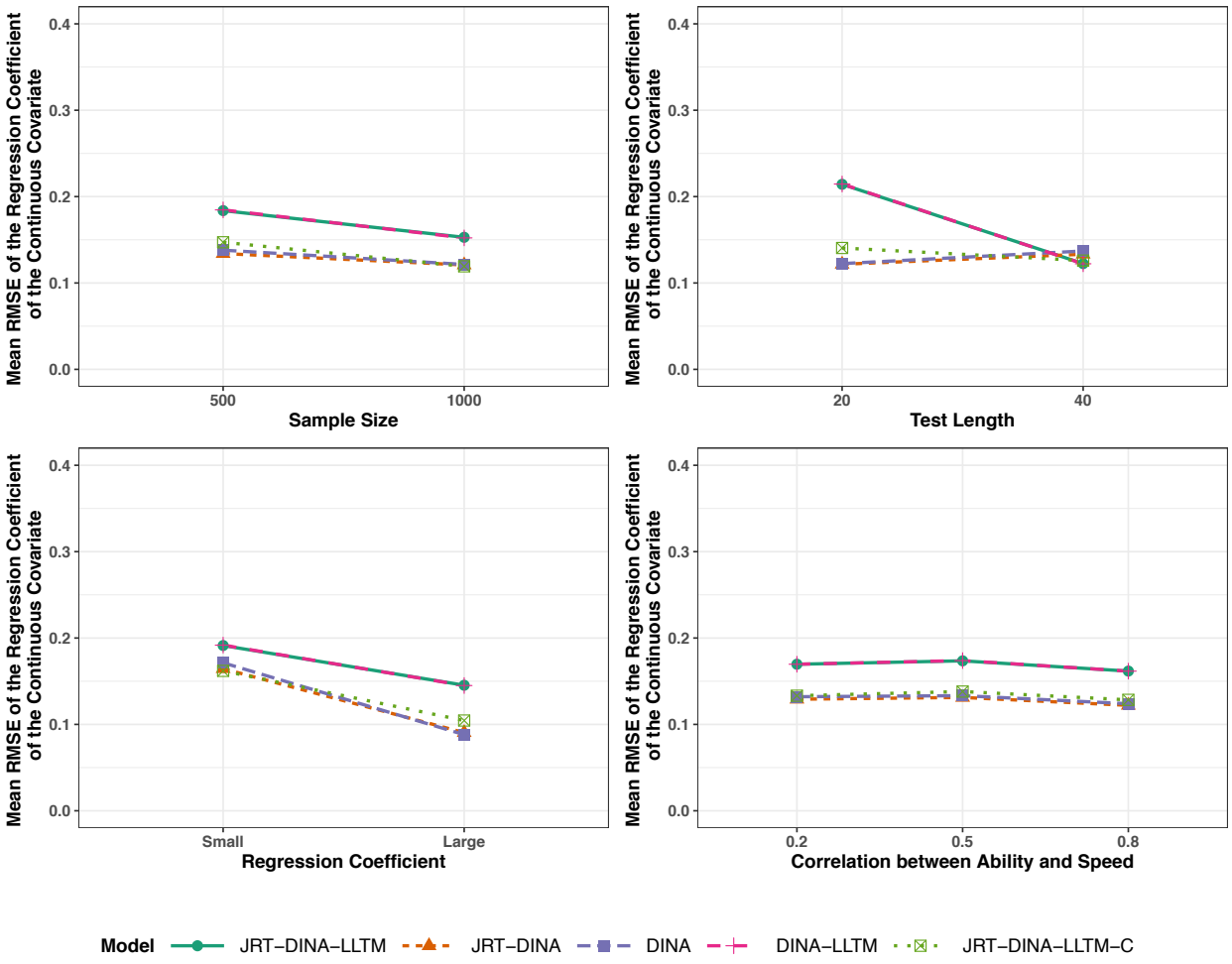


Figure 44. Mean RMSE of the regression coefficient of the continuous covariate of the item interaction parameter.

The mean biases, SE, RMSE of regression coefficient parameter estimates $A_{d(\delta)}$ at each level of the manipulated factors are presented in Figures 45-47. As shown in Figure 45, the mean biases for the three models incorporating ite covariates (i.e., JRT-DINA-LLTM, DINA-LLTM, and JRT-DINA-LLTM-D) were larger than those from the models using two-step estimation (i.e., JRT-DINA and DINA) except when test length was long and explanatory power of item covariates was large. This indicates that JRT-DINA and DINA had less systematic error in the estimation of $A_{d(\delta)}$, which is the same as that of $A_{c(\delta)}$. In addition, JRT-DINA-LLTM and

DINA-LLTM overestimated $A_{d(\delta)}$, while JRT-DINA-LLTM-D underestimated $A_{d(\delta)}$ due to the omission of the continuous covariate at different levels of sample size and correlation between ability and speed.

Similar to the previous mentioned regression coefficient parameters, the three models incorporating item covariates (i.e., JRT-DINA-LLTM, DINA-LLTM, and JRT-DINA-LLTM-D) had larger SE than models using the two-step estimation (i.e., JRT-DINA and DINA) due to model complexity, as shown in Figure 46. In Figure 47, the mean RMSE of the parameter estimates obtained from JRT-DINA-LLTM and DINA-LLTM was larger than JRT-DINA and DINA.

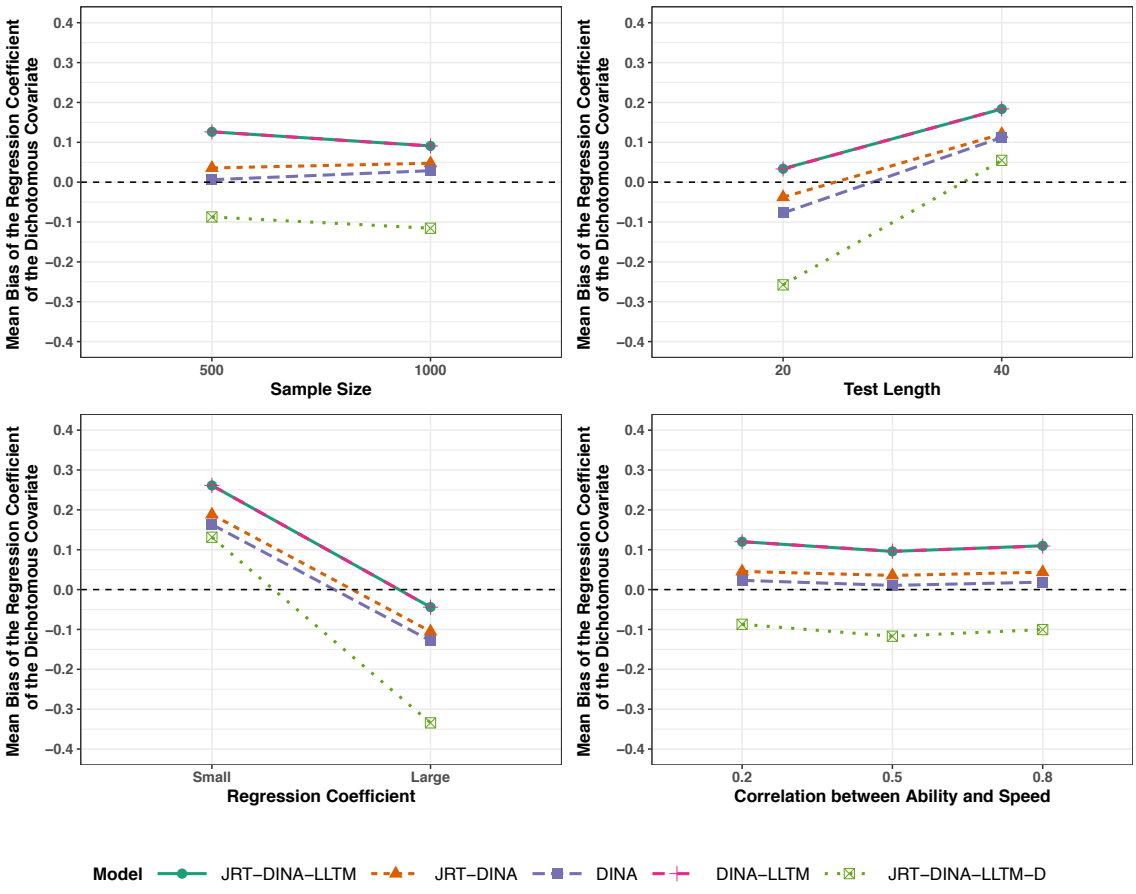


Figure 45. Mean Bias of the regression coefficient of the dichotomous covariate of the item interaction parameter.

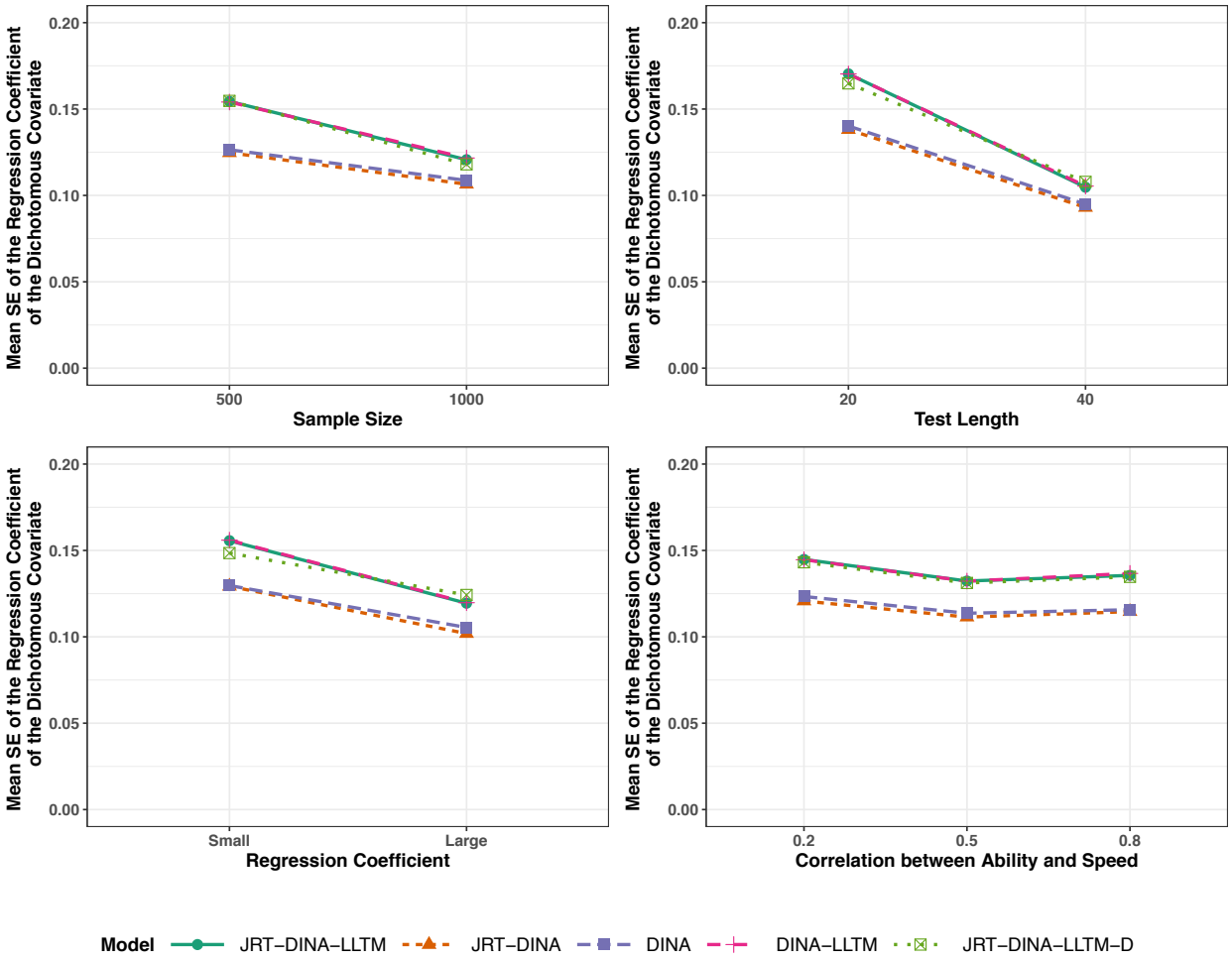


Figure 46. Mean SE of the regression coefficient of the dichotomous covariate of the item interaction parameter.

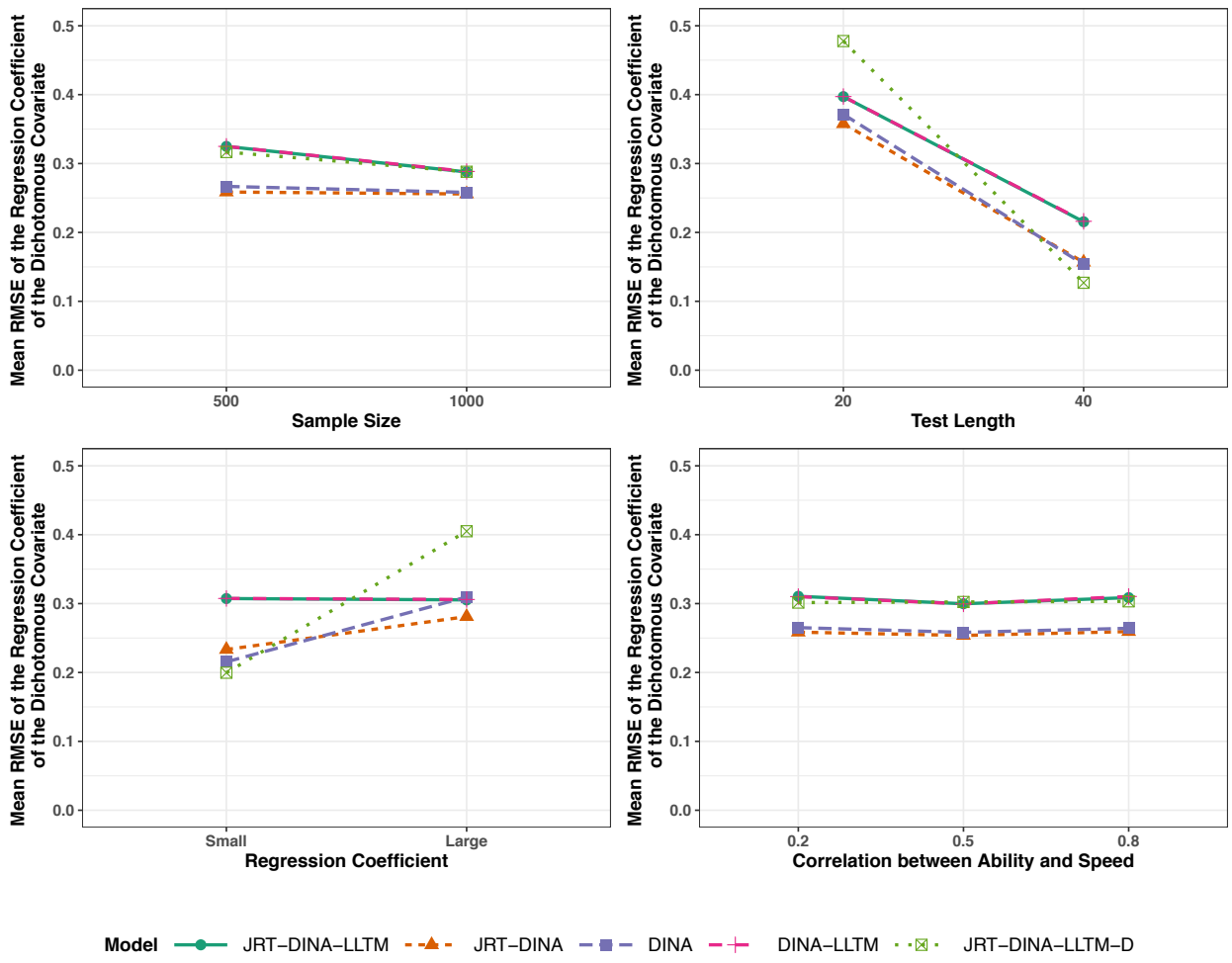


Figure 47. Mean RMSE of the regression coefficient of the dichotomous covariate of the item interaction parameter.

The mean bias, SE and RMSE of the regression intercept parameter estimates ($\mathbf{A}_{0(\delta)}$) at each level of the manipulated factors are shown in Figures 48-50. In terms of bias, model omitting the dichotomous covariate (i.e., JRT-DINA-LLTM-C) tended to overestimate $\mathbf{A}_{0(\delta)}$, while the other models tended to underestimate $\mathbf{A}_{0(\delta)}$. In addition, JRT-DINA-LLTM-C had the largest RMSE and magnitude of bias, which indicates this model had the most systematic error and total error in the estimation of $\mathbf{A}_{0(\delta)}$. Lastly, similar to the regression coefficients of item interaction parameter, models using two-step estimation had the smallest total error.

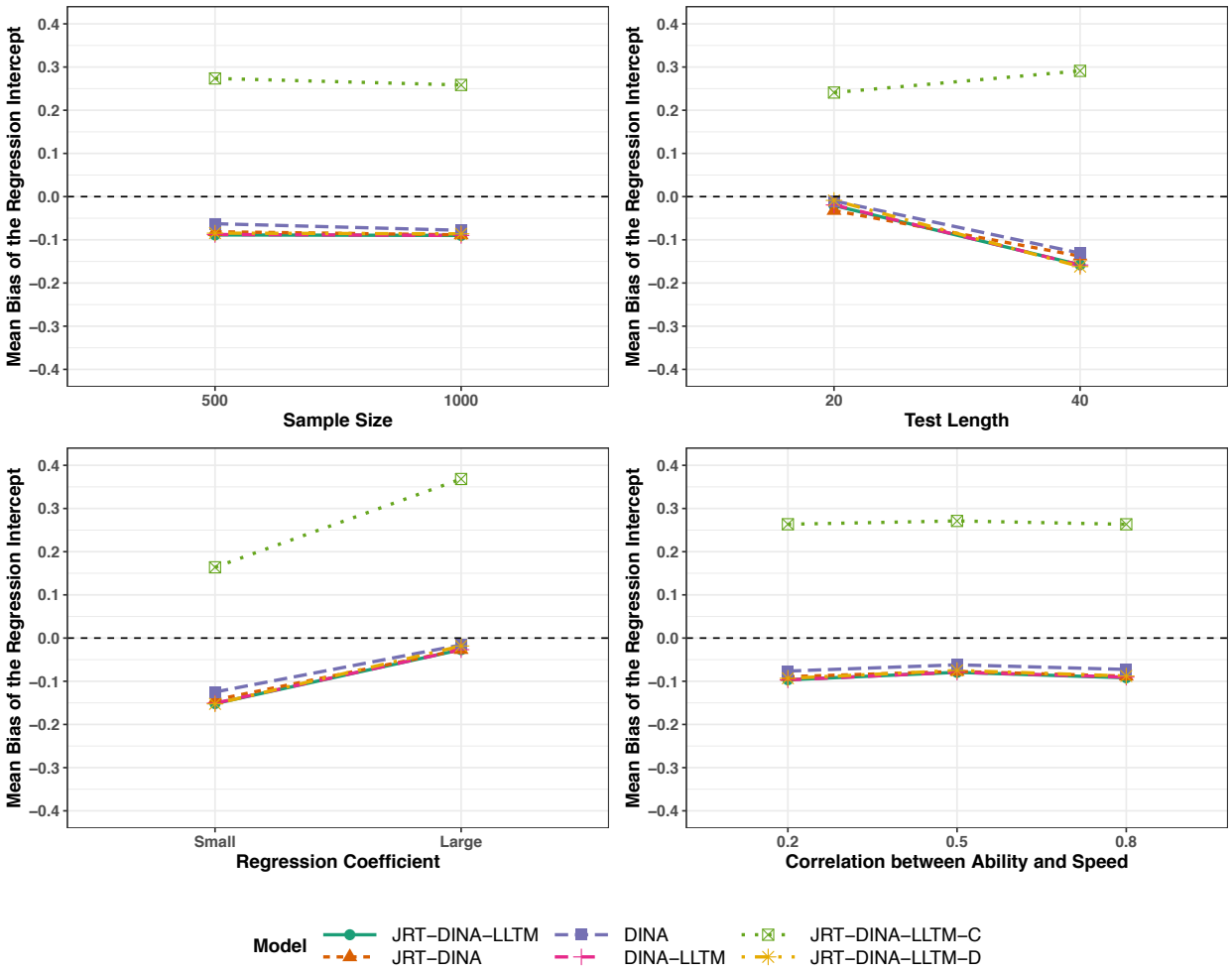


Figure 48. Mean Bias of the regression intercept of the item interaction parameter.

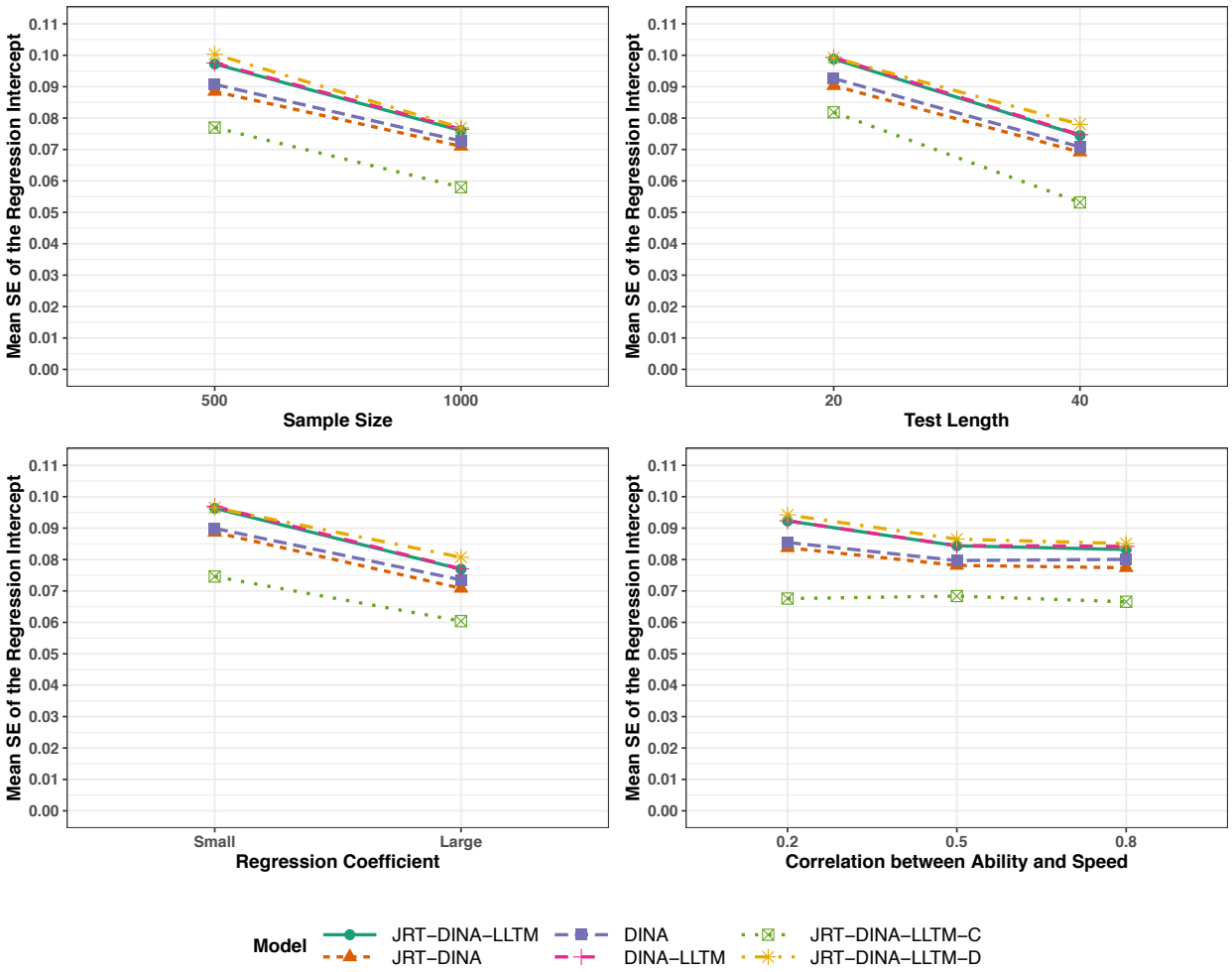


Figure 49. Mean SE of the regression intercept of the item interaction parameter.

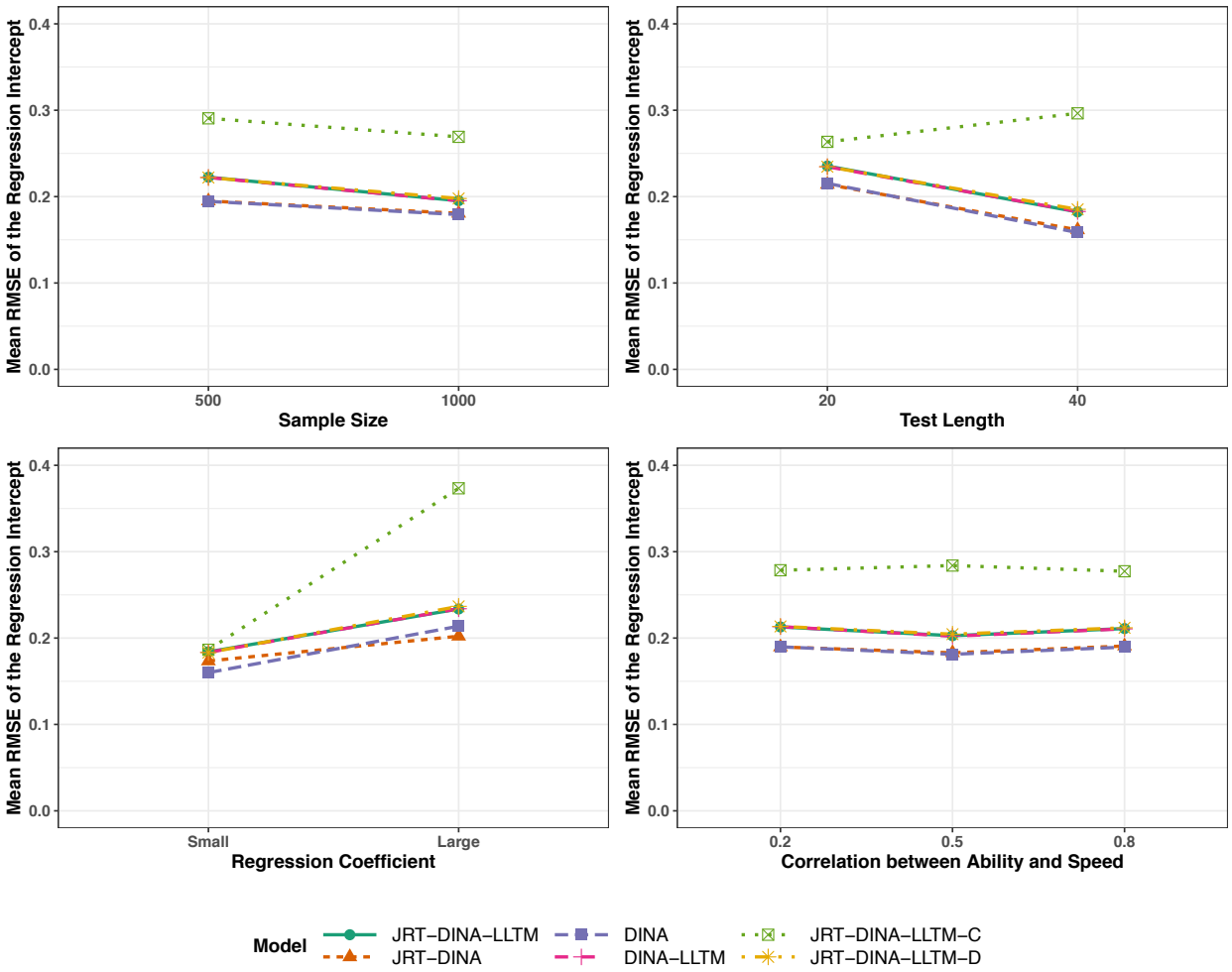


Figure 50. Mean RMSE of the regression intercept of the item interaction parameter.

4.3.3.3 Regression Coefficients for the Item Intensity Parameter

Item intensity parameter is the item parameter in the response time model and indicates the labor required by the item. The regression parameters for the item intensity parameter include the regression coefficient for continuous covariate $A_{c(\zeta)}$, regression coefficient for dichotomous covariate $A_{d(\zeta)}$, and regression intercept $A_{0(\zeta)}$. The mean bias, SE and RMSE of $A_{c(\zeta)}$ are presented in Figures 51-53. As shown in Figure 51, all data fitting models tend to underestimate $A_{c(\zeta)}$, especially the JRT-DINA-LLTM-C. In addition, JRT-DINA-LLTM-C also had the largest

RMSE, while the mean SE was similar among all data fitting models. This indicates that omitting the dichotomous covariate affects the parameter recovery of the regression coefficient in the item intensity parameter. However, the overall magnitude of bias for all data fitting models was smaller compared to that of item intercept parameter and item interaction parameter. This is expected because the item intensity parameter recovered very well with negligible biases.

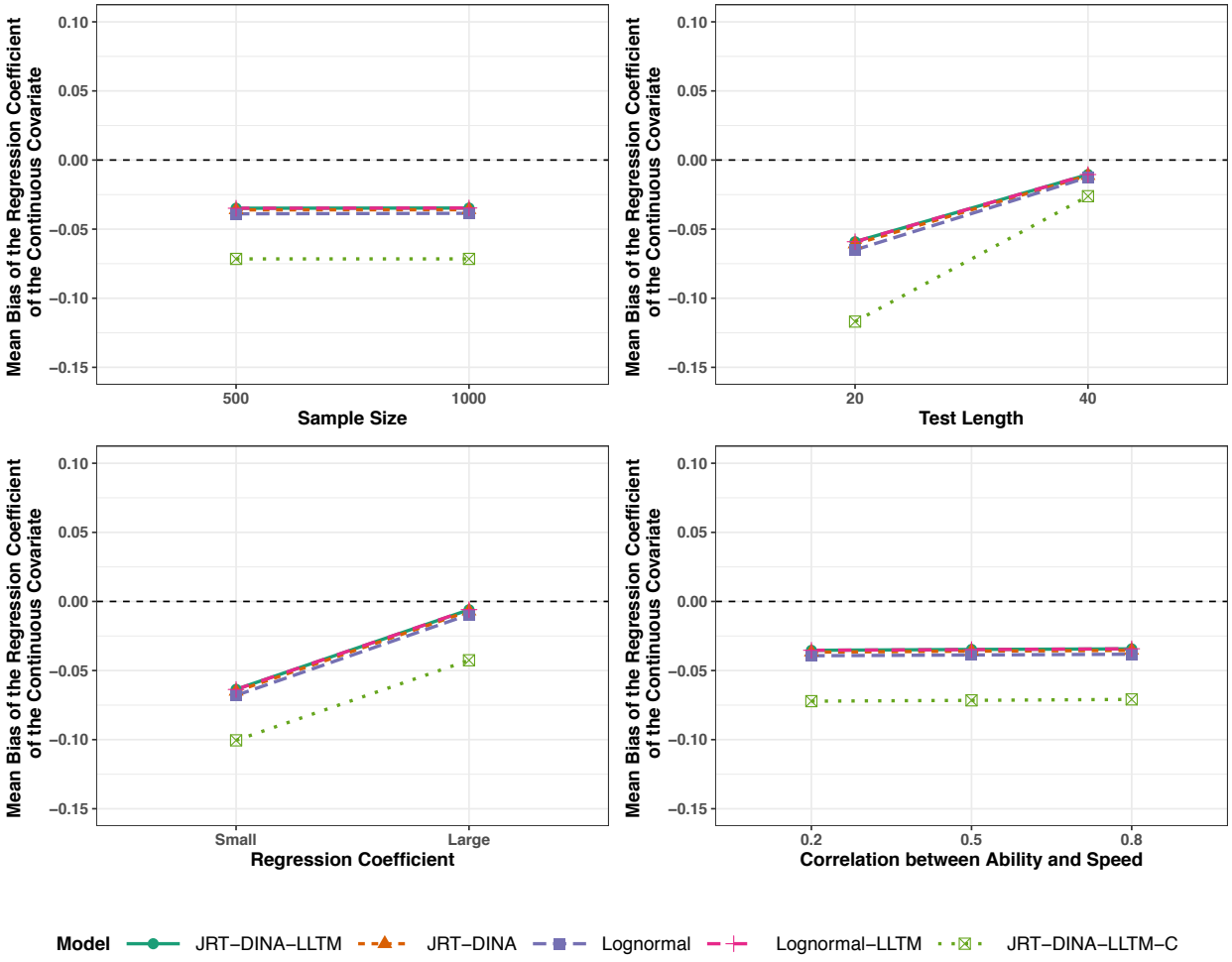


Figure 51. Mean bias of the regression coefficient of the continuous covariate of the item intensity parameter.

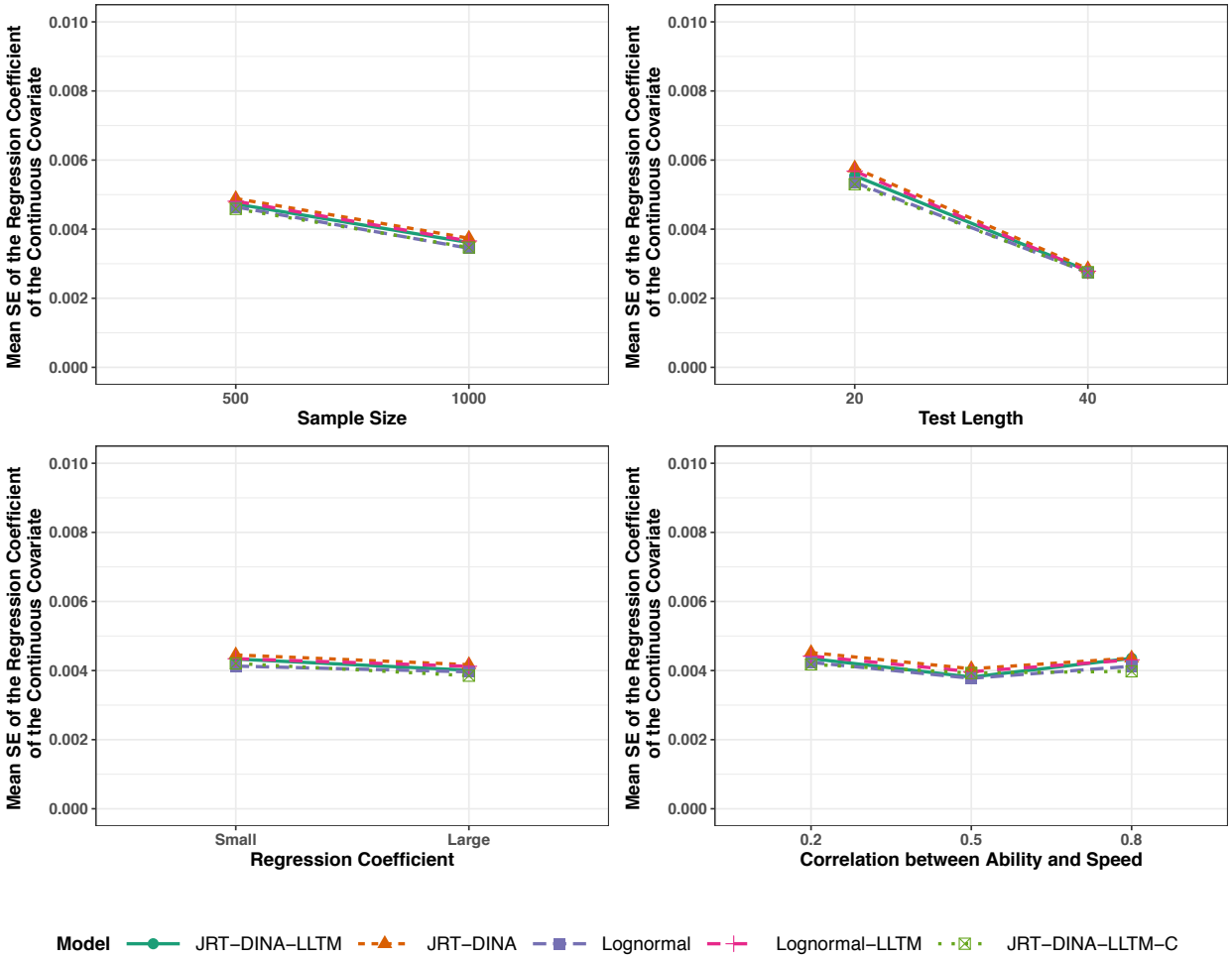


Figure 52. Mean SE of the regression coefficient of the continuous covariate of the item intensity parameter.

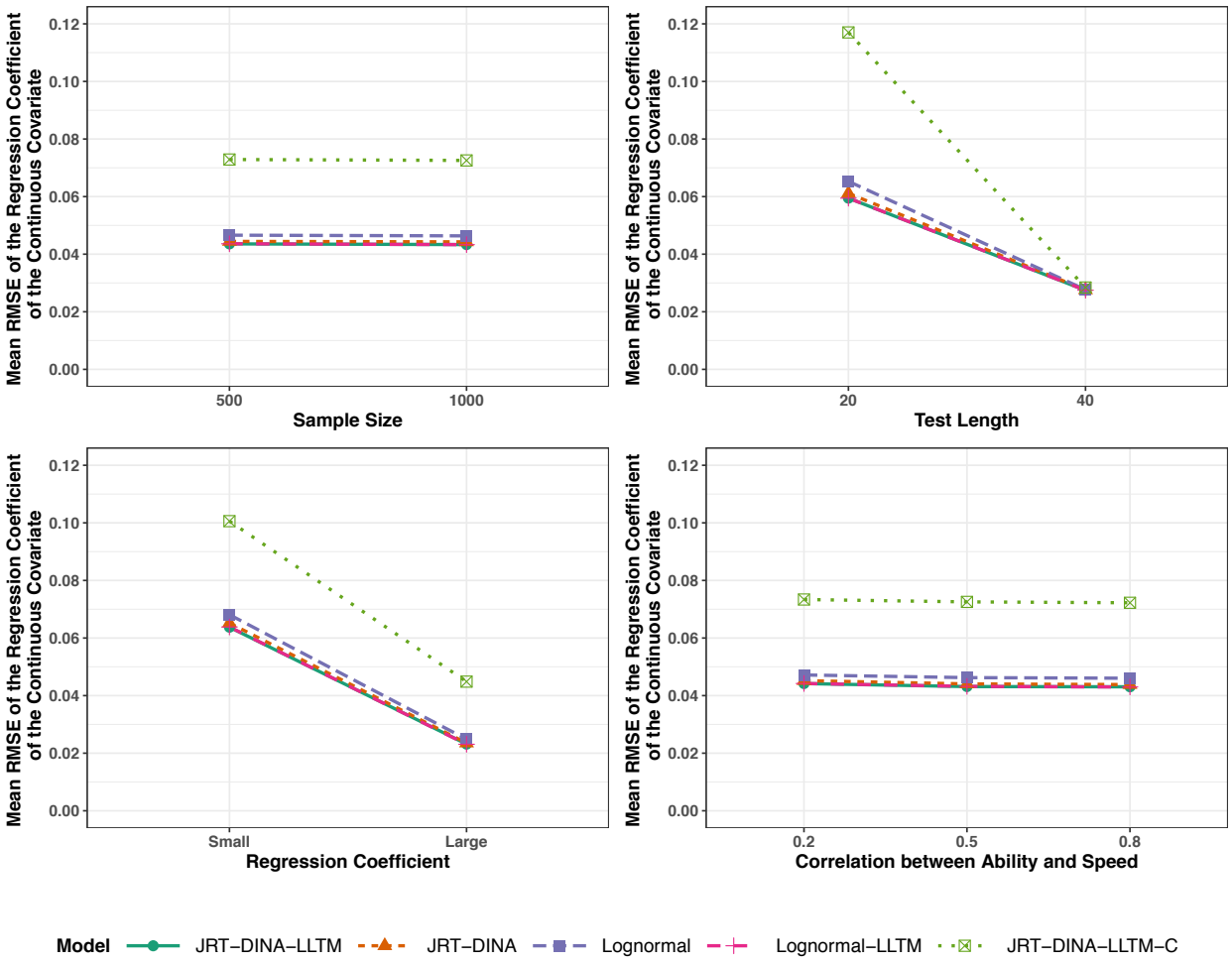


Figure 53. Mean RMSE of the regression coefficient of the continuous covariate of the item intensity parameter.

The mean bias, SE and RMSE of the regression coefficient estimates for the dichotomous covariate ($A_{d(\zeta)}$) at each level of the manipulated factor are shown in Figures 54-56. Similar to the biases of $A_{c(\zeta)}$, the mean biases of $A_{d(\zeta)}$ was also trivial. In addition, model omitting one covariate (i.e., JRT-DINA-LLTM-D) showed more negative bias than the other models, similar to that of $A_{c(\zeta)}$. Lastly, JRT-DINA-LLTM-D also had the largest RMSE, while the other models had very similar RMSE. All data fitting models had similar SE. Therefore, only covariate

misspecification would affect the recovery of the regression coefficients of the item intensity parameter.

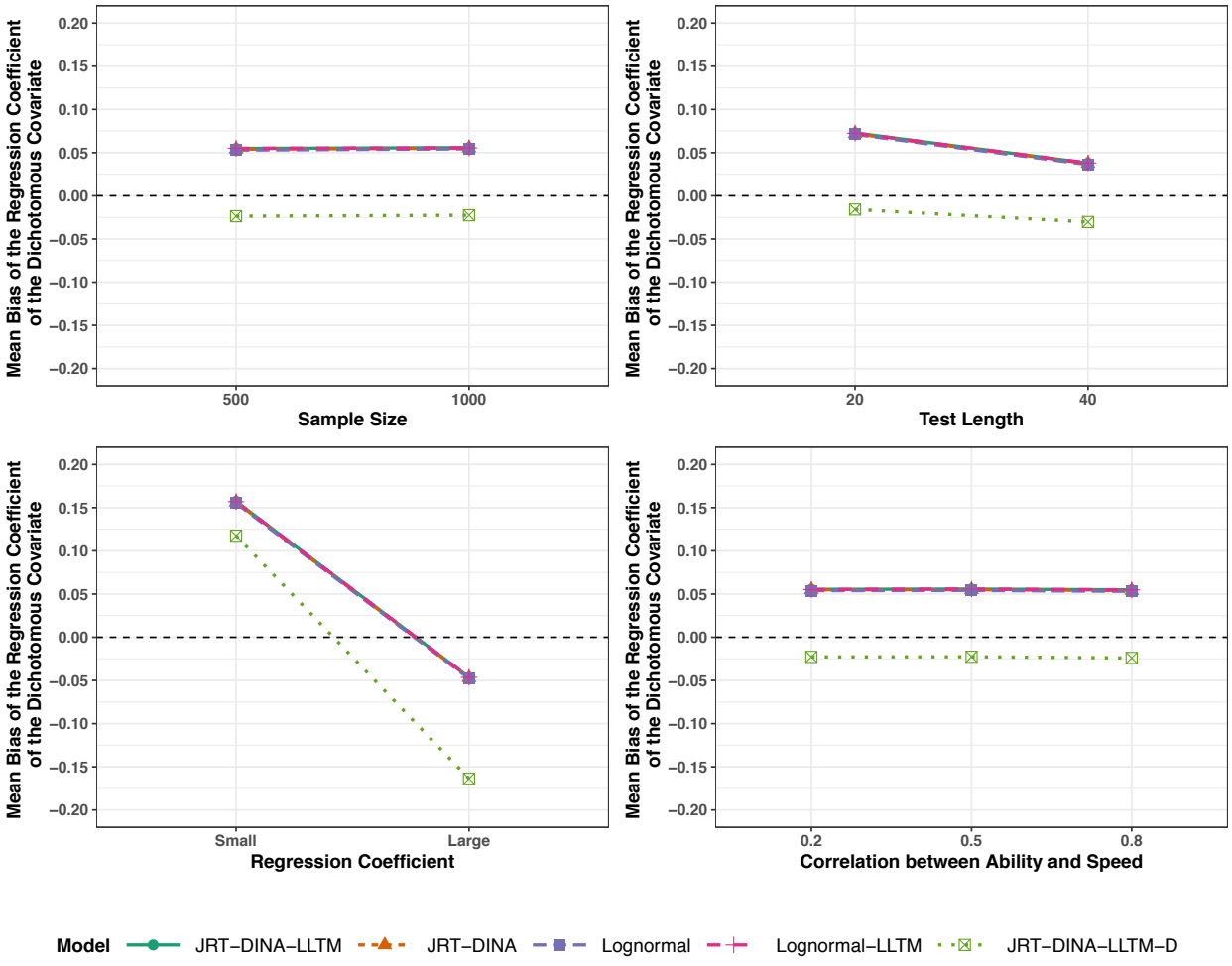


Figure 54. Mean bias of the regression coefficient of the dichotomous covariate of the item intensity parameter.

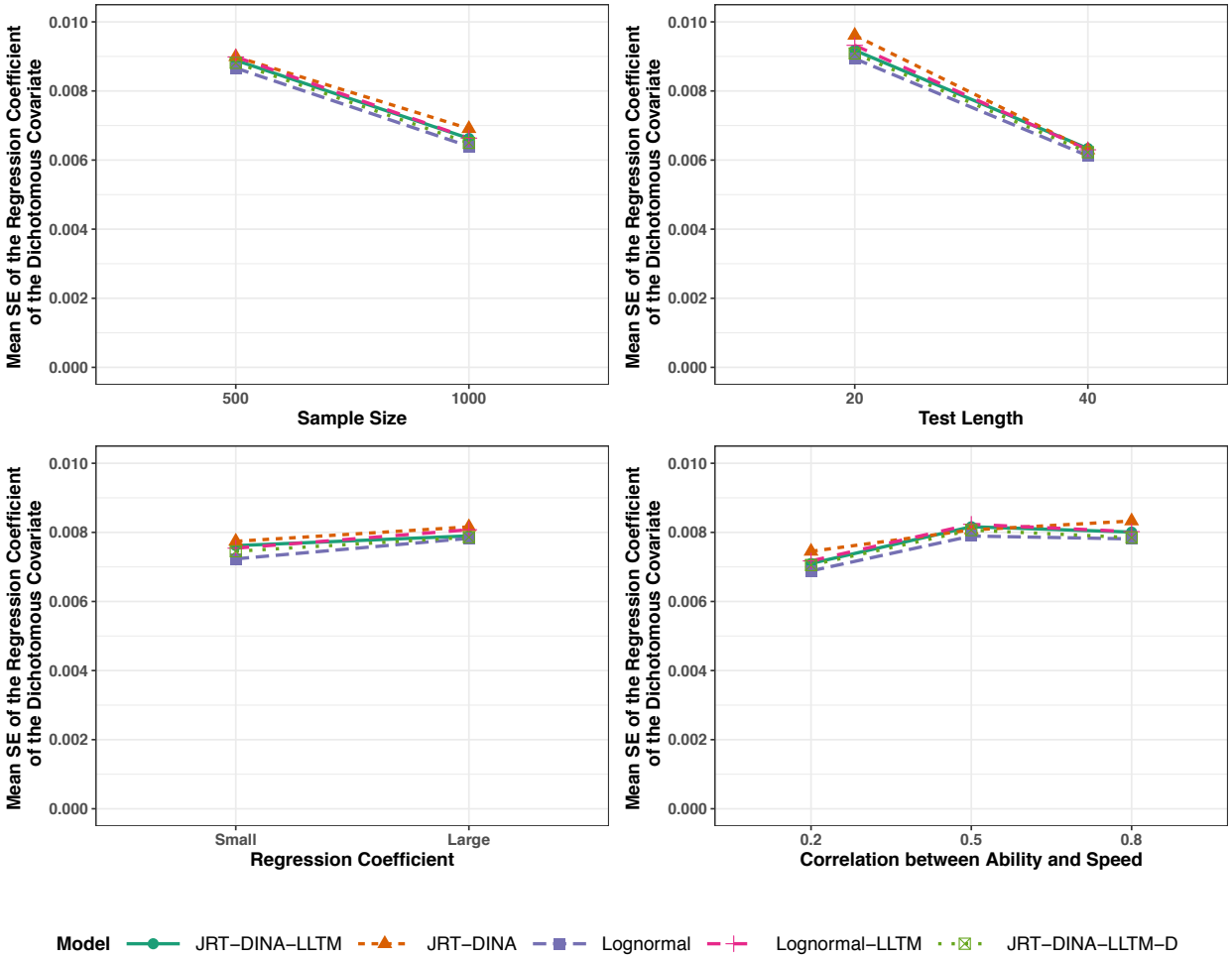


Figure 55. Mean SE of the regression coefficient of the dichotomous covariate of the item intensity parameter.

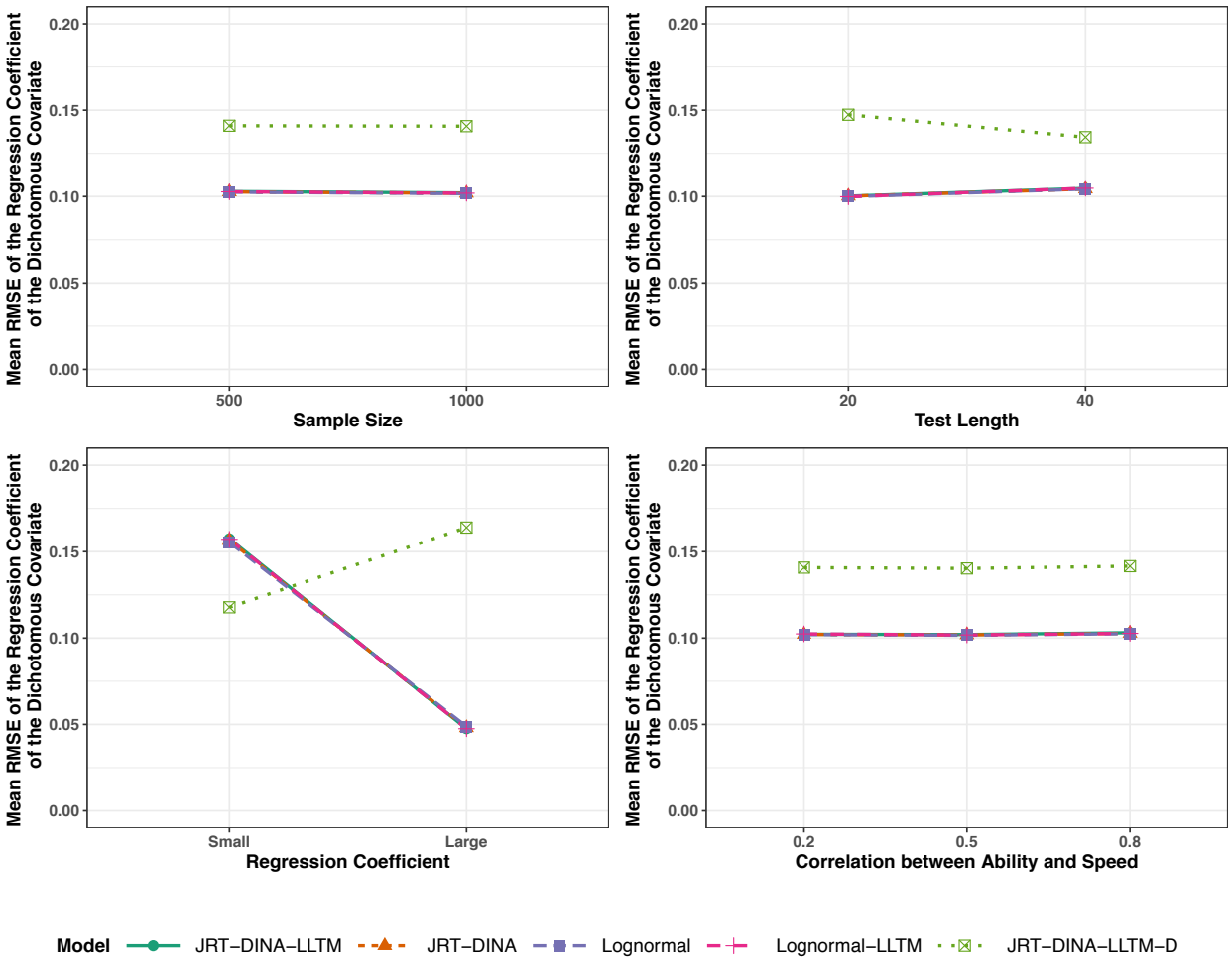


Figure 56. Mean RMSE of the regression coefficient of the dichotomous covariate of the item intensity parameter.

The mean bias, SE and RMSE of the regression intercept parameter estimates $A_{0(\zeta)}$ at different levels of manipulated factors are shown in Figures 57-59. Similar to the regression intercepts of item intercept parameter and item interaction parameter, the JRT-DINA-LLTM-C (i.e., model omitting the dichotomous covariate) had the largest bias and RMSE, while the smallest SE. This indicates that omitting the dichotomous covariate affected the regression intercept the most. This is as expected because the dichotmous covariate was dummy coded and the its existence would affect the value of the regression intercept.

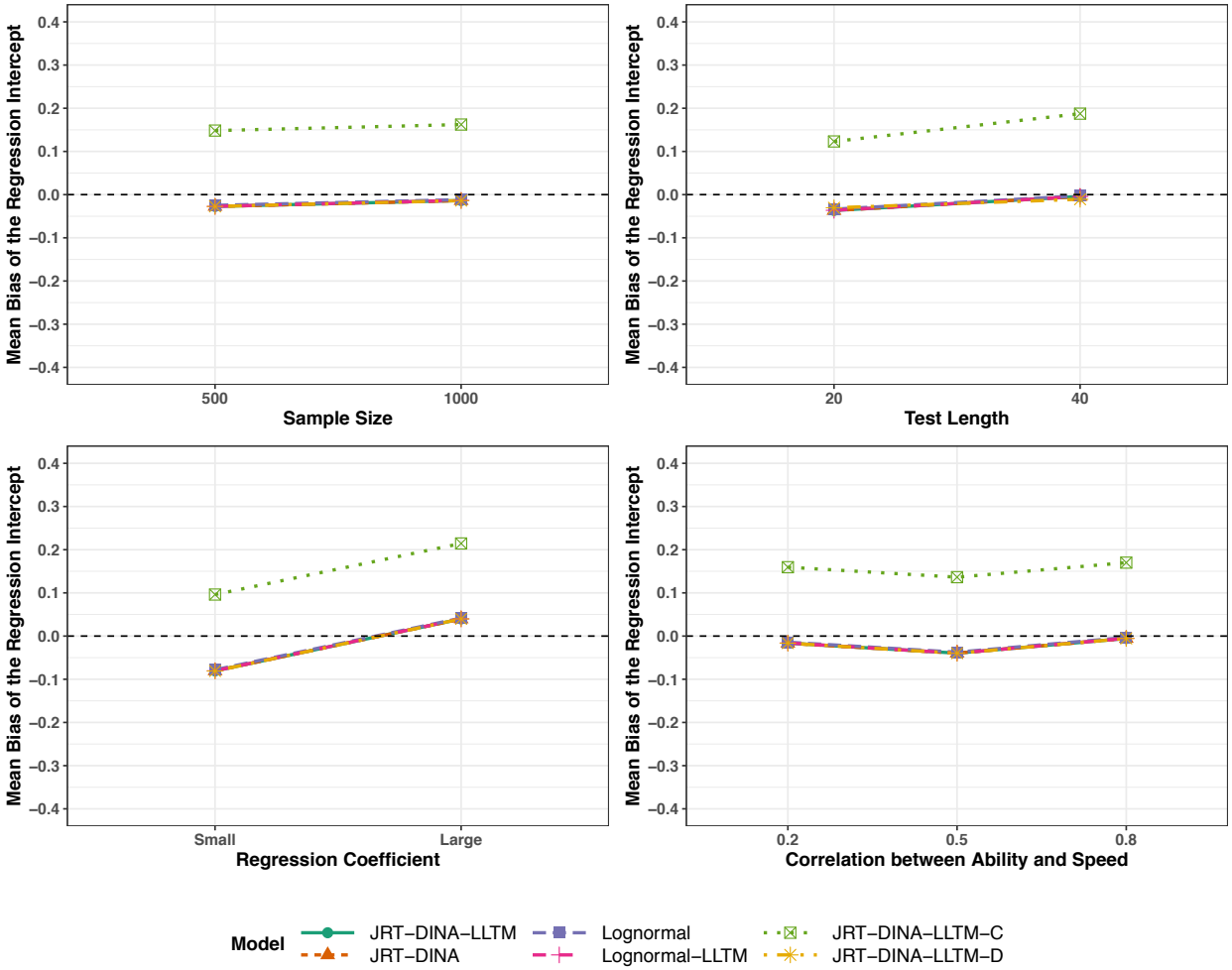


Figure 57. Mean bias of the regression intercept of the item intensity parameter.

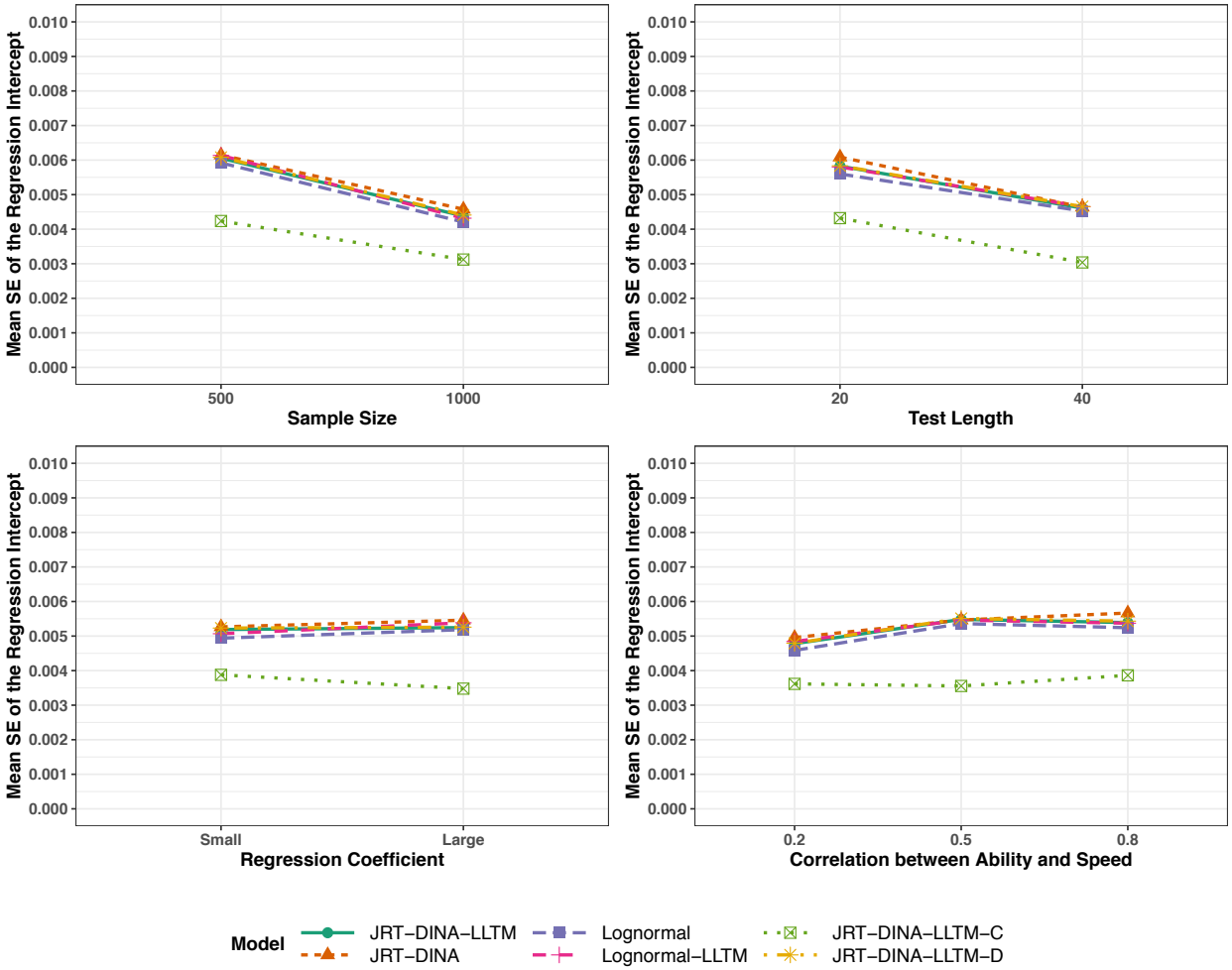


Figure 58. Mean SE of the regression intercept of the item intensity parameter.

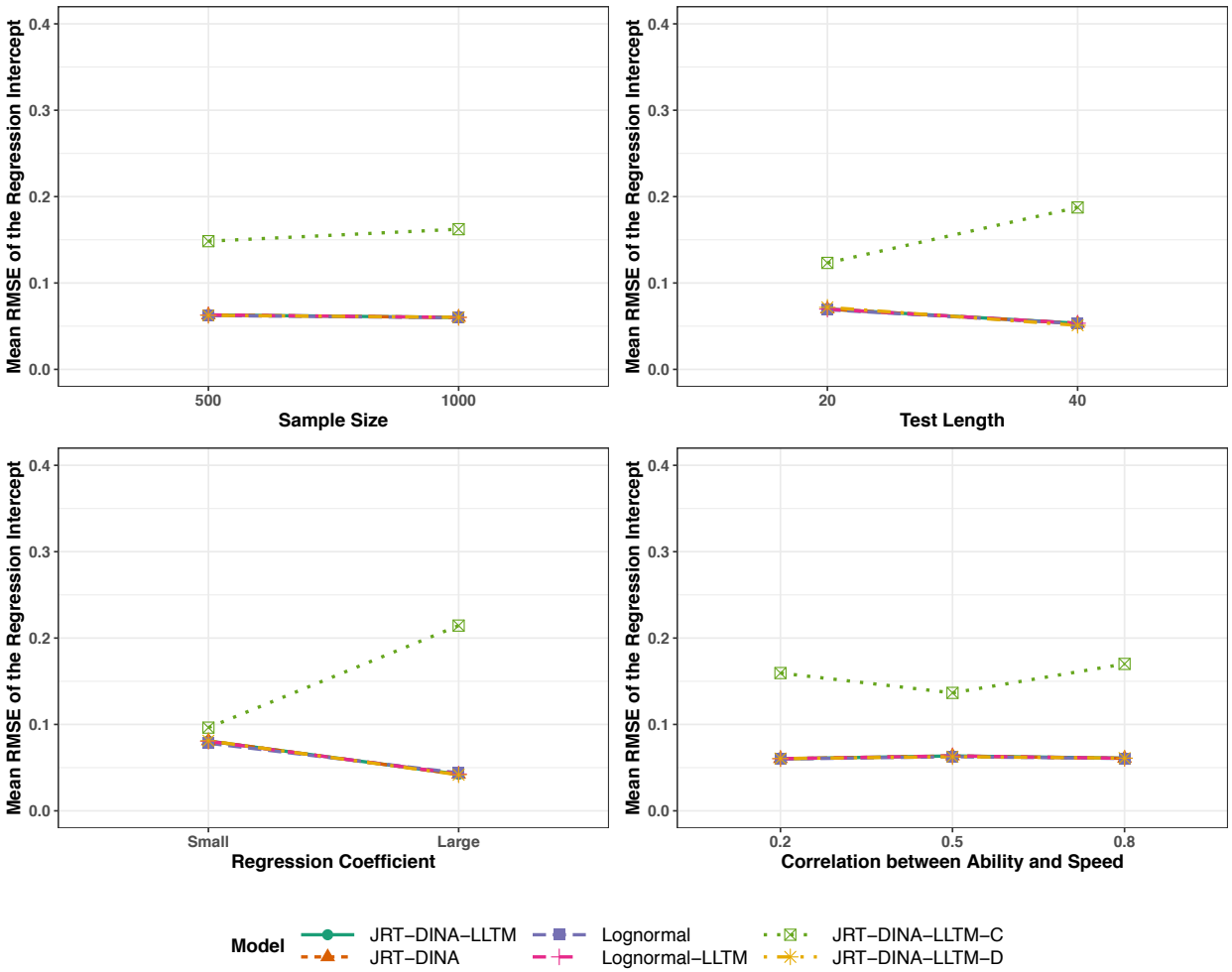


Figure 59. Mean RMSE of the regression intercept of the item intensity parameter.

4.3.3.4 Regression Coefficient Standard Error Biases

Standard error estimation plays an important role in statistical inference. Inaccurate standard error estimates can lead to erroneous conclusions. Therefore, the regression coefficient standard error biases were examined. Among the four manipulated factors, regression coefficient standard errors were affected only by test length and the explanatory power of the item covariates. Therefore, the marginal mean biases of the standard errors were plotted against these two manipulated factors for each covariate in each item parameter, as shown in Figures 60 to 65. There are several common findings for all the regression coefficients. First, all data fitting

models tended to inflate the standard error for all the regression coefficients given the positive mean biases. This indicates the possibility of higher Type II error rates in the statistical inference. That is, it is more difficult to find significant effects of the item covariates. Second, the inflation of the standard error is more severe for the dichotomous covariate than the continuous covariate. Third, the inflation of the standard error became smaller when the test length was long and the explanatory power of the item covariates was large, except for the JRT-DINA-LLTM-D (i.e., model with only dichotomous covariate). Recall that the point estimate of the regression coefficient of the dichotomous covariate in JRT-DINA-LLTM-D was also large. These indicate that it is necessary to include continuous covariates in the model. Lastly, models using two-step estimation (i.e., JRT-DINA, DINA) yielded smaller standard error biases in the item intercept and item interaction parameters than the models using one-step estimation (i.e., JRT-DINA-LLTM, DINA-LLTM). One possible reason is the different ways of setting the positive constraints for the item interaction parameter between these two categories of models. It worth further exploration of the performance of estimation methods other than JAGS.

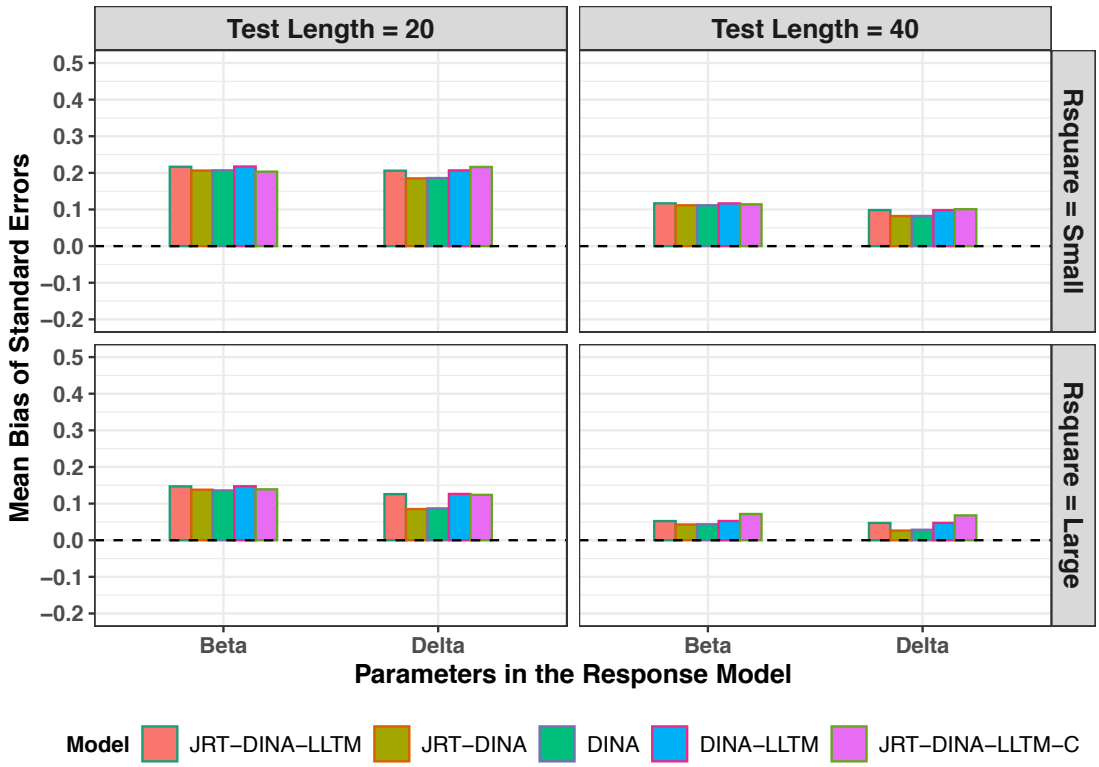


Figure 60. Mean bias of the standard errors of the continuous covariate regression coefficient.
Note. Beta = Item intercept parameter; Delta = Item interaction parameter.

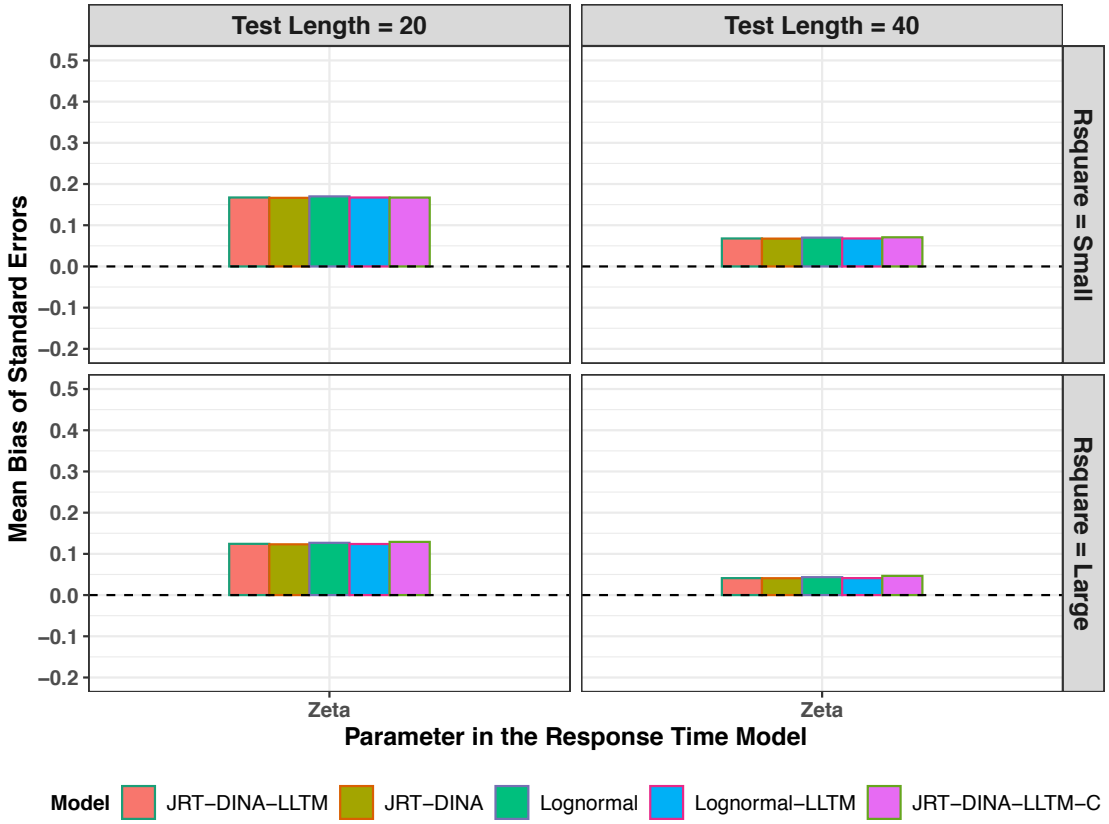


Figure 61. Mean bias of the standard errors of the continuous covariate regression coefficient.
Note. Beta = Item intercept parameter; Delta = Item interaction parameter.

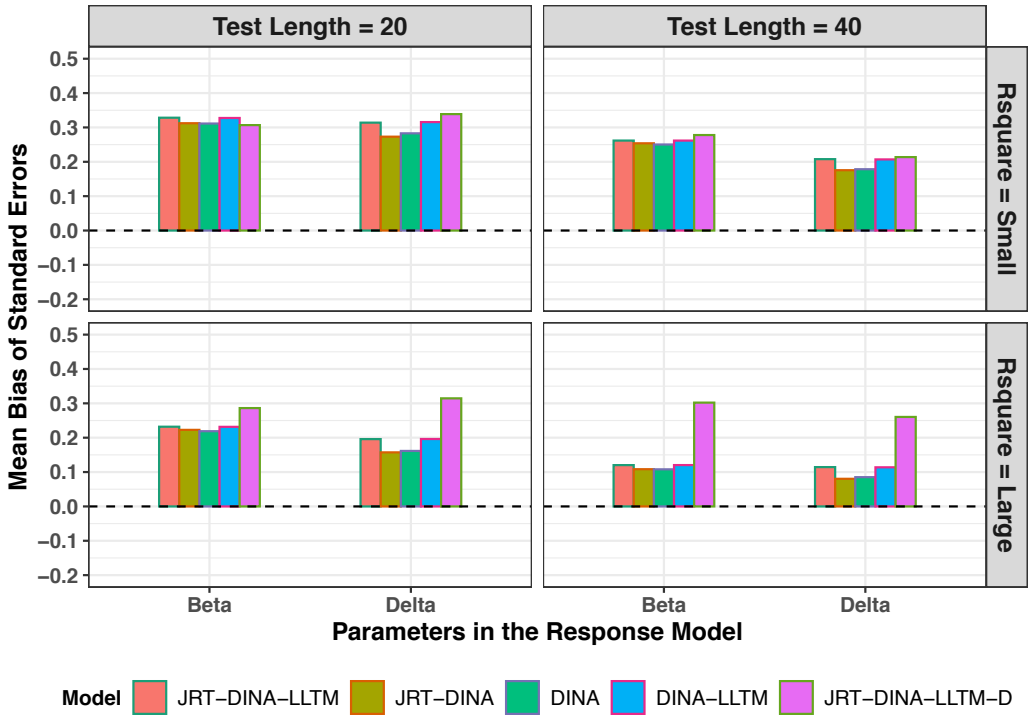


Figure 62. Mean bias of the standard errors of the dichotomous covariate regression coefficient.
Note. Beta = Item intercept parameter; Delta = Item interaction parameter.

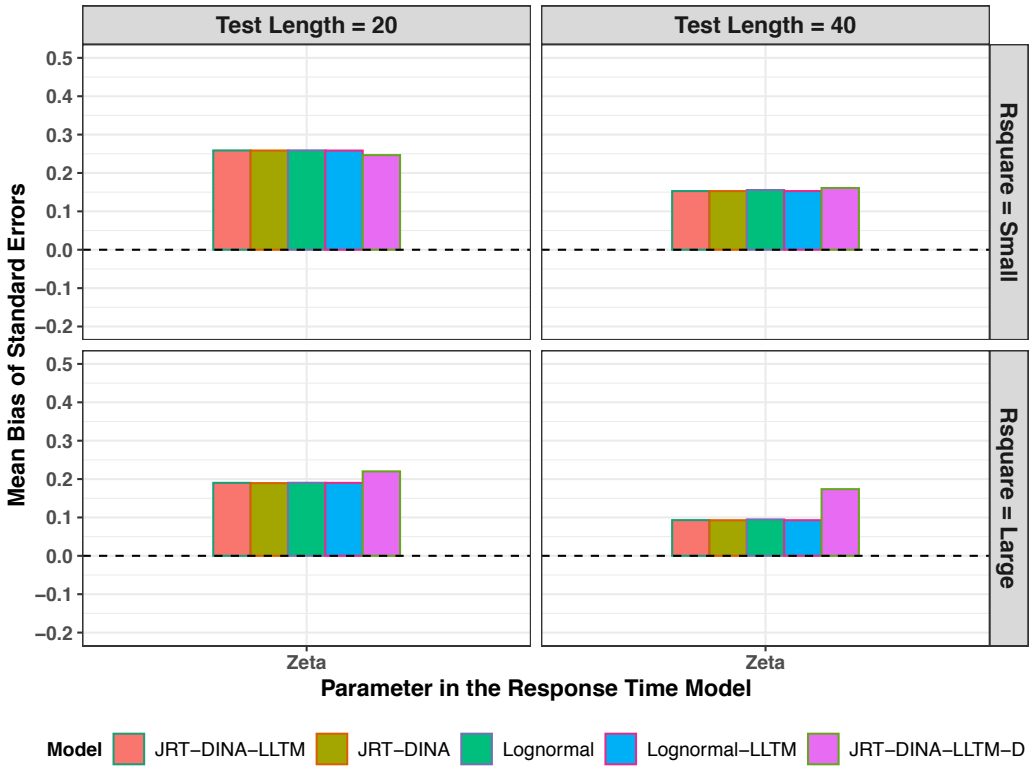


Figure 63. Mean bias of the standard errors of the dichotomous covariate regression coefficient.
Note. Zeta = Item intensity parameter.

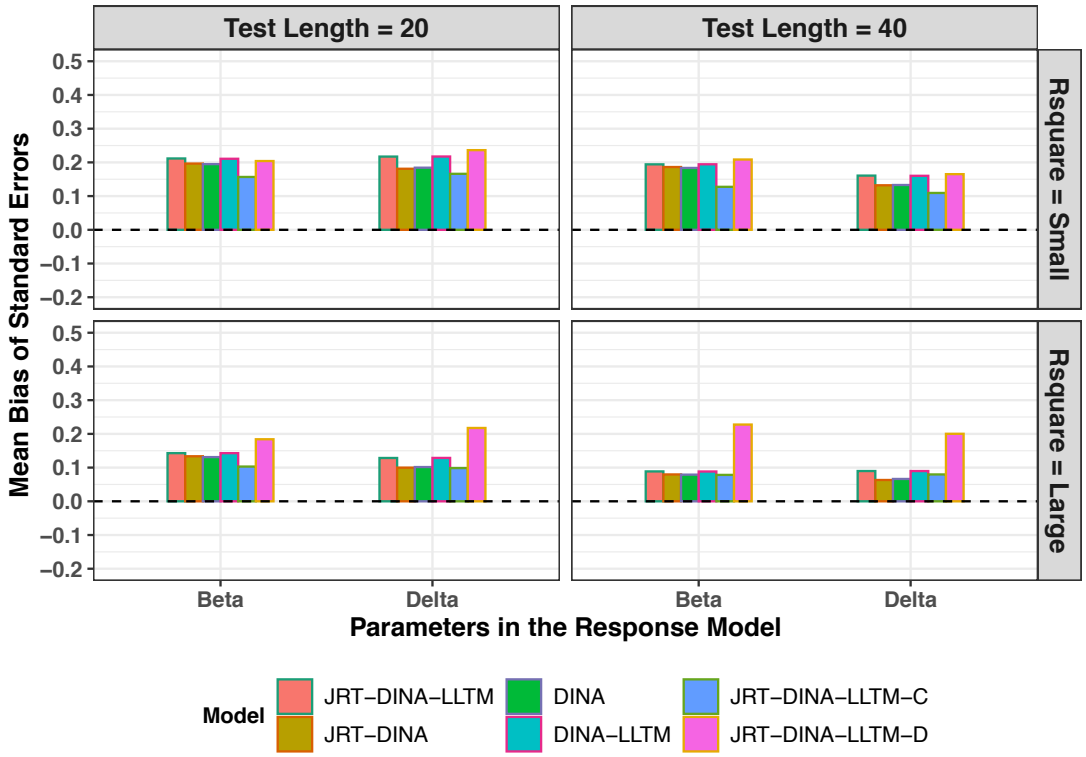


Figure 64. Mean bias of the standard errors of the regression intercept.
Note. Beta = Item intercept parameter; Delta = Item interaction parameter.

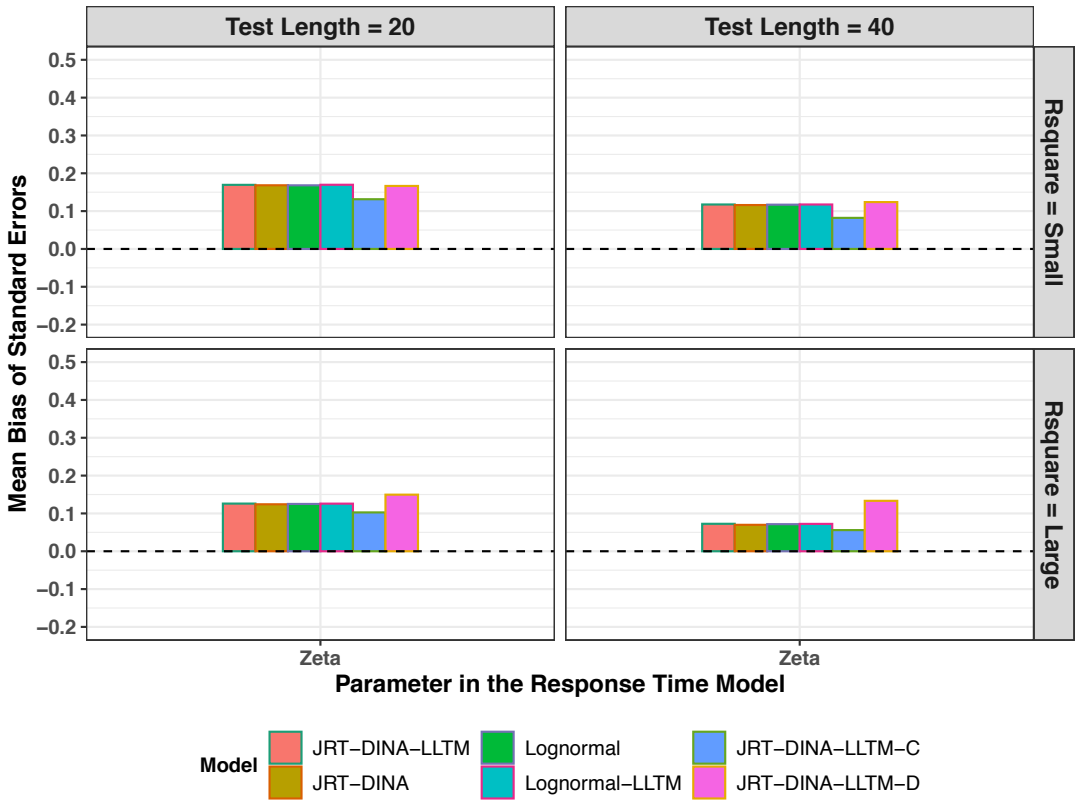


Figure 65. Mean bias of the standard errors of the regression intercept.
Note. Zeta = Item intensity parameter.

4.4 Recovery of the Higher-Order Structural Parameters

The higher-order structural parameters include the attribute intercept parameter (λ_k) and attribute slope parameter (γ_k) (k = 1, 2, 3, 4), which indicate the relation between the higher-order ability and the probability of mastery of each latent attribute in the higher-order structure. The true values for these two parameters are λ_k = 1, 0.5, -0.5, -1 and γ_k = 1 (k = 1, 2, 3, 4). The parameter recovery of these parameters are summarized in Figures 66 to 71 based on mean Bias, SE and RMSE at each level of the manipulated factors.

4.4.1 Attribute intercept parameter

For the attribute intercept parameters, as shown in Figures 66 to 68, discrepancy of mean bias among the models was between the joint models and separate models. Specifically, DINA

and DINA-LLTM had larger mean bias than the four joint models in the attribute intercept parameters except for the intercept parameter of Attribute 4 when sample size was large, test length was long or the correlation between ability and speed was small. This is expected because omitting the second-level structure affected the parameter recovery of person ability parameters and lowered the attribute correct classification rates. It is reasonable that the higher-order structural parameters associated with person parameters had larger systematic errors in DINA and DINA-LLTM. Mean SE, on the other hand, showed more consistent pattern. DINA and DINA-LLTM had slightly larger SE, indicating more random errors, than the four joint models in the estimation of attribute intercept parameters. Mean RMSE, however, did not show much differences among all models, indicating the total error in the estimation of attribute intercept parameters were comparable.

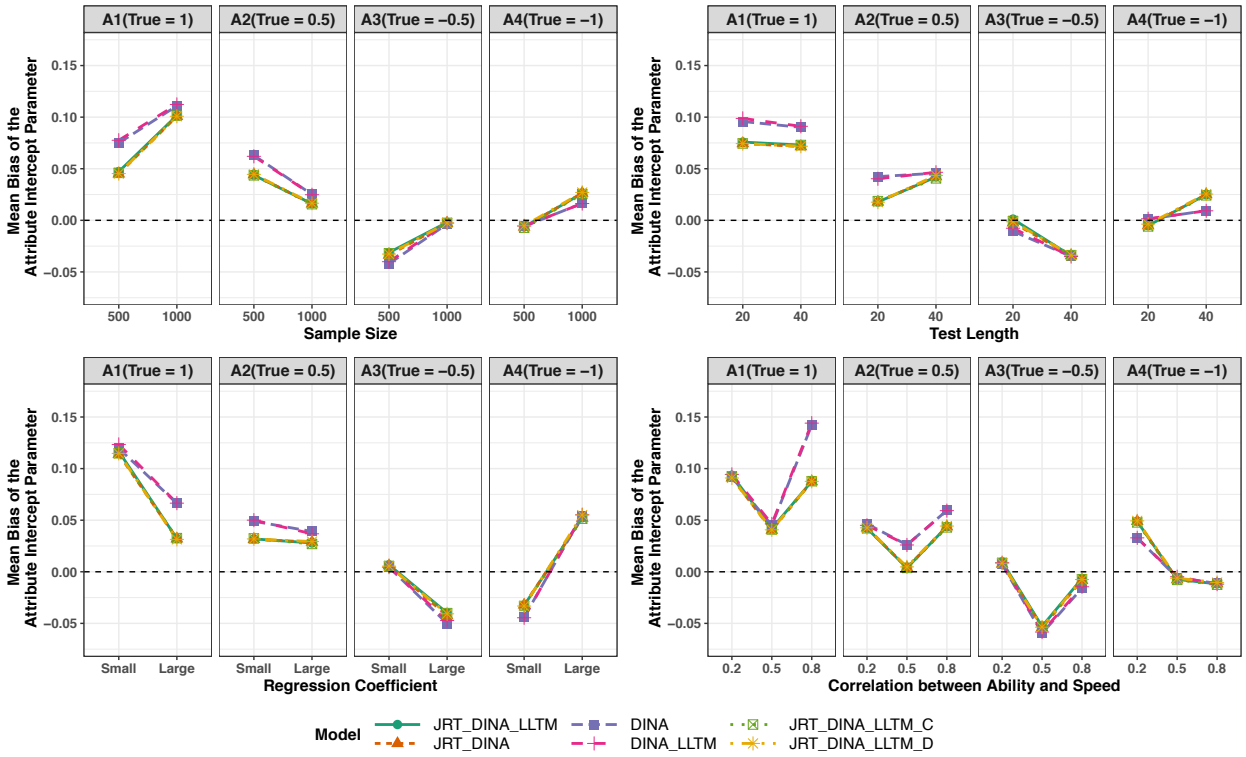


Figure 66. Marginal mean bias of the attribute intercept parameter estimates each level of the manipulated factors.

Note. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute intercept parameters corresponding to different attributes.

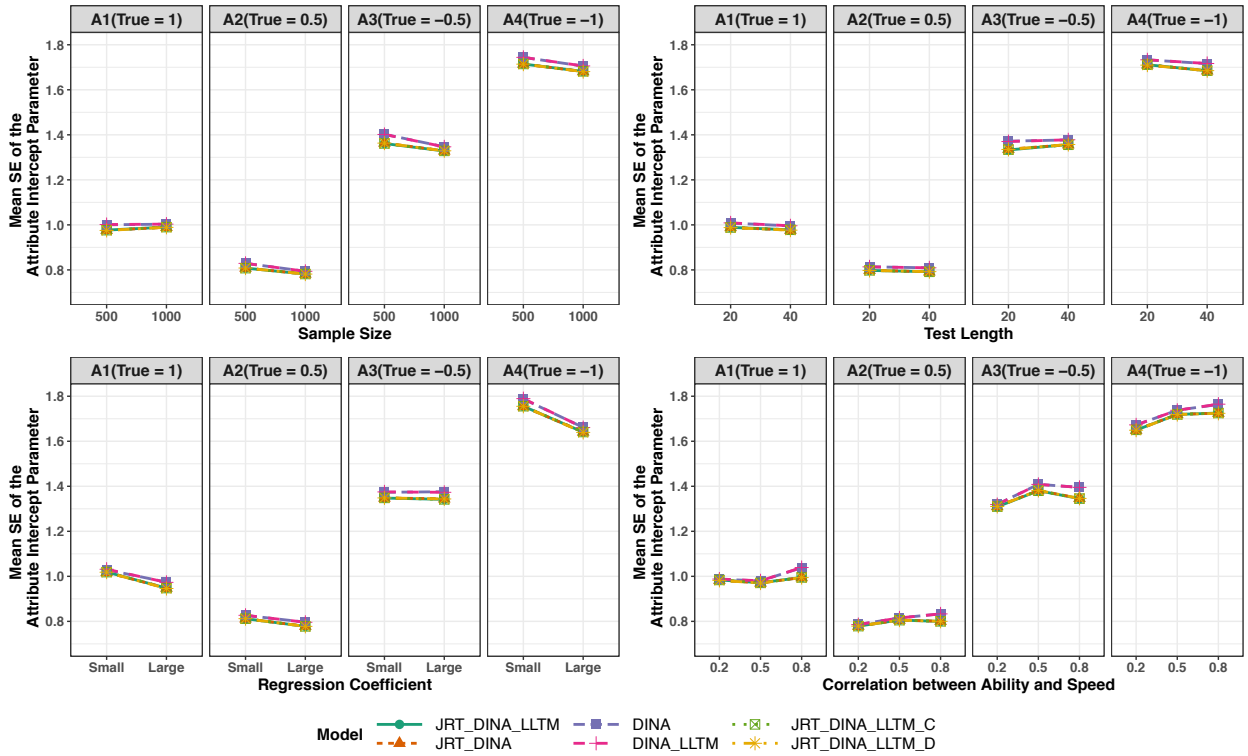


Figure 67. Marginal mean SE of the attribute intercept parameter estimates each level of the manipulated factors.

Note. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute intercept parameters corresponding to different attributes.

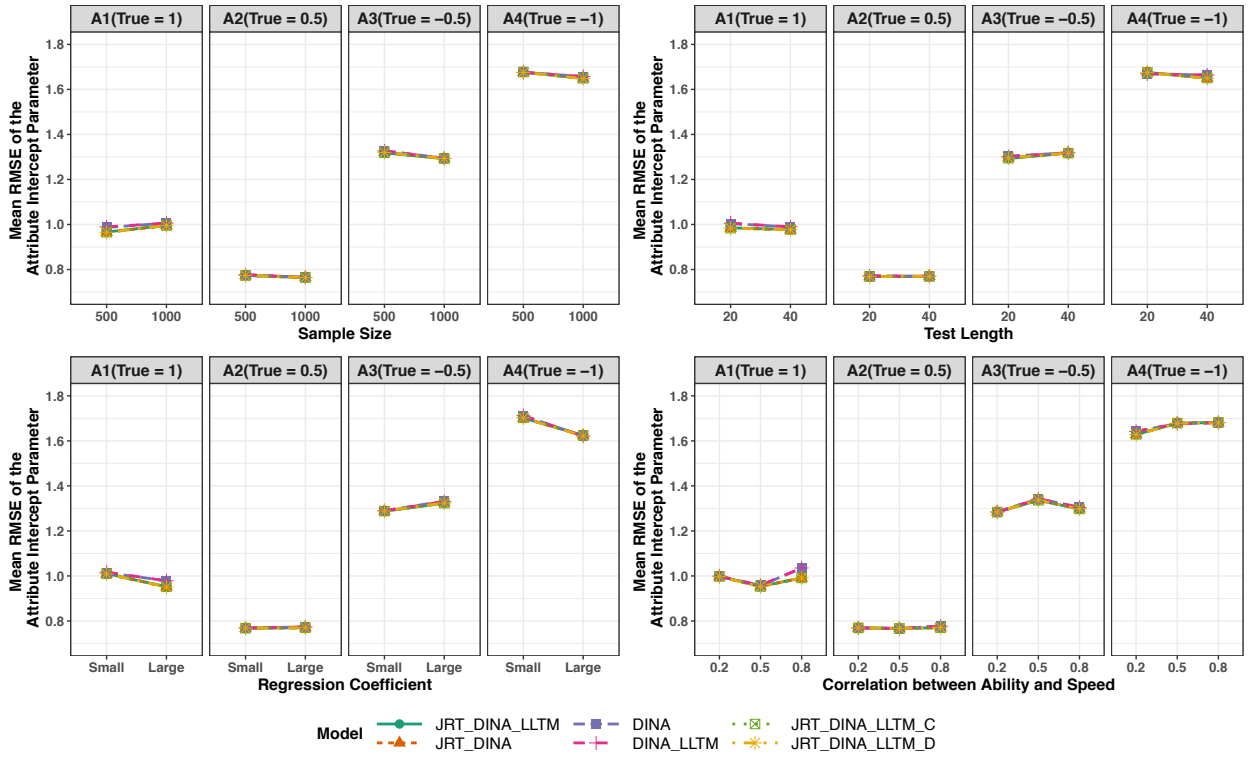


Figure 68. Marginal mean RMSE of the attribute intercept parameter estimates each level of the manipulated factors.
Note. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute intercept parameters corresponding to different attributes.

4.4.2 Attribute slope parameter

Similar to that of the attribute intercept parameter, the main difference of the mean bias for the attribute slope parameter lies between the separate models (i.e., DINA and DINA-LLTM) and joint models (i.e., JRT-DINA-LLTM, JRT-DINA, JRT-DINA-LLTM-C, JRT-DINA-LLTM-D), as shown in Figure 69. For attribute slope parameters of Attribute 1 and Attribute 2, DINA and DINA-LLTM had larger or equal mean biases than the joint models in all levels of the manipulated factors. The mean biases for attribute slope parameters of Attribute 3 and Attribute 4, however, were less consistent across different levels of the manipulated factors. Yet, mean SE and mean RMSE of the parameters shown in Figures 70-71 indicate that the random error and total error in the estimation of attribute slope parameters were larger when the data-fitting models omitted the second-level structure.

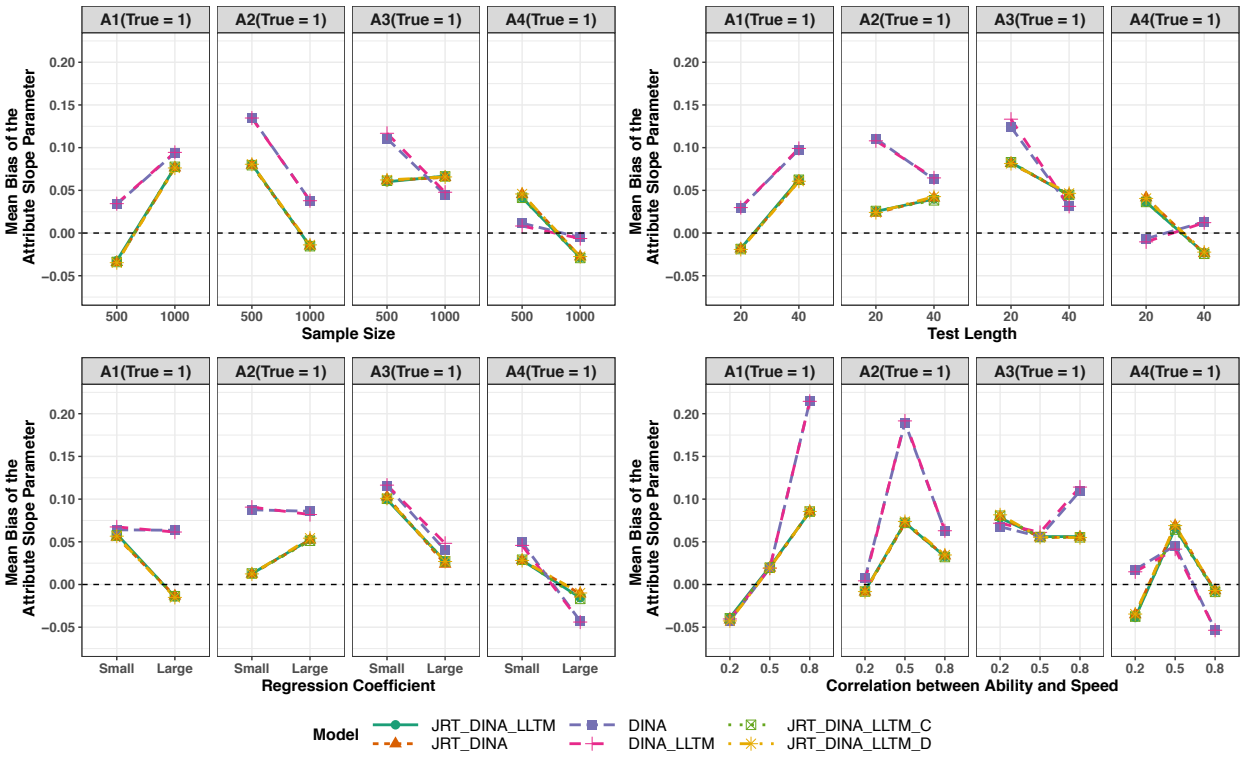


Figure 69. Marginal mean bias of the attribute slope parameter estimates each level of the manipulated factors.

Note. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute slope parameters corresponding to different attributes.

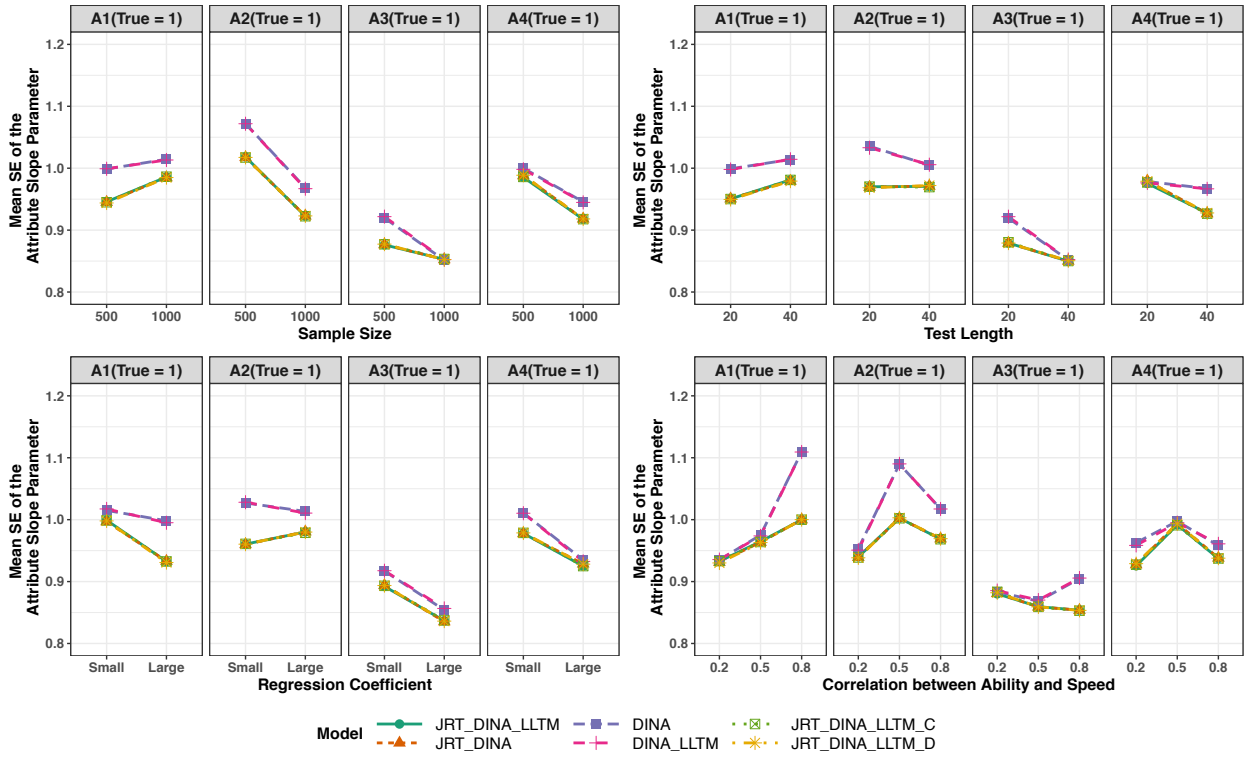


Figure 70. Marginal mean SE of the attribute slope parameter estimates each level of the manipulated factors.

Note. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute slope parameters corresponding to different attributes.

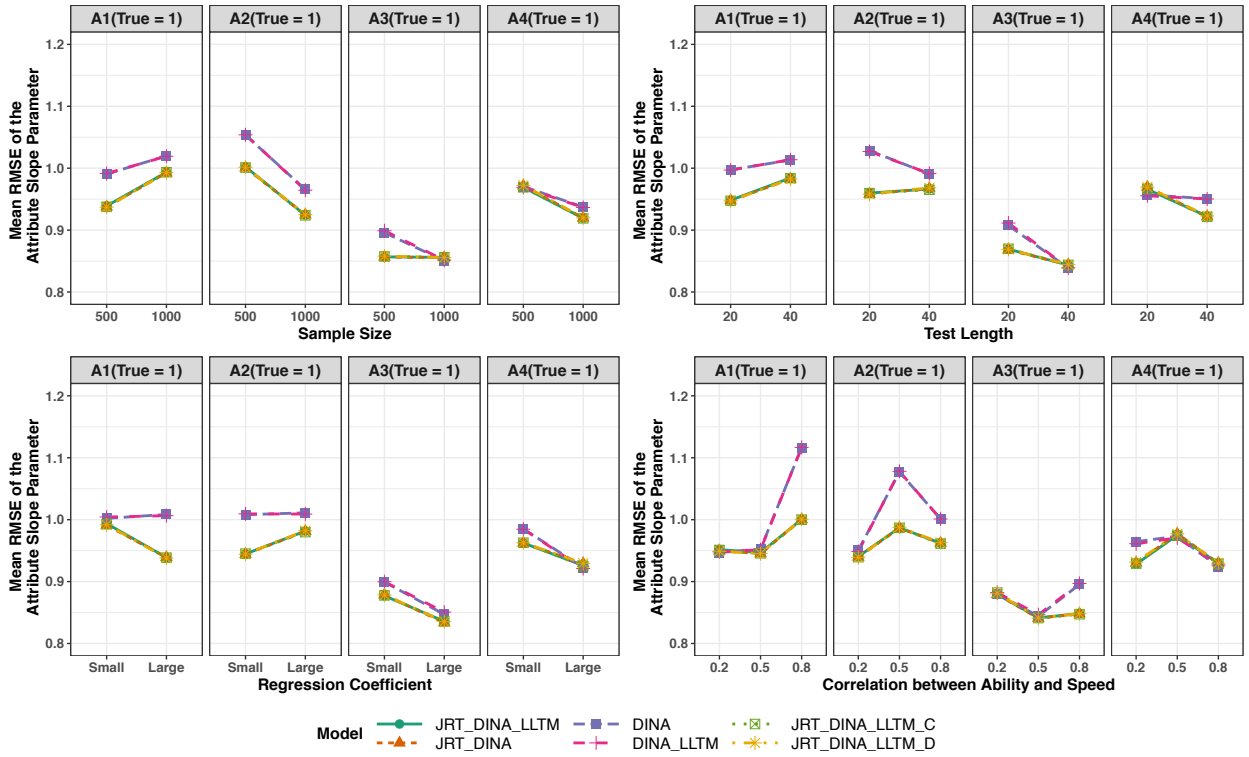


Figure 71. Marginal mean RMSE of the attribute slope parameter estimates each level of the manipulated factors.

Note. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute slope parameters corresponding to different attributes.

Chapter 5: Empirical Data Analysis Results

As an empirical demonstration, the explanatory CDMs incorporating RTs were applied on a dataset obtained from the PISA 2012 problem-solving items (OECD, 2014). The data-fitting models and the number of iterations, the number of burn-in and resulting posterior sample size are presented in Table 39. For all models, the number of chains and thinning were both set as two.

Table 39. *MCMC Iterations in Empirical Data Analysis*

Model	Number of Iterations	Number of Burn-in	Posterior Sample Size
JRT-DINA-LLTM	60000	30000	30000
JRT-DINA	60000	30000	30000
DINA	60000	30000	30000
DINA-LLTM	60000	30000	30000
Lognormal	20000	10000	10000
Lognormal-LLTM	20000	10000	10000
Two-step Regression	10000	5000	5000

The dataset included 2173 respondents who completed 7 items measuring 2 latent attributes. The Q matrix is presented in Table 8. Detailed information on the dataset and analysis procedure is presented in Chapter 3. The description of the two covariates used in the data analysis is provided in Table 40. In the following paragraphs, the model fit, item parameter estimates and attribute estimation are compared among different data-fitting models. In addition, the estimation of regression coefficients of the item covariates is presented and interpretation is provided for each data-fitting model. In general, the empirical data analysis aims to answer two research questions:

1) According to the empirical data analysis, do the item covariates have significant effects on the item parameters from CDM and the RT model, respectively?

2) How is the consistency of the item parameter estimates and attribute estimates among the data-fitting models?

Table 40. *Descriptive Statistics of Item Covariates*

Covariate	Variable Type	Mean	SD	Min	Max
Average steps used to complete the item	Continuous	12.82	12.53	4.64	39.98
Familar or unfamiliar	Dichotomous	0.29	0.49	0	1

5.1 Convergence and Model fit

All models converged and mixed well with $\hat{R} < 1.05$ for all estimated parameters. To further examine the estimation precision of the parameter estimates, percentages of the focal parameters (i.e., person ability, person speed, item intercept, item interaction, item intensity, regression coefficients) with ESS larger than 400 for each model are reported in Table 41. Parameter estimates with ESS < 400 included item intercept and item interaction parameter estimates (no more than 3 for each model). Peripheral parameters with ESS < 400 were the higher-order structural parameters, which are known to be difficult to obtain accurate estimation (e.g., de la Torre & Douglas, 2004). Given that the statistical inference was mainly drawn on the regression parameters (i.e., regression intercepts and coefficients), the ESS of these parameters was further examined. As a result, all regression parameters were estimated with sufficient precision (ESS > 400) as shown in Table 41.

Table 41. *Proportions of Parameters with Effective Sample Sizes > 400.*

Model	Proportion among all parameters	Proportion among regression parameters
JRT-DINA-LLTM	1.000	1.000
JRT-DINA	0.999	1.000
DINA	0.997	1.000
DINA-LLTM	0.999	1.000
Lognormal	1.000	1.000
Lognormal-LLTM	1.000	1.000

The model fit of the data fitting models was evaluated in terms of both absolute fit and relative fit as shown in Table 42. Specifically, the PPP-values for the response models and response time models were close to 0.5 for all models. This indicates that all models had satisfactory absolute model fit on the empirical dataset. Further, although the simulation study suggests that DIC may penalize the RT-DINA-LLTM and DINA-LLTM for their model complexity, both deviance and DIC favored models with covariates over their counterpart without covariates. A possible reason is that the sample size in the empirical dataset was large enough such that the model complexity of JRT-DINA-LLTM and DINA-LLTM was not significantly penalized by DIC.

Table 42. *Data Fitting Models and Model Fit Results in Empirical Data Analysis*

Model	PPP of response model	PPP of response time model	Deviance	DIC
JRT-DINA-LLTM	0.466	0.509	**23321.140**	**39120.401**
JRT-DINA	0.470	0.511	23405.750	44051.319
DINA-LLTM	0.481	-	**9066.952**	**21944.356**
DINA	0.480	-	9080.959	30103.623
Lognormal-LLTM	-	0.511	**14204.924**	**16822.872**
Lognormal	-	0.514	14205.760	18640.803

Note. PPP = Posterior predictive p-value; DIC = deviance information criterion.

5.2 Model Parameter Estimates

The model parameter estimates of interest include the regression parameter estimates, item parameter estimates and attribute estimates obtained from the data-fitting models. In the following paragraphs, the interpretations of the regression coefficients and the consistency of the item parameter estimates and attribute estimates among the data fitting models are provided.

The regression parameter estimates, including point estimates and 95% credible intervals, are reported in Table 43. Point estimates are essentially posterior means, which are referred to as Expected A Posteriori (EAP) in Table 43. Point estimates with 95% credible intervals not covering 0 are considered to significantly deviate from 0. The statistical inferences are consistent

among the data fitting models in general. The only effect of the covariates that significantly deviated from 0 was the effect of the dichotomous covariate (i.e., *familiar* or *unfamiliar*) for the item intercept parameter. Specifically, items with contexts that were familiar to the respondents had item intercept parameter estimates that were larger than the items with contexts that were less familiar to the respondents given the same number of average steps taken. Note that item intercept parameter is the logit form of the guessing parameter in the original parameterization of the DINA model. Therefore, familiar items were more likely to be guessed correctly in the current dataset. It is worthwhile to point out that the effects that did not deviate significantly from 0 also have meaningful interpretations. For example, none of the effects for the item intensity parameter significantly deviated from 0. This indicates that *familiar* or *unfamiliar* items or items required different average number of steps to complete did not lead to significantly different amount of time to complete. In addition, the "non-significance" of the effects in the current analysis may be due to that there were only seven items in the current dataset. It is suggested to use more items in future applications to achieve enough statistical power to detect the effects of the item covariates.

The consistency of item parameter estimates among the data fitting models was very high as shown in Figures 72 to 74. For the intercept parameter, the corresponding guessing parameter estimates on the probability scale shown in Figure 72 indicates that the *familiar items* (CP007Q01, CP007Q02) had higher guessing probabilities than the other items. This has also been suggested based on the regression results shown in Table 43. Similarly, the slipping parameter estimates, shown in Figure 73, also showed that items CP007Q01 and CP007Q02 were less likely to slip than the other items. In addition, the items were considered to be of low quality given that $1\text{-}s\text{-}g < 0.65$ for all items except for CP038Q01 (de la Torre, 2007). Future

explorations may include more item covariates which may shed light on how to revise these item. For the item intensity parameter, all data fitting models yielded the same estimates for all item, as shown in Figure 74. This is consistent with the simulation finding that the item intensity parameter biases were negligible for all data fitting models.

The consistency of the attribute estimates was also high among the data fitting models as shown in Tables 44 and 45. A clear pattern is that the presence of the second-level structure between ability and speed did not affect much of the consistency. Specifically, the correlation of attribute patterns between models with the second-level structure between ability and speed (i.e., JRT-DINA-LLTM and JRT-DINA) was 0.990 and that between models without the second-level structure (i.e., DINA-LLTM and DINA) was 0.982. However, any correlations between a model with the second-level structure and a model without the second-level structure were no higher than 0.932. The same observations were also found for the consistency of single attribute estimates. Given that the absolute value of the $\hat{\rho}_{\theta\tau}$ obtained by the JRT-DINA-LLTM was about 0.19, this finding was consistent with the simulation study that models with the second-level structure had similar attribute classification accuracies than models without the second-level structure when $\rho_{\theta\tau}$ is smaller than 0.5.

Table 43. *Regression Coefficient Estimates of the Covariates*

Description	Parameter	Data Fitting Model							
		JRT-DINA-LLTM		JRT-DINA		DINA + Lognormal		DINA-LLTM + Lognormal-LLTM	
		EAP	95% CI	EAP	95% CI	EAP	95% CI	EAP	95% CI
Intercept	$A_{0(\beta)}$	**-2.24**	[-3.45, -1.03]	**-2.13**	[-3.10, -1.19]	**-1.95**	[-2.95, -.97]	**-1.98**	[-3.14, -.82]
	$A_{0(\delta)}$	**3.22**	[1.94, 4.55]	**2.98**	[1.74, 4.19]	**3.08**	[1.62, 4.54]	**3.19**	[1.92, 4.52]
	$A_{0(\zeta)}$	**4.13**	[3.04, 5.22]	**4.13**	[3.31, 4.94]	**4.14**	[3.22, 4.96]	**4.13**	[.41, 4.95]
Average steps used to complete the item	$A_{c(\beta)}$	-.64	[-1.98, .73]	-.60	[-1.82, .06]	-.59	[-1.87, .52]	-.60	[-1.97, .78]
	$A_{c(\delta)}$	.66	[-.94, 2.43]	.41	[-1.12, 1.46]	.47	[-1.30, 1.60]	.69	[-.92, 2.50]
	$A_{c(\zeta)}$	.01	[-1.33, 1.37]	.02	[-1.01, 1.27]	.02	[-1.02, 1.07]	.01	[-1.00, 1.02]
Familiar or unfamiliar	$A_{d(\beta)}$	**3.29**	[.44, 6.14]	**3.12**	[.69, 5.59]	**3.03**	[.48, 5.53]	**3.10**	[.29, 5.92]
	$A_{d(\delta)}$	-2.17	[-6.31, 1.20]	-1.33	[-4.50, 1.84]	-1.55	[-5.40, 2.21]	-2.27	[-6.59, 1.12]
	$A_{d(\zeta)}$	.20	[-2.58, 3.00]	.21	[-1.90, 2.29]	.20	[-1.89, 2.34]	.21	[-1.89, 2.28]

Note. EAP = Expected A Posteriori estimate; 95% CI = 95% credible interval. EAP estimates with 95% CI not covering 0 are in bold face.

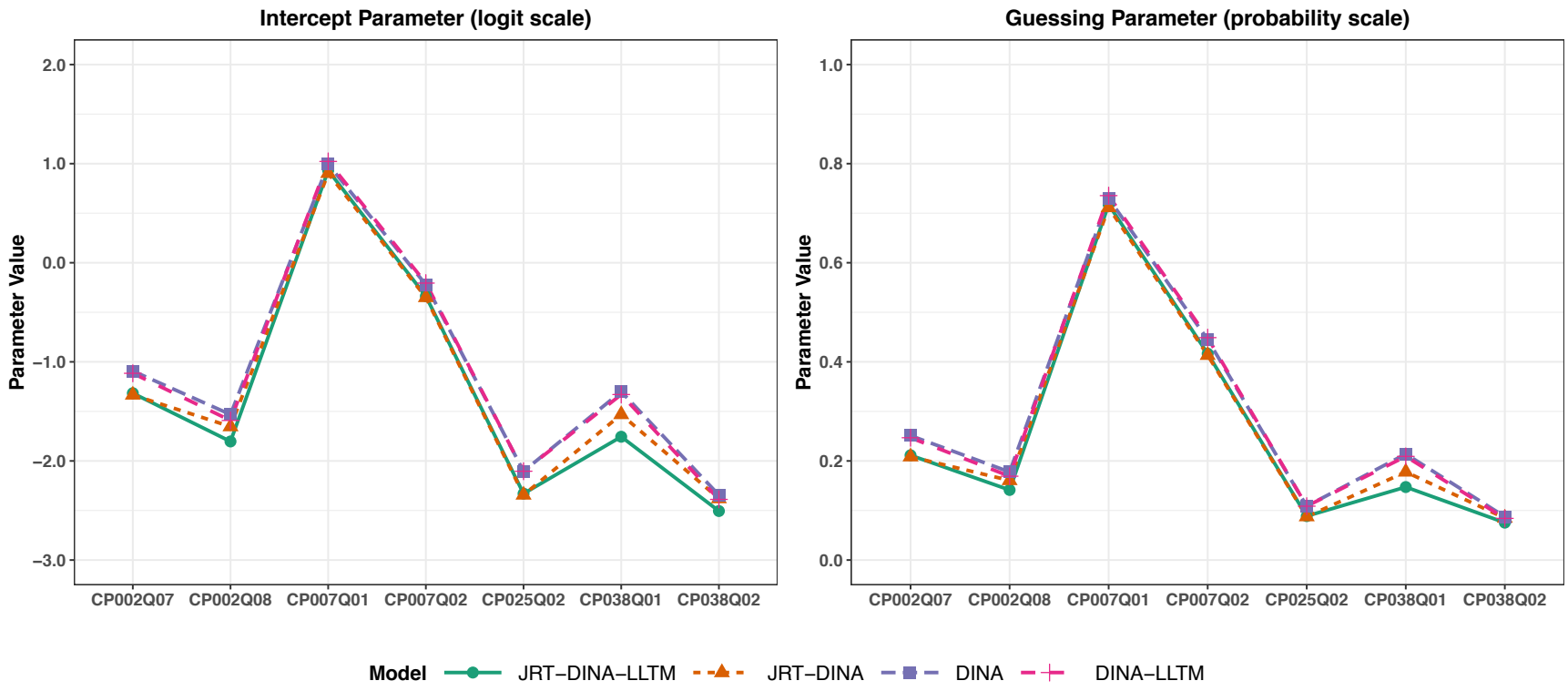


Figure 72. Consistency of the item intercept parameter estimates among data fitting models.

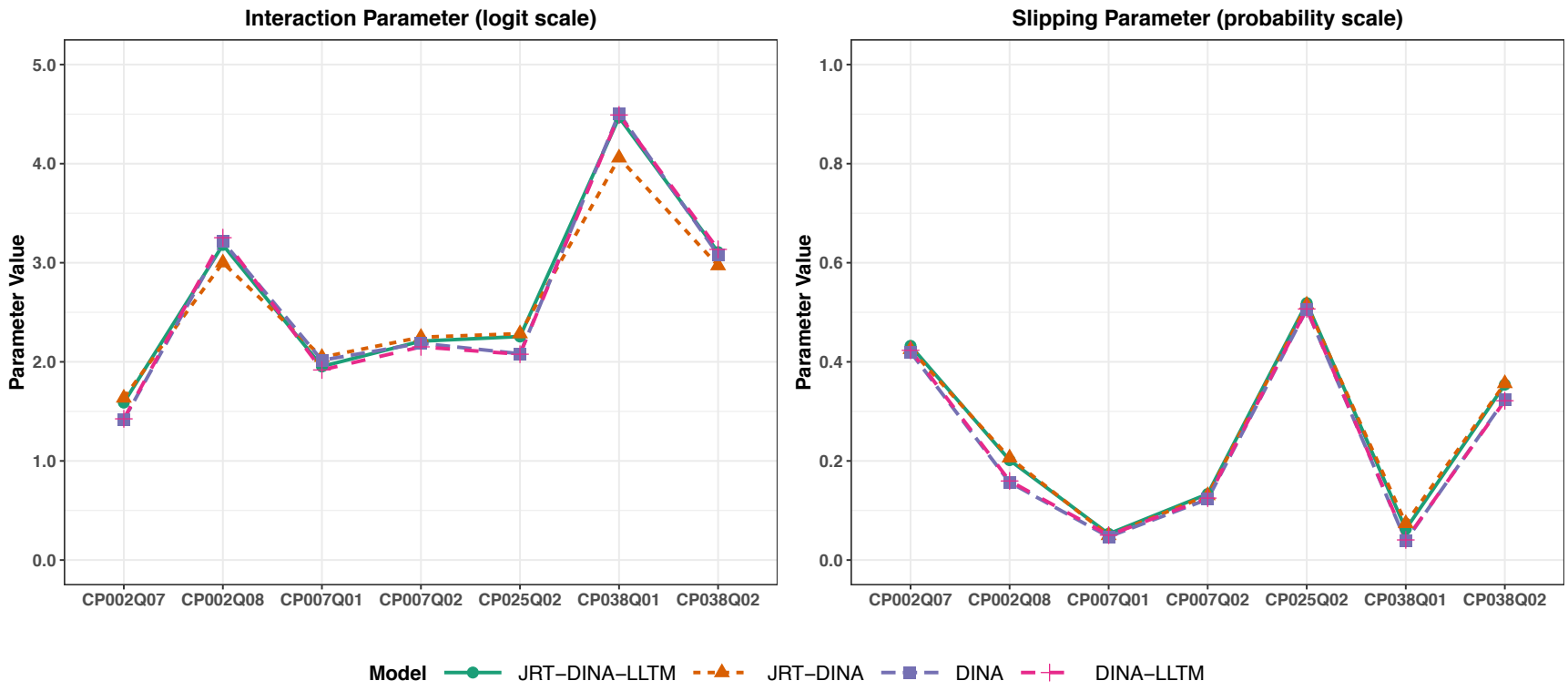


Figure 73. Consistency of the item interaction parameter estimates among data fitting models.

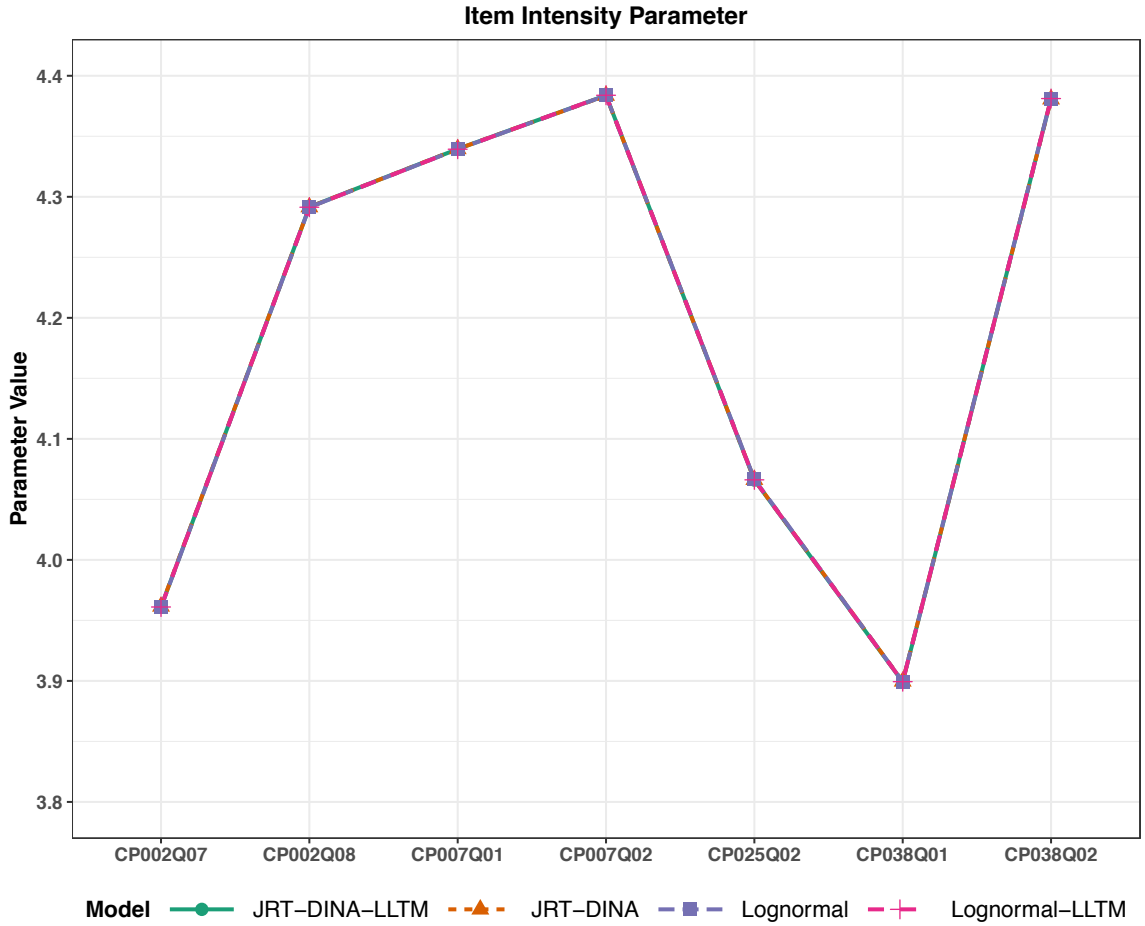


Figure 74. Consistency of the item intensity parameter estimates among data fitting models.

Table 44. *Attribute Profile Classification Consistency Among the Data Fitting Models*

Model	JRT-DINA-LLTM	JRT-DINA	DINA
JRT-DINA	0.990		
DINA	0.923	0.932	
DINA-LLTM	0.922	0.929	0.982

Table 45. *Attribute Classification Consistency Among the Data Fitting Models*

Model	JRT-DINA-LLTM	JRT-DINA	DINA
JRT-DINA	0.998/0.992		
DINA	0.991/0.932	0.993/0.939	
DINA-LLTM	0.999/0.923	0.998/0.931	0.991/0.991

Note. Values in each cell indicate the classification consistency classification consistency for attribute1/attribute 2.

Chapter 6: Discussion

It is challenging for item writers to develop items with targeted characteristics (e.g., slipping, guessing) in diagnostic assessments. With the rapid development of digital-based assessments, it is also necessary to control the time required to complete the items in assessments with time limit. To cater to these needs, the current study proposes the explanatory diagnostic classification modeling incorporating response times to provide information on how the item features are related to item parameters. A simulation study was conducted to evaluate the parameter recovery of the proposed model using the Bayesian MCMC estimation. In addition, an empirical data analysis was conducted to illustrate the application of the proposed method. Results have shown that the proposed model has satisfactory parameter recovery and has the potential in providing information on how the item features are related to item psychometric properties in the test development process. In this chapter, the findings from the simulation study and the empirical data analysis are summarized in section 6.1 and 6.2. Recommendations, limitations and future directions are provided in section 6.3.

6.1 Findings from the Simulation Study

The simulation study aims to address three research questions as stated in section 1.2:

1) How do the relative model fit indices perform in identifying the proposed model as the best data-fitting model in comparison with alternative models with misspecifications?
2) How accurate is the parameter recovery of the proposed model under different simulation conditions (i.e., sample size, test length, predictive power of the item covariates)?
3) How do the model misspecifications (e.g., the two-step estimation method and the one-step estimation) affect parameter recovery compared to that of the proposed model?

The findings from the simulation study are summarized with respect to each research question in as follows.

Research Question 1: How do the relative model fit indices perform in identifying the proposed model as the best data-fitting model in comparison with alternative models with misspecifications?

Relative model fit indices investigated in the current study include AIC, BIC and DIC. Specifically, AIC and BIC are traditionally developed to be used in the frequentist framework, while DIC is the generalization of AIC to be used in the Bayesian framework. All three model fit indices take into account of how well the data fitting model fits the dataset and penalize model complexity. Specifically, BIC has larger penalty of the model complexity than AIC, while DIC had a term that takes into account of effective number of parameters in the model.

In the simulation study, the proposed model JRT-DINA-LLTM, which was also the true data fitting model, had the smallest deviance in all replications under all simulated conditions. Yet, the deviances of the competing models were very close to that of JRT-DINA-LLTM. AIC and BIC tended to choose models omitting one covariate (i.e., JRT-DINA-LLTM-C and JRT-DINA-LLTM-D) under all conditions. This suggests that AIC and BIC penalized the JRT-DINA-LLTM for its model complexity. On the other hand, DIC tended to choose JRT-DINA-LLTM only when test length was long and the explanatory power of the item covariates was large. This also indicates that DIC tended to penalize JRT-DINA-LLTM when the number of items was small due to its model complexity. Furthermore, evidence ratio (Anderson, 2008) was used to determine whether the differences of the model fit among the data fitting models were significant. It was found that JRT-DINA-LLTM was not significantly different from the best fitting model (i.e., models with covariate misspecifications) identified by AIC, while it was

significantly different from the best fitting model (i.e., models with covariate misspecifications) identified by BIC. This is because BIC had more penalty on model complexity than AIC. The evidence ratio findings for DIC were similar as the findings on using DIC itself that DIC tended to select JRT-DINA-LLTM when test length was long and the explanatory power of the item covariates was large. This was also consistent with the model fit results obtained from the empirical data analysis. Therefore, AIC, BIC, and DIC did not perform well in correctly identifying the proposed model among other joint models with misspecifications.

The performance of relative fit indices for DINA and DINA-LLTM was similar with that of the joint models. AIC and BIC chose DINA-LLTM, but the results for DIC were mixed. DIC also tended to choose DINA-LLTM when test length was long and the explanatory power of item covariates was large. However, all three relative model fit indices chose Lognormal-LLTM over the Lognormal model in all replications under all simulated conditions. Evidence ratio findings on the three model fit indices suggested the same conclusions on these separate models.

Research Question 2: How is the parameter recovery of the proposed model under different simulation conditions?

Four factors were manipulated in the simulation study, including sample size, test length, correlation between ability and speed, and explanatory power of the item covariates. In general, the parameters of the proposed model JRT-DINA-LLTM recovered well under all simulation conditions. However, each of the four manipulated factors had impact on at least one type of parameters in terms of the parameter recovery, which is discussed in the following paragraphs.

First, test length and sample size were found to have significant ordinal interaction effects on person parameters (i.e., person speed, person ability, latent attributes) and item parameters (i.e., item intercept, item interaction, item intensity), respectively. Specifically, long test length

was found to lead to smaller estimation error in person parameters while large sample size was found to lead to smaller estimation error in item parameters (i.e., item intercept, item interaction, item intensity). These findings are consistent with the literature (e.g., Kang, 2016; Liao, 2018, Marianti, 2015; Suh, 2010; Wang, et al., 2013). It is also expected from the frequentist estimation perspective that larger sample size and longer test length would yield more accurate item parameter estimates and person parameter estimates.

Second, large correlation between ability and speed yielded smaller total error in person ability parameter estimation and better attribute classification accuracy. This is consistent with existing findings that more information shared between the latent traits lead to more accurate measurement precision (Klein Entink, 2009; Patton, 2015). Yet, the speed estimates were not significantly affected by correlation between ability and speed because the information is shared asymmetrically (Liao, 2018).

Third, the explanatory power of the item covariates was manipulated to be small or large in the current study. It is found that large the explanatory power of the item covariates led to less estimation error in person ability parameter, latent attributes, and the item parameters. Given that the same set of item covariates were used for all item parameters, the magnitude of the explanatory power of the item covariates reflects the correlations among the item parameters. Therefore, this indicates that the correlated random item effects affected the parameter estimates to some extent. Molenaar et al. (2014), however, showed that omission of the correlated random item effects did not affect parameter estimates substantially. It is necessary to point out several differences in the simulation setup between the current study and Molenaar et al. (2014). First, the response model used in the current study is the DINA model, while Molenaar et al. (2014) used 2PL model. Second, Molenaar et al. (2014) investigated correlations among item

parameters from 0 to 0.6, while the correlations investigated in the current study ranged from 0.1 to 0.8. Third, Molenaar et al. (2014) used marginal maximum likelihood (Bock & Aitkin, 1981) to estimate the models omitting the random item effects, while the current study used Bayesian MCMC estimation. Lastly, Molenaar et al. (2014) focused on the biases of the item parameter estimates, while the current study mainly found the impact of explanatory power of item covariates on the SE and RMSE of the item parameter estimates. Therefore, it is important to consider the inclusion of the correlated random item effects especially when the substantive research context suggests so.

Lastly, the parameter recovery of the regression coefficients in the proposed model is of particular interest in the current study. Both the recovery of the point estimates and standard errors of the regression coefficients were examined. It has been found that longer test length and larger explanatory power of the item covariates yielded more accurate and precise regression coefficient estimates. In general, the parameter recovery of the regression coefficients of item intensity parameter was better than that of the item parameters in the DINA model because the item intensity parameter had trivial estimation errors. For the point estimates of regression coefficients of item parameters in the DINA model, marginal mean biases when test length was 20 and explanatory power of the item was small were as high as 0.2 to 0.3. Furthermore, the proposed model overestimated the standard errors of the regression coefficients given the positive marginal mean biases of the standard errors. However, the biases were negligible when test length was long and explanatory power of item covariates was large especially for continuous covariates. Therefore, it is crucial to include item covariates with large explanatory power and use enough items to estimate the regression coefficients. In addition, further explorations are required on the inflated standard error estimates. It is worth pointing out that the

current study used noninformative priors for the regression coefficients. It is always possible to use more informative priors to make the posterior SDs of the regression coefficients smaller. Another possible reason is the estimation method using JAGS in the current study. JAGS implements slice-within-Gibbs sampling and makes it difficult to use the truncated multivariate normal distribution to set the positive constraint on the item interaction parameter. Thus, other sampling methods, such as the acceptance-rejection sampling, may be explored in the future.

Research Question 3: How do the model misspecifications (e.g., the two-step estimation method and the one-step estimation) affect parameter recovery compared to that of the proposed model?

The current study considered three competing models (i.e., models with misspecifications): 1) separate models that omit the second-level structural component between the latent traits; 2) models with covariate misspecifications; 3) models with two-step estimation. The differences of parameter recovery between the proposed model and that of the competing models are summarized in the following paragraphs.

First, separate models (i.e., DINA, DINA-LLTM) had less accurate person ability parameter estimates, person speed parameter estimates, latent attribute estimates, the variance of speed parameter estimates and high-order structural parameters. These parameters are all on the person side. The discrepancy became larger especially when the correlation between ability and speed was larger than 0.5. It has been discussed above that shared information between the two latent traits lead to higher measurement precision (Klein Entink, 2009; Patton, 2015). Therefore, it is suggested that joint models should be considered when modeling both item responses and response times incorporating item covariates. Admittedly, response times can be incorporated in

other ways, such as adding as a covariate, in the response model. However, the benefit of joint modeling is that the correlation between person ability and person speed can be estimated.

Second, two types of models with covariate misspecifications were investigated: JRT-DINA-LLTM-C (i.e., omitting the dichotomous covariate) and JRT-DINA-LLTM-D (i.e., omitting the continuous covariate). Both types of covariate misspecifications yielded larger estimation errors in item parameters in the DINA model and regression coefficients, but JRT-DINA-LLTM-D was more problematic. Specifically, JRT-DINA-LLTM-D had the largest random error and total error in the item intercept and item interaction parameter estimates among all data fitting models. In addition, the regression coefficient estimates for the dichotomous covariate obtained from JRT-DINA-LLTM-D had large mean biases and RMSE especially when test length was short and the explanatory power of the item covariates was large. Further, JRT-DINA-LLTM-D yielded the largest standard error biases of the regression coefficient of the dichotomous covariates. However, the current study investigated an extreme scenario where only two covariates (i.e., one continuous and one dichotomous) exist. Future studies are needed to examine scenarios where more covariates are included.

Third, models using two-step estimation (i.e., JRT-DINA, DINA + Lognormal) and one-step estimation (i.e., JRT-DINA-LLTM, DINA-LLTM + Lognormal-LLTM) were included in the simulation study to compare the parameter recovery of item parameters and regression coefficients. Given that the advantages of using joint models have been discussed in previous paragraphs, this paragraph focuses on the difference between JRT-DINA and JRT-DINA-LLTM. In terms of item parameters, both models yielded satisfactory parameter recovery in terms of bias, SE and RMSE. This also explains the high consistency of item parameter estimates obtained in the empirical data analysis from different data fitting models that will be discussed in

the next section. High consistency of the item parameter estimates between the two-step and one-step estimations indicates the potential of using the scored item covariates to predict item parameter. In terms of regression coefficients, JRT-DINA and JRT-DINA-LLTM yielded similarly well parameter recovery for item intercept parameter and item intensity parameter. However, JRT-DINA had smaller mean biases in both the point estimate and standard errors of the regression coefficients of the item interaction parameter than JRT-DINA-LLTM. One possible reason is that different ways were used to set the positive constraints on the item interaction parameter in the two models. It is worthwhile to explore other estimation methods than JAGS in the estimation of the two models. To conclude, two-step estimation and one-step estimation yielded similar parameter estimates, which is consistent with the findings on IRT models (Green & Smith, 1987). As in Green and Smith (1987), two-step estimation provides additional convenience in fitting regression models with interaction terms and polynomials in readily available software programs.

6.2 Findings from the Empirical Data Analysis

To illustrate the application of the proposed method, the JRT-DINA-LLTM and competing models including JRT-DINA, DINA-LLTM + Lognormal-LLTM, and DINA + Lognormal were applied to analyze an empirical dataset from PISA 2012 problem-solving items. Two research questions were addressed in the empirical data analysis:

4) According to the empirical data analysis, do the item covariates have significant effects on the item parameters from CDM and the RT model, respectively?
5) How is the consistency of the item parameter estimates and attribute estimates among the data-fitting models?

Research Question 4: Do the item covariates have significant effects on the item parameters from CDM and the RT model, respectively? There were two item covariates used in the empirical study: average steps taken to complete the item and item type (familiar context or unfamiliar context). Only the regression coefficient of the item type in the item intercept parameter was found to significantly deviate from 0 for all the data fitting models. This indicates that item type with familiar context is more likely to be guessed correctly than the item type with unfamiliar context to the respondents. In fact, the guessing probability of one item with familiar context was about 0.75, which indicates that it is too easy to be guessed correctly. This information can help item writers to revise this item type, although it would be ideal to have more item covariates included in the models to inform item writers of which aspects of the item type should be revised. In addition, the nonsignificant item covariates indicate that the items do not differ much from each other in terms of the item covariates included.

Several limitations in the empirical data analysis remains. First, there were only 7 items included in the analysis due to the limit access to a more suitable dataset. Thus, the non-significance can be due to lack of statistical power. Second, some items shared the same context and, thus, item local dependence may exist. This can be tackled by extending the current model to include a testlet structure. However, this was not considered due to the purpose and scope of the current study.

Research Question 5: How is the consistency of the item parameter estimates and attribute estimates among the data-fitting models? Item parameter estimates are very consistent among data fitting models, especially item intensity parameter estimates. As shown in the simulation study, the mean biases of the item parameter estimates were similarly small for all the data fitting models. This also indicates that the models with covariates yield similar item

parameter estimates with models using the second step estimation. This further shows the potential of using scored item covariates to predict item parameters in a variety of testing scenarios. For example, in the CD-CAT scenario, it can be challenging to develop items for respondents at the two end of the ability continuum. Thus, the ability estimation for these respondents can be less precise than those respondents with ability levels in the middle. It would be beneficial to generate items under this scenario using scored item covariates.

The attribute and attribute pattern classifications were almost perfectly aligned between the two joint models and between the two separate models respectively. In addition, the classification between joint models and separate models were consistent as well. Given that the correlation between ability and speed estimate in the empirical dataset was about 0.2, the attribute classification between joint models and separate models was about 0.95, while the attribute pattern classification between joint models and separate models was about 0.90. This finding also aligns with the finding from simulation study that the attribute classification accuracy of the separate models was similar to that of the joint models when the true population model indicating a correlation between ability and speed smaller than 0.5.

6.3 Limitations and Future Studies

The current study has several limitations and future directions which could be addressed in further studies. Specifically, five aspects are identified related to model assumptions, posterior predictive model check, ways to extract item covariates, and generalizability of simulation study.

Local dependence. The current study considers RA and RT to be both locally independent as in van der Linden (2007). However, as reviewed in literature in Chapter 2, it is reasonable to explore the local dependence between the two data sources (e.g., Bolsinova & Tijmstra, 2017; Molenaar, et al., 2016). Specifically, the measurement model for RT can be

dependent on response data and the distribution of responses can vary according to RTs. In addition, as shown in Liao (2018), location shift of RT between correct and incorrect responses for each item is strongly correlated with the item difficulty. This shifting effect can be explored in the joint modeling of DCM and can be explained by the item covariates in the LLTM manner. Furthermore, the distribution of RTs has been shown to be mixture of subpopulations of respondents (e.g., Wang & Xu, 2015). The existence of mixture of RTs may imply aberrant behavior such as low motivation and cheating. Therefore, in the application of the proposed model, it is worthwhile to incorporate the mixture component in RTs if necessary.

Posterior predictive model check. The posterior predictive model check was conducted as the absolute model fit measure of the data fitting models. The discrepancy measure used in the current study was the sum of squared residuals. Then a PPP-value was calculated to summarize the comparison between the discrepancy measure obtained from the observed data and the replicated data. However, different discrepancy measures capture different aspects of the model-data fit. Future explorations are needed to investigate the performance of other discrepancy measures. For example, total-score distribution, odds ratios between item pairs, and correlations between attribute pairs have been investigated as the discrepancy measures in the application of DCMs (Park, Johnson, & Lee, 2015). In addition, sample size may affect the power of the posterior predictive model check. When sample size is small, the power of the posterior predictive model check can be affected by the priors (Berkhof et al., 2000). Furthermore, given that PPP-values measure the statistical significance of the model-data fit, it is suggested to consider practical significance of the difference between the model and the dataset which depends on the substantive interest of the study (Gelman et al., 2014).

Methods to extract item covariates. In real-world applications of the proposed model, one important question to answer is how to extract item covariates with large predictive power. As indicated by the findings in the simulation study, large predictive power of item covariates led to smaller estimation error of the item parameters and regression coefficients. Some recommendations on how to extract item covariates are given as follows. First, item covariates can be extracted based on test design frameworks. Usually, test design frameworks specify item stimulus features and define the psychometric properties of the items (Embretsen, 1999). Such item covariates may include item types and maximum score category which can be used to predict guessing, slipping and time intensity of the items. Second, in digital-based assessments, data obtained during students' problem-solving processes can be used item covariates. As shown in the empirical study, average steps used to complete the items was used as an item covariate to indicate the cognitive complexity of the items. Third, it is possible to extract linguistic covariates from the item stems using techniques such as natural language processing (NLP). For example, linguistic covariates may include the number of words and the number of sentences which can be extracted as raw tokens. In addition, vocabulary complexity and phrase complexity of the item stem can be determined by comparing the vocabulary or phrase with existing word list, such as Dale-Chall word list (Dale & Chall, 1948). Using the linguistic covariates from the item stems can also reveal construct irrelevant factors if the high slipping probability of an item is due to vocabulary complexity of the item stems rather than the construct being measured. Lastly, after obtaining a large number of item covariates, it is necessary to select the ones with large predictive power. Statistical learning methods, such as lasso regression, can be used to implement covariate selection.

Generalizability of the simulation study. The simulation study design set several constraints to make sure the work can be done in a certain amount of time frame. However, future studies can explore more factors to make the simulation study more generalizable. For example, the simulation study only included two item covariates, one continuous and one dichotomous. Models with covariate misspecifications can be extended to scenarios where more covariates exist. Additionally, it is worthwhile to consider the multicollinearity issue among the item covariates (Green & Smith, 1987). As shown in Green and Smith (1987), collinearity issues can lead to less accurate point estimates and standard errors of the regression coefficients especially when sample size was less than 200. Therefore, collinearity issues may as well affect the estimation of regression coefficients in the DCM and RT modeling framework.

Despite the limitations, the current study provides modeling framework for the explanatory diagnostic classification model incorporating response times. The simulation study presented satisfactory parameter recovery of the proposed model. The empirical data analysis demonstrated the application of explanatory diagnostic classification models and RT models. In the digital assessment age, both item responses and response times play an important role in the test development procedures. The utilities of the proposed method are two-fold. First, the scored item covariates can be used to predict item parameters in the diagnostic classification model and the RT model, which can be beneficial in a variety of testing scenarios when the calibration sample is not large enough. Second, the regression coefficient estimates of the item covariates can help item writers better understand how the item features determine the psychometric properties of the items and how to improve the measure. In sum, the current study builds a good foundation for future studies on explanatory diagnostic classification models and response time models and demonstrates promising applications in digital-based assessments.

Appendix A: Classification Accuracy, Bias, SE and RMSE Results by the Simulated Conditions (Person Parameters)

Table A. 1. *Attribute Profile Correct Classification Rate (Attribute 1)*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.952	.951	.951	.951	.951	.951
			Large	.972	.972	.972	.972	.973	.972
		0.5	Small	.957	.957	.958	.958	.957	.957
			Large	.977	.977	.977	.977	.977	.977
		0.8	Small	.962	.963	.960	.960	.962	.962
			Large	.978	.977	.976	.976	.977	.978
	40	0.2	Small	.987	.987	.987	.987	.987	.987
			Large	.993	.993	.993	.993	.993	.993
		0.5	Small	.989	.989	.988	.989	.989	.989
			Large	.994	.994	.994	.994	.994	.994
		0.8	Small	.989	.989	.989	.989	.989	.989
			Large	.995	.995	.995	.995	.995	.995
1000	20	0.2	Small	.963	.963	.963	.963	.963	.963
			Large	.973	.973	.972	.973	.972	.972
		0.5	Small	.961	.961	.962	.961	.961	.961
			Large	.976	.976	.976	.976	.976	.976
		0.8	Small	.963	.963	.961	.961	.963	.963
			Large	.979	.978	.977	.978	.978	.978
	40	0.2	Small	.988	.988	.988	.988	.988	.988
			Large	.994	.994	.994	.994	.994	.994
		0.5	Small	.988	.988	.987	.987	.988	.988
			Large	.994	.994	.994	.994	.994	.994
		0.8	Small	.988	.988	.988	.988	.988	.987
			Large	.993	.993	.993	.993	.993	.993

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 2. *Attribute Profile Correct Classification Rate (Attribute 2)*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.953	.953	.953	.953	.953	.953
			Large	.960	.960	.962	.960	.960	.960
		0.5	Small	.946	.946	.946	.945	.946	.946
			Large	.966	.965	.965	.965	.965	.965
		0.8	Small	.945	.945	.944	.944	.945	.945
			Large	.968	.968	.963	.962	.968	.968
	40	0.2	Small	.989	.989	.988	.989	.989	.989
			Large	.987	.987	.986	.987	.987	.986
		0.5	Small	.991	.991	.990	.990	.991	.991
			Large	.988	.988	.989	.989	.989	.988
		0.8	Small	.993	.993	.991	.991	.993	.993
			Large	.987	.987	.988	.988	.987	.987
1000	20	0.2	Small	.949	.949	.949	.949	.949	.949
			Large	.963	.963	.963	.963	.963	.963
		0.5	Small	.956	.956	.956	.956	.956	.956
			Large	.964	.964	.962	.962	.964	.964
		0.8	Small	.952	.952	.952	.952	.952	.953
			Large	.965	.965	.962	.962	.965	.965
	40	0.2	Small	.988	.988	.988	.988	.988	.988
			Large	.988	.988	.988	.988	.988	.988
		0.5	Small	.991	.991	.991	.991	.991	.991
			Large	.986	.986	.986	.986	.986	.986
		0.8	Small	.991	.991	.990	.990	.991	.991
			Large	.987	.987	.987	.987	.987	.986

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 3. *Attribute Profile Correct Classification Rate (Attribute 3)*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.939	.939	.940	.939	.939	.939
			Large	.967	.967	.967	.967	.967	.967
		0.5	Small	.948	.948	.945	.944	.948	.947
			Large	.968	.968	.968	.968	.968	.968
		0.8	Small	.944	.944	.939	.939	.944	.943
			Large	.967	.967	.965	.966	.967	.967
	40	0.2	Small	.990	.990	.990	.990	.990	.990
			Large	.996	.996	.996	.996	.996	.996
		0.5	Small	.992	.992	.992	.992	.992	.992
			Large	.995	.995	.995	.995	.995	.995
		0.8	Small	.995	.995	.994	.993	.995	.995
			Large	.996	.996	.996	.996	.996	.996
1000	20	0.2	Small	.950	.950	.948	.948	.950	.950
			Large	.969	.969	.969	.969	.969	.969
		0.5	Small	.948	.949	.948	.949	.948	.948
			Large	.967	.967	.967	.967	.967	.967
		0.8	Small	.945	.945	.942	.942	.945	.945
			Large	.967	.967	.966	.966	.967	.967
	40	0.2	Small	.992	.992	.992	.992	.992	.992
			Large	.996	.996	.996	.996	.996	.996
		0.5	Small	.992	.992	.991	.991	.992	.992
			Large	.996	.996	.996	.996	.996	.996
		0.8	Small	.993	.993	.993	.993	.993	.993
			Large	.996	.996	.996	.996	.996	.996

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 4. *Attribute Profile Correct Classification Rate (Attribute 4)*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.953	.953	.954	.953	.953	.953
			Large	.970	.969	.970	.970	.971	.970
		0.5	Small	.958	.958	.958	.958	.959	.959
			Large	.971	.971	.970	.970	.971	.971
		0.8	Small	.958	.959	.957	.957	.958	.958
			Large	.971	.970	.969	.969	.972	.970
	40	0.2	Small	.990	.990	.990	.990	.990	.990
			Large	.987	.987	.987	.987	.987	.986
		0.5	Small	.992	.993	.993	.993	.993	.993
			Large	.988	.988	.988	.988	.988	.987
		0.8	Small	.993	.993	.992	.993	.993	.993
			Large	.988	.987	.985	.986	.988	.987
1000	20	0.2	Small	.957	.958	.958	.958	.957	.957
			Large	.972	.971	.971	.972	.972	.971
		0.5	Small	.957	.957	.957	.957	.957	.957
			Large	.968	.967	.967	.967	.968	.968
		0.8	Small	.960	.961	.959	.958	.960	.960
			Large	.974	.974	.973	.973	.974	.974
	40	0.2	Small	.991	.991	.991	.991	.991	.991
			Large	.987	.987	.987	.987	.987	.987
		0.5	Small	.992	.992	.992	.992	.992	.992
			Large	.986	.985	.985	.985	.986	.986
		0.8	Small	.992	.992	.992	.992	.992	.992
			Large	.987	.987	.987	.987	.987	.987

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 5. *Mean and Standard Deviation of Bias of the Person Ability Parameter Estimates*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.000	.000	.000	.000	.000	.000	.730	.730	.741	.741	.730	.730
			Large	.000	.000	.000	.000	.000	.000	.716	.715	.723	.723	.716	.715
		0.5	Small	.000	.000	.000	.000	.000	.000	.682	.682	.766	.766	.682	.682
			Large	.000	.000	.000	.000	.000	.000	.691	.691	.773	.773	.691	.691
		0.8	Small	.000	.000	.000	.000	.000	.000	.530	.530	.775	.775	.530	.530
			Large	.000	.000	.000	.000	.000	.000	.536	.536	.738	.738	.536	.536
	40	0.2	Small	.000	.000	.000	.000	.000	.000	.756	.756	.767	.767	.756	.756
			Large	.000	.000	.000	.000	.000	.000	.761	.761	.771	.771	.761	.761
		0.5	Small	.000	.000	.000	.000	.000	.000	.668	.668	.744	.744	.668	.668
			Large	.000	.000	.000	.000	.000	.000	.675	.675	.743	.743	.675	.675
		0.8	Small	.000	.000	.000	.000	.000	.000	.524	.524	.777	.777	.524	.524
			Large	.000	.000	.000	.000	.000	.000	.530	.530	.754	.754	.530	.530
1000	20	0.2	Small	.000	.000	.000	.000	.000	.000	.758	.758	.758	.758	.758	.758
			Large	.000	.000	.000	.000	.000	.000	.751	.751	.761	.761	.751	.751
		0.5	Small	.000	.000	.000	.000	.000	.000	.705	.705	.772	.772	.705	.705
			Large	.000	.000	.000	.000	.000	.000	.685	.685	.755	.755	.685	.685
		0.8	Small	.000	.000	.000	.000	.000	.000	.501	.501	.726	.726	.501	.501
			Large	.000	.000	.000	.000	.000	.000	.505	.507	.707	.707	.506	.507
	40	0.2	Small	.000	.000	.000	.000	.000	.000	.745	.745	.755	.755	.745	.745
			Large	.000	.000	.000	.000	.000	.000	.748	.748	.759	.759	.748	.748
		0.5	Small	.000	.000	.000	.000	.000	.000	.688	.688	.764	.764	.688	.688
			Large	.000	.000	.000	.000	.000	.000	.702	.702	.781	.781	.702	.702
		0.8	Small	.000	.000	.000	.000	.000	.000	.524	.524	.778	.778	.524	.524
			Large	.000	.000	.000	.000	.000	.000	.508	.508	.736	.736	.508	.508

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 6. *Mean and Standard Deviation of SE of the Person Ability Parameter Estimates*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.199	.199	.208	.207	.199	.200	.064	.065	.069	.069	.065	.065
			Large	.191	.192	.197	.198	.191	.191	.066	.066	.068	.068	.066	.066
		0.5	Small	.154	.154	.172	.171	.154	.154	.054	.054	.075	.076	.054	.054
			Large	.134	.134	.144	.143	.134	.135	.053	.053	.068	.067	.053	.053
		0.8	Small	.180	.179	.229	.229	.181	.179	.031	.031	.079	.079	.032	.031
			Large	.196	.195	.298	.298	.196	.195	.036	.036	.109	.109	.036	.036
	40	0.2	Small	.092	.092	.094	.094	.092	.092	.054	.053	.058	.058	.054	.054
			Large	.080	.080	.082	.081	.080	.081	.043	.043	.047	.047	.043	.043
		0.5	Small	.079	.079	.074	.074	.079	.079	.039	.039	.055	.055	.039	.039
			Large	.085	.085	.085	.084	.085	.085	.031	.031	.048	.048	.031	.031
		0.8	Small	.099	.099	.080	.080	.098	.099	.019	.019	.050	.050	.019	.019
			Large	.131	.132	.195	.194	.131	.131	.036	.036	.115	.115	.036	.036
1000	20	0.2	Small	.157	.157	.164	.164	.157	.157	.063	.063	.067	.067	.063	.063
			Large	.134	.134	.138	.137	.135	.135	.062	.062	.066	.065	.062	.061
		0.5	Small	.149	.149	.163	.163	.149	.148	.048	.048	.070	.070	.048	.048
			Large	.134	.134	.137	.137	.134	.134	.047	.047	.065	.065	.047	.047
		0.8	Small	.169	.170	.202	.202	.169	.169	.038	.038	.075	.075	.038	.037
			Large	.189	.185	.257	.257	.185	.185	.039	.036	.104	.104	.036	.036
	40	0.2	Small	.081	.082	.081	.082	.081	.081	.054	.054	.056	.056	.054	.054
			Large	.075	.074	.075	.074	.075	.075	.049	.049	.054	.054	.049	.049
		0.5	Small	.085	.084	.077	.078	.085	.085	.038	.038	.055	.055	.038	.038
			Large	.081	.081	.073	.073	.081	.081	.034	.034	.053	.053	.034	.034
		0.8	Small	.126	.127	.135	.135	.126	.126	.025	.025	.071	.071	.025	.025
			Large	.108	.108	.073	.073	.108	.108	.018	.018	.053	.052	.019	.019

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 7. *Mean and Standard Deviation of RMSE of the Person Ability Parameter Estimates*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.640	.640	.651	.650	.640	.641	.410	.410	.419	.419	.410	.410
			Large	.629	.629	.638	.638	.629	.629	.398	.397	.401	.401	.398	.398
		0.5	Small	.583	.583	.655	.655	.583	.583	.389	.389	.439	.439	.389	.389
			Large	.593	.593	.663	.663	.593	.593	.383	.382	.427	.427	.383	.383
		0.8	Small	.482	.481	.684	.684	.482	.481	.289	.289	.438	.438	.288	.288
			Large	.492	.492	.695	.695	.492	.492	.293	.293	.404	.404	.293	.293
	40	0.2	Small	.615	.615	.625	.625	.615	.615	.454	.454	.459	.459	.454	.454
			Large	.622	.622	.628	.628	.622	.623	.449	.449	.459	.459	.449	.449
		0.5	Small	.558	.558	.620	.620	.558	.558	.378	.378	.421	.421	.378	.378
			Large	.555	.555	.614	.614	.555	.555	.394	.394	.430	.430	.394	.394
		0.8	Small	.436	.436	.628	.628	.436	.436	.308	.308	.468	.469	.309	.308
			Large	.458	.458	.656	.656	.458	.458	.300	.300	.436	.436	.300	.300
1000	20	0.2	Small	.646	.646	.650	.650	.646	.646	.431	.430	.428	.428	.430	.430
			Large	.633	.633	.644	.644	.634	.634	.429	.429	.432	.433	.429	.429
		0.5	Small	.600	.600	.657	.657	.600	.600	.403	.403	.442	.442	.403	.403
			Large	.569	.569	.627	.627	.569	.569	.407	.406	.446	.446	.406	.406
		0.8	Small	.449	.449	.632	.632	.449	.449	.280	.280	.418	.418	.280	.280
			Large	.465	.464	.647	.646	.464	.464	.275	.277	.398	.398	.277	.277
	40	0.2	Small	.602	.602	.608	.608	.602	.602	.449	.449	.458	.458	.449	.449
			Large	.605	.605	.617	.617	.605	.606	.449	.449	.452	.452	.449	.449
		0.5	Small	.563	.563	.624	.624	.563	.563	.405	.405	.451	.451	.405	.405
			Large	.572	.572	.636	.636	.572	.572	.415	.415	.461	.461	.415	.415
		0.8	Small	.452	.453	.651	.651	.452	.452	.293	.293	.451	.451	.293	.293
			Large	.430	.430	.593	.593	.430	.430	.291	.291	.446	.446	.291	.291

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 8. *Mean and Standard Deviation of Bias of the Person Speed Parameter Estimates*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	-.004	-.004	-.004	-.004	-.004	-.004	.030	.030	.030	.030	.030	.030
			Large	-.004	-.004	-.004	-.004	-.004	-.004	.031	.031	.031	.031	.031	.031
		0.5	Small	-.002	-.002	-.003	-.002	-.002	-.002	.031	.031	.031	.030	.031	.031
			Large	-.001	-.002	-.002	-.002	-.002	-.002	.031	.031	.030	.030	.031	.031
		0.8	Small	.006	.006	.005	.006	.006	.006	.032	.032	.030	.030	.032	.032
			Large	.006	.006	.005	.006	.006	.006	.033	.033	.031	.031	.033	.033
	40	0.2	Small	-.004	-.004	-.004	-.004	-.004	-.004	.018	.018	.018	.018	.018	.018
			Large	-.004	-.004	-.004	-.004	-.004	-.004	.019	.019	.019	.019	.019	.019
		0.5	Small	-.001	-.001	-.001	-.002	-.002	-.002	.020	.020	.020	.020	.020	.020
			Large	-.002	-.002	-.001	-.002	-.002	-.002	.020	.020	.019	.019	.020	.020
		0.8	Small	.006	.006	.006	.006	.006	.006	.019	.019	.018	.018	.019	.019
			Large	.006	.006	.006	.006	.006	.006	.019	.019	.018	.018	.019	.019
1000	20	0.2	Small	-.007	-.007	-.007	-.007	-.007	-.007	.030	.030	.029	.029	.030	.030
			Large	-.007	-.007	-.007	-.007	-.007	-.007	.031	.031	.031	.031	.031	.031
		0.5	Small	-.005	-.005	-.006	-.005	-.005	-.005	.031	.031	.031	.031	.031	.031
			Large	-.005	-.005	-.006	-.005	-.005	-.005	.031	.031	.030	.030	.031	.031
		0.8	Small	.004	.004	.004	.004	.004	.004	.034	.034	.032	.032	.034	.034
			Large	.004	.004	.004	.004	.004	.004	.032	.032	.030	.030	.033	.033
	40	0.2	Small	-.007	-.007	-.008	-.007	-.007	-.007	.018	.018	.018	.018	.018	.018
			Large	-.007	-.007	-.008	-.007	-.007	-.007	.019	.019	.019	.018	.019	.019
		0.5	Small	-.005	-.005	-.006	-.005	-.005	-.005	.019	.019	.019	.019	.019	.019
			Large	-.005	-.005	-.006	-.005	-.005	-.005	.019	.019	.018	.018	.019	.019
		0.8	Small	.004	.004	.004	.004	.004	.004	.019	.019	.018	.018	.019	.019
			Large	.004	.004	.004	.004	.004	.004	.020	.020	.019	.019	.020	.020

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 9. *Mean and Standard Deviation of SE of the Person Speed Parameter Estimates*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.104	.104	.104	.104	.104	.104	.014	.014	.014	.014	.014	.014
			Large	.104	.104	.104	.104	.104	.104	.014	.014	.014	.014	.014	.014
		0.5	Small	.103	.103	.103	.103	.103	.103	.014	.014	.014	.014	.014	.014
			Large	.105	.105	.105	.105	.105	.105	.013	.013	.013	.013	.013	.013
		0.8	Small	.103	.103	.104	.104	.103	.103	.013	.013	.014	.014	.013	.013
			Large	.102	.102	.104	.104	.102	.102	.013	.013	.014	.014	.013	.013
	40	0.2	Small	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
			Large	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
		0.5	Small	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
			Large	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
		0.8	Small	.074	.074	.075	.075	.074	.074	.010	.010	.010	.010	.010	.010
			Large	.075	.075	.075	.075	.075	.075	.009	.009	.009	.009	.010	.009
1000	20	0.2	Small	.104	.104	.104	.104	.104	.104	.014	.014	.014	.014	.014	.014
			Large	.104	.104	.104	.104	.104	.104	.014	.014	.014	.014	.014	.014
		0.5	Small	.103	.103	.104	.104	.103	.103	.014	.014	.014	.014	.014	.014
			Large	.104	.104	.105	.105	.104	.104	.013	.013	.013	.013	.013	.013
		0.8	Small	.102	.102	.104	.104	.102	.102	.013	.013	.014	.014	.013	.013
			Large	.102	.102	.104	.104	.102	.102	.014	.014	.014	.014	.014	.014
	40	0.2	Small	.075	.075	.075	.075	.075	.075	.009	.009	.009	.009	.009	.009
			Large	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
		0.5	Small	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
			Large	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
		0.8	Small	.075	.075	.075	.075	.075	.075	.010	.010	.010	.010	.010	.010
			Large	.074	.074	.075	.075	.074	.074	.010	.010	.010	.010	.010	.010

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 10. *Mean and Standard Deviation of RMSE of the Person Speed Parameter Estimates*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.109	.109	.109	.109	.109	.109	.014	.014	.014	.014	.014	.014
			Large	.108	.108	.108	.108	.108	.108	.014	.014	.014	.014	.014	.014
		0.5	Small	.107	.107	.108	.108	.107	.107	.015	.015	.015	.015	.015	.015
			Large	.109	.109	.109	.109	.109	.109	.014	.014	.014	.014	.014	.014
		0.8	Small	.108	.108	.108	.109	.108	.108	.014	.014	.014	.014	.014	.014
			Large	.107	.107	.108	.108	.107	.107	.015	.015	.014	.014	.015	.015
	40	0.2	Small	.077	.077	.077	.077	.077	.077	.010	.010	.010	.010	.010	.010
			Large	.078	.078	.078	.078	.078	.078	.010	.010	.010	.010	.010	.010
		0.5	Small	.078	.078	.078	.078	.078	.078	.010	.010	.010	.010	.010	.010
			Large	.077	.077	.077	.077	.077	.077	.010	.010	.010	.010	.010	.010
		0.8	Small	.077	.077	.077	.077	.077	.077	.011	.011	.011	.011	.011	.011
			Large	.077	.077	.078	.078	.077	.077	.010	.010	.010	.010	.010	.010
1000	20	0.2	Small	.108	.108	.108	.108	.108	.108	.015	.015	.015	.015	.015	.015
			Large	.108	.108	.108	.108	.108	.108	.015	.015	.015	.015	.015	.015
		0.5	Small	.108	.108	.108	.108	.108	.108	.014	.014	.014	.014	.014	.014
			Large	.109	.109	.109	.109	.109	.109	.014	.014	.014	.014	.014	.014
		0.8	Small	.108	.108	.108	.108	.108	.107	.014	.014	.014	.015	.014	.014
			Large	.107	.107	.108	.108	.107	.107	.015	.015	.015	.015	.015	.015
	40	0.2	Small	.077	.077	.077	.077	.077	.077	.010	.010	.010	.009	.010	.010
			Large	.078	.078	.078	.078	.078	.078	.010	.010	.010	.010	.010	.010
		0.5	Small	.077	.077	.077	.077	.077	.077	.010	.010	.010	.010	.010	.010
			Large	.077	.077	.077	.077	.077	.077	.010	.010	.010	.010	.010	.010
		0.8	Small	.077	.077	.077	.078	.077	.077	.010	.010	.010	.010	.010	.010
			Large	.077	.077	.078	.078	.077	.077	.010	.010	.010	.010	.010	.010

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 11. *Mean Bias of the Person Speed Variance Parameter Estimates*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	0.008	0.008	0.010	0.010	0.008	0.008
			Large	0.007	0.007	0.009	0.009	0.007	0.007
		0.5	Small	0.023	0.023	0.025	0.025	0.023	0.023
			Large	0.024	0.024	0.026	0.025	0.024	0.024
		0.8	Small	0.032	0.032	0.033	0.033	0.032	0.032
			Large	0.031	0.031	0.032	0.032	0.031	0.031
	40	0.2	Small	0.008	0.008	0.010	0.010	0.008	0.008
			Large	0.008	0.008	0.009	0.009	0.008	0.008
		0.5	Small	0.023	0.023	0.025	0.025	0.023	0.023
			Large	0.023	0.023	0.025	0.025	0.023	0.023
		0.8	Small	0.033	0.033	0.034	0.034	0.033	0.033
			Large	0.033	0.033	0.034	0.034	0.033	0.033
1000	20	0.2	Small	0.005	0.005	0.006	0.006	0.005	0.005
			Large	0.004	0.004	0.005	0.005	0.004	0.004
		0.5	Small	0.013	0.013	0.014	0.014	0.013	0.013
			Large	0.014	0.014	0.015	0.015	0.014	0.014
		0.8	Small	-0.007	-0.007	-0.007	-0.006	-0.007	-0.007
			Large	-0.005	-0.006	-0.005	-0.005	-0.006	-0.006
	40	0.2	Small	0.005	0.005	0.006	0.005	0.005	0.005
			Large	0.004	0.004	0.005	0.005	0.004	0.004
		0.5	Small	0.013	0.013	0.014	0.014	0.013	0.013
			Large	0.014	0.014	0.015	0.014	0.014	0.014
		0.8	Small	-0.006	-0.006	-0.005	-0.005	-0.006	-0.006
			Large	-0.007	-0.007	-0.006	-0.006	-0.007	-0.007

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 12. *Mean SE of the Person Speed Variance Parameter Estimates*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.5	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
	40	0.2	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.5	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
1000	20	0.2	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.5	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
	40	0.2	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.5	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	Small	0.000	0.000	0.000	0.000	0.000	0.000
			Large	0.000	0.000	0.000	0.000	0.000	0.000

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 13. *Mean RMSE of the Person Speed Variance Parameter Estimates*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	0.008	0.008	0.010	0.010	0.008	0.008
			Large	0.007	0.007	0.009	0.009	0.007	0.007
		0.5	Small	0.023	0.023	0.025	0.025	0.023	0.023
			Large	0.024	0.024	0.026	0.025	0.024	0.024
		0.8	Small	0.032	0.032	0.033	0.033	0.032	0.032
			Large	0.031	0.031	0.032	0.032	0.031	0.031
	40	0.2	Small	0.008	0.008	0.010	0.010	0.008	0.008
			Large	0.008	0.008	0.009	0.009	0.008	0.008
		0.5	Small	0.023	0.023	0.025	0.025	0.023	0.023
			Large	0.023	0.023	0.025	0.025	0.023	0.023
		0.8	Small	0.033	0.033	0.034	0.034	0.033	0.033
			Large	0.033	0.033	0.034	0.034	0.033	0.033
1000	20	0.2	Small	0.005	0.005	0.006	0.006	0.005	0.005
			Large	0.004	0.004	0.005	0.005	0.004	0.004
		0.5	Small	0.013	0.013	0.014	0.014	0.013	0.013
			Large	0.014	0.014	0.015	0.015	0.014	0.014
		0.8	Small	0.007	0.007	0.007	0.006	0.007	0.007
			Large	0.005	0.006	0.005	0.005	0.006	0.006
	40	0.2	Small	0.005	0.005	0.006	0.005	0.005	0.005
			Large	0.004	0.004	0.005	0.005	0.004	0.004
		0.5	Small	0.013	0.013	0.014	0.014	0.013	0.013
			Large	0.014	0.014	0.015	0.014	0.014	0.014
		0.8	Small	0.006	0.006	0.005	0.005	0.006	0.006
			Large	0.007	0.007	0.006	0.006	0.007	0.007

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 14. *Mean Bias, SE, RMSE of the Correlation between Ability and Speed Parameter Estimates*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	0.025	0.024	0.024	0.025	0.000	0.000	0.000	0.000	0.025	0.024	0.024	0.025
			Large	0.026	0.027	0.026	0.027	0.000	0.000	0.000	0.000	0.026	0.027	0.026	0.027
		0.5	Small	-0.010	-0.011	-0.010	-0.011	0.000	0.000	0.000	0.000	0.010	0.011	0.010	0.011
			Large	-0.080	-0.080	-0.080	-0.080	0.000	0.000	0.000	0.000	0.080	0.080	0.080	0.080
		0.8	Small	0.065	0.063	0.066	0.064	0.000	0.000	0.000	0.000	0.065	0.063	0.066	0.064
			Large	0.088	0.087	0.087	0.087	0.000	0.000	0.000	0.000	0.088	0.087	0.087	0.087
	40	0.2	Small	0.048	0.048	0.048	0.048	0.000	0.000	0.000	0.000	0.048	0.048	0.048	0.048
			Large	0.040	0.040	0.040	0.040	0.000	0.000	0.000	0.000	0.040	0.040	0.040	0.040
		0.5	Small	-0.013	-0.014	-0.013	-0.013	0.000	0.000	0.000	0.000	0.013	0.014	0.013	0.013
			Large	0.069	0.069	0.069	0.068	0.000	0.000	0.000	0.000	0.069	0.069	0.069	0.068
		0.8	Small	-0.004	-0.004	-0.004	-0.005	0.000	0.000	0.000	0.000	0.004	0.004	0.004	0.005
			Large	0.039	0.039	0.040	0.038	0.000	0.000	0.000	0.000	0.039	0.039	0.040	0.038
1000	20	0.2	Small	0.073	0.073	0.073	0.073	0.000	0.000	0.000	0.000	0.073	0.073	0.073	0.073
			Large	0.009	0.009	0.009	0.009	0.000	0.000	0.000	0.000	0.009	0.009	0.009	0.009
		0.5	Small	0.048	0.049	0.049	0.048	0.000	0.000	0.000	0.000	0.048	0.049	0.049	0.048
			Large	0.015	0.016	0.016	0.016	0.000	0.000	0.000	0.000	0.015	0.016	0.016	0.016
		0.8	Small	-0.038	-0.037	-0.037	-0.037	0.000	0.000	0.000	0.000	0.038	0.037	0.037	0.037
			Large	0.050	0.050	0.049	0.050	0.000	0.000	0.000	0.000	0.050	0.050	0.049	0.050
	40	0.2	Small	-0.001	-0.001	-0.001	-0.001	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001
			Large	0.007	0.007	0.008	0.007	0.000	0.000	0.000	0.000	0.007	0.007	0.008	0.007
		0.5	Small	0.010	0.010	0.010	0.010	0.000	0.000	0.000	0.000	0.010	0.010	0.010	0.010
			Large	0.005	0.005	0.005	0.004	0.000	0.000	0.000	0.000	0.005	0.005	0.005	0.004
		0.8	Small	0.016	0.015	0.014	0.016	0.000	0.000	0.000	0.000	0.016	0.015	0.014	0.016
			Large	0.016	0.017	0.017	0.015	0.000	0.000	0.000	0.000	0.016	0.017	0.017	0.015

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table A. 15. *Mean Bias of the Attribute-specific Slope and Intercept parameters.*

				Bias of $\hat{\gamma}_k$						Bias of $\hat{\lambda}_k$					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	-.082	-.078	-.045	-.045	-.077	-.076	.035	.031	.021	.024	.031	.030
			Large	-.046	-.046	-.029	-.030	-.046	-.048	.009	.007	.008	.008	.008	.007
		0.5	Small	.065	.063	.112	.114	.064	.064	.015	.013	.033	.034	.014	.014
			Large	.090	.092	.114	.115	.089	.091	-.115	-.117	-.121	-.119	-.115	-.115
		0.8	Small	-.030	-.028	.035	.038	-.031	-.030	.110	.109	.152	.154	.109	.109
			Large	.049	.051	.110	.110	.049	.049	.074	.074	.117	.117	.074	.075
	40	0.2	Small	.152	.149	.163	.164	.150	.148	.162	.161	.156	.157	.161	.160
			Large	-.119	-.119	-.107	-.107	-.120	-.115	.017	.016	.021	.022	.015	.018
		0.5	Small	.166	.166	.193	.193	.166	.165	-.100	-.100	-.088	-.088	-.100	-.100
			Large	.020	.021	.039	.039	.019	.024	-.033	-.033	-.048	-.047	-.034	-.032
		0.8	Small	.066	.066	.112	.113	.065	.066	-.054	-.054	-.068	-.068	-.054	-.055
			Large	.118	.119	.178	.179	.116	.122	.041	.042	.086	.087	.041	.044
1000	20	0.2	Small	.074	.073	.092	.092	.074	.073	.041	.040	.037	.038	.040	.040
			Large	-.031	-.032	-.021	-.021	-.030	-.031	.009	.009	.013	.013	.009	.009
		0.5	Small	.041	.041	.074	.075	.041	.041	.010	.009	.008	.009	.009	.010
			Large	.034	.035	.042	.042	.033	.034	.082	.081	.079	.079	.081	.082
		0.8	Small	.160	.159	.184	.184	.160	.159	-.007	-.007	-.011	-.010	-.007	-.007
			Large	.054	.055	.107	.108	.054	.054	.008	.007	.050	.050	.006	.007
	40	0.2	Small	.055	.056	.059	.060	.056	.055	.068	.067	.063	.064	.067	.067
			Large	-.016	-.015	-.011	-.011	-.016	-.014	.050	.049	.036	.036	.050	.051
		0.5	Small	.041	.041	.058	.059	.041	.041	.077	.077	.086	.086	.076	.076
			Large	-.028	-.028	-.010	-.009	-.029	-.026	.036	.036	.063	.063	.035	.036
		0.8	Small	-.111	-.111	-.087	-.086	-.109	-.111	.012	.012	-.001	-.001	.012	.012
			Large	.025	.024	.029	.029	.024	.027	.042	.042	.024	.024	.041	.043

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients. $\hat{\gamma}_k$ = Attribute-specific slope parameter; $\hat{\lambda}_k$ = Attribute-specific intercept parameter.

Table A. 16. *Mean SE of the Attribute-specific Slope and Intercept parameters.*

				SE of $\hat{\gamma}_k$						SE of $\hat{\lambda}_k$					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.216	.218	.263	.259	.220	.223	.081	.082	.090	.087	.081	.082
			Large	.203	.206	.234	.237	.204	.199	.065	.065	.072	.074	.065	.063
		0.5	Small	.134	.135	.231	.226	.132	.134	.074	.073	.095	.093	.073	.074
			Large	.129	.131	.205	.197	.128	.130	.047	.047	.056	.056	.047	.048
		0.8	Small	.101	.100	.235	.239	.102	.101	.090	.089	.112	.115	.090	.090
			Large	.109	.111	.209	.208	.109	.110	.072	.071	.086	.086	.071	.072
	40	0.2	Small	.095	.097	.108	.108	.095	.094	.044	.044	.046	.046	.043	.043
			Large	.068	.068	.086	.085	.069	.071	.025	.025	.027	.028	.025	.026
		0.5	Small	.059	.057	.082	.083	.059	.059	.025	.024	.027	.027	.025	.025
			Large	.066	.066	.112	.112	.066	.067	.029	.029	.037	.037	.030	.030
		0.8	Small	.051	.051	.096	.095	.048	.050	.026	.025	.038	.038	.025	.025
			Large	.091	.093	.217	.216	.095	.093	.062	.062	.073	.072	.063	.063
1000	20	0.2	Small	.141	.140	.178	.180	.140	.141	.064	.063	.070	.071	.064	.064
			Large	.108	.108	.129	.125	.111	.112	.037	.037	.040	.039	.037	.038
		0.5	Small	.097	.098	.159	.161	.097	.097	.052	.052	.063	.063	.052	.053
			Large	.069	.069	.103	.102	.069	.069	.037	.037	.042	.042	.037	.037
		0.8	Small	.097	.097	.157	.158	.098	.098	.059	.059	.066	.066	.059	.059
			Large	.073	.072	.189	.186	.073	.073	.057	.054	.086	.086	.053	.053
	40	0.2	Small	.063	.065	.069	.071	.063	.063	.026	.027	.028	.028	.027	.027
			Large	.048	.047	.052	.052	.049	.049	.017	.017	.017	.016	.017	.017
		0.5	Small	.046	.046	.060	.062	.047	.046	.023	.023	.025	.025	.023	.023
			Large	.034	.033	.051	.051	.034	.034	.017	.017	.022	.022	.017	.017
		0.8	Small	.045	.046	.070	.070	.045	.045	.017	.017	.022	.022	.017	.017
			Large	.036	.037	.052	.052	.037	.037	.018	.018	.021	.020	.018	.018

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients; $\hat{\gamma}_k$ = Attribute-specific slope parameter; $\hat{\lambda}_k$ = Attribute-specific intercept parameter.

Table A. 17. *Mean RMSE of the Attribute-specific Slope and Intercept parameters.*

				RMSE of $\hat{\gamma}_k$						RMSE of $\hat{\lambda}_k$					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.364	.373	.423	.417	.372	.375	.135	.138	.138	.135	.136	.136
			Large	.314	.313	.311	.309	.315	.307	.092	.091	.092	.093	.092	.090
		0.5	Small	.209	.207	.341	.343	.206	.208	.119	.118	.123	.123	.117	.118
			Large	.196	.198	.285	.281	.194	.199	.125	.127	.136	.133	.125	.125
		0.8	Small	.163	.161	.317	.324	.164	.162	.175	.175	.194	.198	.175	.176
			Large	.126	.128	.275	.273	.126	.127	.106	.105	.148	.147	.106	.106
	40	0.2	Small	.200	.197	.229	.230	.199	.197	.182	.180	.188	.189	.181	.180
			Large	.151	.152	.144	.144	.154	.153	.100	.100	.094	.094	.098	.101
		0.5	Small	.287	.287	.317	.317	.286	.287	.129	.129	.133	.133	.128	.129
			Large	.083	.084	.168	.168	.083	.086	.116	.115	.125	.125	.114	.116
		0.8	Small	.126	.127	.236	.238	.125	.126	.113	.113	.136	.137	.113	.113
			Large	.207	.208	.397	.397	.207	.210	.112	.113	.169	.168	.113	.114
1000	20	0.2	Small	.200	.200	.203	.204	.200	.200	.151	.150	.165	.166	.150	.151
			Large	.220	.217	.240	.237	.223	.222	.065	.067	.073	.070	.066	.068
		0.5	Small	.121	.122	.219	.220	.120	.121	.099	.099	.120	.121	.099	.100
			Large	.095	.095	.116	.115	.095	.095	.092	.091	.091	.091	.092	.092
		0.8	Small	.190	.189	.247	.247	.190	.189	.102	.101	.096	.097	.102	.102
			Large	.157	.156	.325	.321	.157	.156	.109	.109	.165	.164	.108	.109
	40	0.2	Small	.202	.202	.198	.199	.201	.201	.106	.106	.106	.107	.106	.105
			Large	.134	.134	.109	.109	.137	.134	.086	.085	.072	.073	.086	.086
		0.5	Small	.111	.112	.194	.196	.111	.111	.087	.087	.093	.093	.087	.087
			Large	.090	.090	.154	.154	.092	.090	.042	.042	.068	.068	.042	.043
		0.8	Small	.125	.125	.192	.193	.123	.125	.030	.030	.056	.056	.030	.030
			Large	.102	.102	.079	.081	.102	.104	.058	.058	.048	.048	.057	.058

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficien

Appendix B: Bias, SE and RMSE Results by the Simulated Conditions (Item Parameters)

Table B. 1. *Mean and Standard Deviation of Bias of the Item Intercept Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	-.011	.002	-.001	-.012	-.013	-.010	.067	.084	.080	.067	.058	.064
			Large	-.026	-.023	-.027	-.025	-.025	-.014	.055	.074	.077	.056	.060	.092
		0.5	Small	-.004	.009	.006	-.006	-.006	-.003	.049	.073	.071	.047	.046	.053
			Large	-.021	-.020	-.024	-.021	-.020	-.010	.057	.067	.066	.057	.061	.082
		0.8	Small	-.014	.001	-.005	-.017	-.016	-.013	.042	.066	.055	.038	.036	.042
			Large	-.009	-.005	-.008	-.007	-.008	.002	.061	.068	.071	.063	.067	.086
	40	0.2	Small	.010	.020	.020	.010	.011	.017	.154	.182	.193	.155	.167	.180
			Large	-.012	-.005	-.012	-.012	-.005	.017	.083	.083	.063	.083	.104	.110
		0.5	Small	.004	.013	.014	.004	.005	.011	.145	.175	.187	.146	.159	.176
			Large	-.011	-.004	-.010	-.011	-.004	.019	.079	.078	.067	.079	.119	.111
		0.8	Small	-.004	.004	.008	-.002	-.003	.004	.146	.167	.178	.145	.157	.170
			Large	.005	.014	.006	.003	.014	.036	.073	.081	.060	.073	.105	.112
1000	20	0.2	Small	-.031	-.021	-.021	-.028	-.032	-.030	.031	.045	.034	.028	.030	.035
			Large	-.007	-.007	-.009	-.007	-.008	-.002	.032	.045	.050	.033	.040	.057
		0.5	Small	-.025	-.017	-.021	-.028	-.027	-.025	.032	.044	.037	.032	.034	.038
			Large	.002	.004	.001	.002	.002	.008	.043	.044	.038	.044	.045	.050
		0.8	Small	-.009	-.001	-.001	-.008	-.010	-.009	.028	.048	.043	.029	.024	.030
			Large	.002	.004	-.002	-.001	.003	.009	.044	.051	.048	.041	.039	.057
	40	0.2	Small	.003	.008	.009	.003	.004	.008	.104	.122	.129	.104	.114	.122
			Large	-.007	-.002	-.004	-.006	-.002	.011	.054	.051	.043	.054	.068	.068
		0.5	Small	-.001	.005	.004	-.002	-.001	.004	.094	.115	.121	.095	.104	.114
			Large	.004	.009	.007	.004	.009	.022	.057	.053	.038	.058	.065	.065
		0.8	Small	-.002	.004	.005	-.001	-.001	.003	.099	.116	.125	.099	.110	.119
			Large	-.004	.001	-.001	-.004	.001	.013	.056	.054	.044	.056	.065	.073

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 2. *Mean and Standard Deviation of SE of the Item Intercept Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.185	.166	.168	.185	.185	.185	.053	.039	.039	.053	.054	.052
			Large	.161	.170	.178	.161	.158	.169	.041	.048	.057	.041	.041	.046
		0.5	Small	.188	.171	.173	.189	.187	.187	.050	.033	.036	.051	.050	.048
			Large	.178	.187	.197	.179	.176	.187	.048	.053	.063	.048	.047	.051
		0.8	Small	.196	.176	.180	.198	.196	.195	.063	.042	.045	.062	.062	.059
			Large	.168	.177	.184	.168	.166	.176	.046	.051	.058	.045	.045	.050
	40	0.2	Small	.194	.188	.191	.194	.193	.193	.063	.056	.058	.063	.062	.060
			Large	.150	.177	.196	.150	.164	.178	.039	.066	.087	.039	.051	.063
		0.5	Small	.193	.188	.191	.193	.192	.193	.049	.046	.048	.049	.048	.050
			Large	.152	.178	.194	.152	.163	.177	.041	.063	.081	.041	.049	.058
		0.8	Small	.199	.196	.199	.199	.198	.198	.074	.070	.074	.075	.074	.072
			Large	.152	.176	.193	.152	.163	.176	.043	.064	.083	.043	.054	.061
1000	20	0.2	Small	.150	.141	.142	.150	.149	.150	.055	.044	.045	.054	.055	.054
			Large	.131	.137	.141	.131	.130	.137	.049	.053	.057	.049	.048	.053
		0.5	Small	.145	.138	.139	.145	.145	.145	.056	.047	.048	.056	.057	.056
			Large	.131	.136	.140	.132	.130	.136	.041	.043	.047	.041	.040	.042
		0.8	Small	.145	.135	.138	.146	.144	.144	.051	.040	.042	.052	.051	.050
			Large	.134	.136	.141	.133	.131	.136	.041	.040	.044	.040	.037	.039
	40	0.2	Small	.152	.149	.151	.151	.151	.151	.064	.060	.062	.064	.064	.062
			Large	.115	.129	.137	.115	.123	.130	.036	.050	.059	.036	.043	.049
		0.5	Small	.152	.151	.153	.153	.152	.152	.061	.058	.060	.061	.061	.058
			Large	.119	.134	.142	.119	.127	.135	.045	.062	.072	.045	.054	.061
		0.8	Small	.152	.149	.151	.152	.151	.151	.064	.061	.062	.064	.063	.063
			Large	.122	.137	.144	.122	.130	.137	.038	.052	.060	.037	.044	.051

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 3. *Mean and Standard Deviation of RMSE of the Item Intercept Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.195	.184	.184	.195	.193	.194	.060	.046	.045	.059	.056	.057
			Large	.171	.184	.192	.171	.170	.188	.044	.054	.066	.045	.045	.059
		0.5	Small	.193	.183	.185	.194	.192	.192	.055	.044	.046	.056	.055	.053
			Large	.188	.198	.207	.188	.187	.201	.050	.058	.069	.051	.049	.061
		0.8	Small	.201	.186	.188	.201	.200	.199	.065	.046	.045	.064	.062	.061
			Large	.179	.188	.196	.178	.178	.193	.047	.057	.062	.047	.048	.059
	40	0.2	Small	.221	.222	.226	.221	.223	.225	.129	.150	.161	.129	.139	.150
			Large	.167	.191	.203	.166	.185	.199	.057	.079	.092	.057	.074	.092
		0.5	Small	.220	.224	.228	.220	.222	.228	.110	.134	.148	.111	.123	.137
			Large	.169	.191	.203	.169	.191	.200	.050	.070	.087	.050	.079	.085
		0.8	Small	.225	.227	.231	.225	.227	.229	.123	.139	.153	.123	.133	.143
			Large	.165	.189	.199	.164	.187	.200	.055	.078	.088	.055	.077	.092
1000	20	0.2	Small	.156	.148	.148	.155	.155	.155	.056	.047	.045	.055	.057	.056
			Large	.136	.143	.148	.136	.136	.146	.048	.055	.059	.048	.048	.057
		0.5	Small	.150	.145	.145	.151	.151	.151	.057	.049	.048	.057	.059	.058
			Large	.138	.141	.145	.139	.137	.143	.041	.046	.048	.041	.041	.047
		0.8	Small	.148	.142	.144	.149	.147	.147	.050	.042	.042	.051	.050	.049
			Large	.140	.143	.147	.139	.136	.144	.043	.047	.051	.041	.038	.050
	40	0.2	Small	.166	.167	.170	.165	.167	.169	.102	.113	.120	.102	.109	.115
			Large	.125	.137	.142	.125	.136	.141	.044	.055	.062	.044	.056	.062
		0.5	Small	.166	.168	.171	.167	.168	.169	.090	.104	.110	.090	.098	.103
			Large	.128	.141	.146	.128	.139	.145	.055	.070	.075	.055	.065	.075
		0.8	Small	.165	.166	.169	.165	.167	.168	.097	.108	.117	.097	.104	.112
			Large	.131	.144	.149	.131	.140	.149	.048	.060	.065	.047	.057	.068

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 4. *Mean and Standard Deviation of Bias of the Item Interaction Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.036	.001	.011	.036	.023	.023	.231	.247	.236	.232	.186	.198
			Large	.061	.019	.022	.062	.047	.050	.154	.190	.174	.155	.147	.188
		0.5	Small	.044	.013	.027	.047	.038	.036	.139	.188	.160	.140	.114	.112
			Large	.079	.041	.044	.080	.068	.065	.103	.138	.133	.105	.123	.121
		0.8	Small	.053	.015	.032	.060	.042	.040	.160	.197	.163	.160	.110	.134
			Large	.073	.033	.034	.073	.059	.060	.149	.196	.185	.149	.165	.179
	40	0.2	Small	-.015	-.028	-.024	-.016	-.018	-.019	.269	.293	.298	.269	.282	.278
			Large	.024	.009	.014	.024	.037	.048	.154	.161	.149	.154	.159	.208
		0.5	Small	-.010	-.024	-.020	-.010	-.013	-.013	.255	.278	.280	.256	.267	.263
			Large	.019	.004	.009	.019	.030	.036	.152	.154	.151	.152	.162	.176
		0.8	Small	-.003	-.017	-.016	-.005	-.006	-.007	.263	.284	.282	.265	.272	.268
			Large	.002	-.011	-.005	.003	.011	.020	.139	.152	.133	.139	.171	.184
1000	20	0.2	Small	.018	.003	.007	.018	.015	.015	.104	.131	.099	.104	.086	.096
			Large	.033	.008	.008	.033	.025	.025	.089	.096	.089	.089	.092	.089
		0.5	Small	.060	.043	.049	.061	.056	.056	.088	.117	.092	.088	.064	.077
			Large	.014	-.010	-.010	.013	.005	.005	.084	.104	.094	.086	.094	.079
		0.8	Small	.006	-.009	.000	.010	.003	.003	.099	.136	.110	.098	.079	.094
			Large	.022	-.002	.000	.020	.012	.011	.097	.104	.096	.095	.102	.097
	40	0.2	Small	-.016	-.025	-.022	-.017	-.017	-.018	.185	.197	.200	.185	.196	.187
			Large	.018	.011	.013	.018	.026	.030	.090	.092	.086	.091	.103	.126
		0.5	Small	-.007	-.015	-.010	-.005	-.008	-.008	.192	.204	.202	.191	.200	.193
			Large	.002	-.006	-.004	.002	.010	.013	.102	.100	.089	.102	.119	.112
		0.8	Small	-.009	-.018	-.015	-.010	-.010	-.011	.185	.194	.196	.186	.197	.183
			Large	.015	.008	.010	.015	.024	.027	.097	.087	.077	.098	.088	.117

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 5. *Mean and Standard Deviation of SE of the Item Interaction Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.379	.341	.351	.378	.380	.381	.088	.090	.092	.089	.088	.088
			Large	.277	.286	.302	.277	.272	.316	.057	.059	.062	.057	.056	.073
		0.5	Small	.366	.329	.339	.367	.366	.362	.107	.080	.089	.109	.097	.093
			Large	.270	.282	.297	.270	.266	.304	.049	.038	.044	.049	.047	.059
		0.8	Small	.386	.339	.357	.387	.385	.383	.088	.077	.083	.088	.081	.081
			Large	.278	.293	.306	.276	.271	.320	.044	.049	.049	.045	.040	.059
	40	0.2	Small	.335	.327	.331	.334	.333	.334	.072	.080	.079	.072	.072	.075
			Large	.249	.295	.318	.249	.290	.342	.046	.059	.071	.045	.062	.087
		0.5	Small	.321	.312	.318	.321	.320	.323	.076	.076	.077	.076	.077	.077
			Large	.249	.294	.314	.249	.288	.338	.047	.056	.065	.047	.066	.093
		0.8	Small	.326	.323	.328	.326	.325	.327	.069	.072	.071	.068	.068	.069
			Large	.242	.284	.306	.241	.277	.328	.049	.060	.069	.049	.064	.090
1000	20	0.2	Small	.283	.267	.272	.283	.282	.282	.074	.068	.069	.074	.073	.074
			Large	.233	.243	.251	.233	.231	.256	.047	.047	.052	.047	.044	.065
		0.5	Small	.298	.279	.284	.298	.295	.296	.081	.077	.077	.082	.078	.080
			Large	.232	.242	.250	.233	.230	.254	.037	.037	.038	.036	.036	.045
		0.8	Small	.278	.258	.266	.281	.277	.277	.066	.060	.064	.069	.064	.065
			Large	.213	.218	.227	.212	.208	.232	.041	.039	.039	.040	.037	.051
	40	0.2	Small	.260	.257	.261	.259	.260	.260	.060	.059	.062	.060	.063	.059
			Large	.205	.229	.239	.205	.226	.250	.034	.043	.050	.034	.044	.059
		0.5	Small	.254	.252	.256	.254	.255	.255	.054	.054	.055	.054	.055	.054
			Large	.201	.227	.238	.201	.224	.250	.043	.056	.063	.044	.056	.069
		0.8	Small	.273	.270	.274	.273	.273	.274	.078	.079	.082	.078	.081	.080
			Large	.204	.232	.241	.204	.229	.252	.038	.047	.052	.037	.049	.061

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 6. *Mean and Standard Deviation of RMSE of the Item Interaction Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.432	.404	.407	.432	.417	.422	.130	.141	.138	.130	.109	.112
			Large	.315	.330	.341	.315	.307	.361	.088	.105	.090	.089	.076	.107
		0.5	Small	.392	.373	.373	.394	.386	.382	.111	.094	.094	.113	.093	.086
			Large	.296	.312	.326	.297	.296	.334	.064	.057	.049	.065	.064	.052
		0.8	Small	.415	.381	.388	.416	.401	.404	.111	.112	.101	.110	.087	.091
			Large	.318	.347	.355	.317	.314	.368	.068	.074	.064	.067	.076	.070
	40	0.2	Small	.401	.403	.408	.400	.402	.405	.167	.190	.194	.167	.182	.171
			Large	.287	.332	.348	.287	.328	.393	.074	.076	.085	.074	.077	.121
		0.5	Small	.380	.383	.388	.380	.382	.387	.168	.183	.185	.168	.179	.168
			Large	.288	.329	.346	.288	.327	.377	.066	.067	.075	.066	.083	.109
		0.8	Small	.391	.396	.399	.391	.392	.395	.161	.180	.180	.163	.172	.160
			Large	.275	.319	.331	.275	.321	.369	.064	.074	.078	.064	.081	.111
1000	20	0.2	Small	.298	.293	.287	.299	.294	.296	.084	.082	.077	.084	.076	.081
			Large	.250	.260	.265	.250	.248	.271	.055	.054	.055	.054	.047	.066
		0.5	Small	.313	.301	.300	.314	.307	.309	.091	.090	.083	.091	.079	.085
			Large	.246	.260	.265	.247	.247	.266	.038	.053	.048	.038	.040	.046
		0.8	Small	.293	.288	.287	.295	.288	.291	.073	.071	.066	.076	.064	.068
			Large	.233	.239	.245	.232	.229	.251	.049	.048	.046	.047	.047	.051
	40	0.2	Small	.296	.298	.302	.295	.299	.298	.132	.140	.144	.132	.142	.130
			Large	.223	.246	.253	.223	.248	.276	.042	.049	.056	.042	.053	.079
		0.5	Small	.296	.298	.300	.296	.298	.298	.127	.137	.136	.126	.136	.125
			Large	.223	.247	.253	.223	.251	.271	.052	.060	.066	.052	.064	.077
		0.8	Small	.308	.308	.312	.308	.311	.308	.140	.147	.151	.140	.149	.139
			Large	.224	.246	.252	.224	.245	.275	.048	.053	.056	.048	.056	.077

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 7. *Mean and Standard Deviation of Bias of the Item Intensity Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	-.003	-.003	-.004	-.003	-.003	-.003	.004	.004	.004	.004	.004	.004
			Large	-.003	-.003	-.004	-.003	-.003	-.003	.004	.004	.004	.004	.004	.004
		0.5	Small	-.001	-.001	-.002	-.001	-.001	-.001	.004	.004	.004	.004	.004	.004
			Large	-.002	-.002	-.002	-.002	-.002	-.002	.003	.003	.003	.003	.003	.003
		0.8	Small	.005	.005	.005	.005	.005	.005	.004	.004	.004	.004	.004	.004
			Large	.007	.007	.006	.007	.007	.007	.005	.005	.005	.005	.005	.005
	40	0.2	Small	-.004	-.004	-.004	-.004	-.004	-.004	.005	.005	.005	.005	.005	.005
			Large	-.004	-.004	-.004	-.004	-.004	-.004	.004	.004	.004	.004	.004	.004
		0.5	Small	-.001	-.001	-.001	-.002	-.001	-.001	.005	.005	.005	.005	.005	.005
			Large	-.001	-.001	-.001	-.001	-.001	-.001	.004	.004	.004	.004	.004	.004
		0.8	Small	.006	.006	.005	.005	.006	.006	.005	.005	.005	.005	.005	.005
			Large	.006	.007	.007	.007	.007	.007	.004	.004	.004	.004	.004	.004
1000	20	0.2	Small	-.006	-.006	-.007	-.006	-.006	-.006	.003	.003	.003	.003	.003	.003
			Large	-.007	-.007	-.007	-.007	-.007	-.007	.003	.003	.003	.003	.003	.003
		0.5	Small	-.005	-.005	-.005	-.005	-.005	-.005	.003	.003	.003	.003	.003	.003
			Large	-.005	-.005	-.006	-.005	-.005	-.005	.002	.002	.002	.002	.002	.002
		0.8	Small	.005	.005	.004	.005	.005	.005	.003	.003	.003	.003	.003	.003
			Large	.005	.005	.005	.005	.005	.005	.005	.004	.004	.004	.004	.004
	40	0.2	Small	-.006	-.006	-.007	-.007	-.006	-.006	.003	.003	.003	.003	.003	.003
			Large	-.006	-.006	-.007	-.006	-.006	-.006	.003	.003	.003	.003	.003	.003
		0.5	Small	-.005	-.005	-.006	-.005	-.005	-.005	.003	.003	.003	.003	.003	.003
			Large	-.005	-.005	-.005	-.005	-.005	-.005	.003	.003	.003	.003	.003	.003
		0.8	Small	.005	.005	.004	.005	.005	.005	.003	.003	.003	.003	.003	.003
			Large	.004	.004	.004	.004	.004	.004	.003	.003	.003	.003	.003	.003

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 8. *Mean and Standard Deviation of SE of the Item Intensity Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.021	.021	.021	.021	.021	.021	.002	.002	.002	.002	.002	.002
			Large	.022	.022	.022	.022	.022	.022	.003	.003	.003	.003	.003	.003
		0.5	Small	.021	.021	.021	.021	.021	.021	.003	.003	.003	.003	.003	.003
			Large	.022	.022	.022	.022	.022	.022	.004	.004	.004	.004	.004	.004
		0.8	Small	.021	.021	.021	.021	.021	.021	.003	.003	.003	.003	.003	.003
			Large	.021	.021	.021	.021	.021	.021	.003	.003	.003	.003	.003	.003
	40	0.2	Small	.022	.022	.022	.022	.022	.022	.002	.002	.002	.002	.002	.002
			Large	.021	.021	.021	.021	.021	.021	.003	.003	.003	.003	.003	.003
		0.5	Small	.022	.022	.022	.022	.022	.022	.003	.003	.003	.003	.003	.003
			Large	.021	.021	.021	.021	.021	.021	.003	.003	.003	.003	.003	.003
		0.8	Small	.023	.023	.023	.023	.023	.023	.003	.003	.003	.003	.003	.003
			Large	.022	.022	.022	.022	.022	.022	.003	.003	.003	.003	.003	.003
1000	20	0.2	Small	.015	.015	.015	.015	.015	.015	.002	.002	.002	.002	.002	.002
			Large	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002
		0.5	Small	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002
			Large	.015	.015	.015	.015	.015	.015	.002	.002	.002	.002	.002	.002
		0.8	Small	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002
			Large	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002
	40	0.2	Small	.015	.015	.015	.015	.015	.015	.002	.002	.002	.002	.002	.002
			Large	.015	.015	.015	.015	.015	.015	.002	.002	.002	.002	.002	.002
		0.5	Small	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002
			Large	.015	.015	.015	.015	.015	.015	.002	.002	.002	.002	.002	.002
		0.8	Small	.015	.015	.015	.015	.015	.015	.002	.002	.002	.002	.002	.002
			Large	.015	.015	.015	.015	.015	.015	.002	.002	.002	.002	.002	.002

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 9. *Mean and Standard Deviation of RMSE of the Item Intensity Parameter Estimates.*

				Mean						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA_LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.022	.022	.022	.022	.022	.022	.002	.002	.002	.002	.002	.002
			Large	.023	.023	.023	.023	.023	.023	.003	.003	.003	.003	.003	.003
		0.5	Small	.021	.021	.021	.021	.021	.021	.003	.003	.003	.003	.003	.003
			Large	.022	.022	.022	.022	.022	.022	.004	.004	.004	.004	.004	.004
		0.8	Small	.022	.022	.022	.022	.022	.022	.003	.004	.004	.003	.003	.004
			Large	.023	.023	.023	.023	.023	.023	.003	.003	.003	.003	.003	.003
	40	0.2	Small	.023	.023	.023	.023	.023	.023	.003	.003	.003	.003	.003	.003
			Large	.022	.022	.022	.022	.022	.022	.003	.003	.003	.003	.003	.003
		0.5	Small	.023	.023	.023	.023	.023	.023	.003	.003	.003	.003	.003	.003
			Large	.022	.022	.022	.022	.022	.022	.003	.003	.003	.003	.003	.003
		0.8	Small	.024	.024	.024	.024	.024	.024	.003	.003	.003	.003	.003	.003
			Large	.023	.023	.023	.023	.023	.023	.003	.003	.003	.003	.003	.003
1000	20	0.2	Small	.017	.017	.017	.017	.017	.017	.003	.003	.003	.003	.003	.003
			Large	.017	.017	.017	.017	.017	.017	.002	.002	.002	.002	.002	.002
		0.5	Small	.017	.017	.017	.017	.017	.017	.002	.002	.002	.002	.002	.002
			Large	.016	.016	.016	.016	.016	.016	.001	.002	.001	.001	.001	.001
		0.8	Small	.017	.017	.017	.017	.017	.017	.002	.002	.002	.002	.002	.002
			Large	.018	.018	.018	.018	.018	.018	.002	.002	.002	.002	.002	.002
	40	0.2	Small	.017	.017	.017	.017	.017	.017	.002	.002	.002	.002	.002	.002
			Large	.017	.017	.017	.017	.017	.017	.002	.002	.002	.002	.002	.002
		0.5	Small	.017	.017	.017	.017	.017	.017	.002	.002	.002	.002	.002	.002
			Large	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002
		0.8	Small	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002
			Large	.016	.016	.016	.016	.016	.016	.002	.002	.002	.002	.002	.002

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 10. *Mean Bias of the Response Time Variance Parameter Estimates.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.001	.001	.001	.001	.001	.001
			Large	.000	.000	.000	.000	.000	.000
		0.5	Small	.000	.000	.000	.000	.000	.000
			Large	.002	.002	.002	.002	.002	.002
		0.8	Small	.001	.001	.001	.001	.001	.001
			Large	.001	.001	.001	.001	.001	.001
	40	0.2	Small	.000	.000	.000	.000	.000	.000
			Large	.000	.000	.000	.000	.000	.000
		0.5	Small	.001	.001	.001	.001	.001	.001
			Large	.000	.000	.000	.000	.000	.000
		0.8	Small	-.001	-.001	-.001	-.001	-.001	-.001
			Large	.000	.000	.000	.000	.000	.000
1000	20	0.2	Small	.000	.000	.001	.000	.000	.000
			Large	.000	.000	.000	.000	.000	.000
		0.5	Small	.000	.000	.000	.000	.000	.000
			Large	-.001	-.001	-.001	-.001	-.001	-.001
		0.8	Small	.000	.000	.000	.000	.000	.000
			Large	.000	.000	.000	.000	.000	.000
	40	0.2	Small	.000	.000	.000	.000	.000	.000
			Large	.000	.000	.000	.000	.000	.000
		0.5	Small	.000	.000	.000	.000	.000	.000
			Large	.000	.000	.000	.000	.000	.000
		0.8	Small	.000	.000	.000	.000	.000	.000
			Large	.000	.000	.000	.000	.000	.000

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 11. *Mean SE of the Response Time Variance Parameter Estimates.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.004	.004	.004	.004	.004	.004
			Large	.004	.004	.004	.004	.004	.004
		0.5	Small	.004	.004	.004	.004	.004	.004
			Large	.004	.004	.004	.004	.004	.004
		0.8	Small	.004	.004	.004	.004	.004	.004
			Large	.004	.004	.004	.004	.004	.004
	40	0.2	Small	.002	.002	.002	.002	.002	.002
			Large	.003	.003	.003	.003	.003	.003
		0.5	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002
		0.8	Small	.002	.002	.002	.002	.002	.002
			Large	.003	.003	.003	.003	.003	.003
1000	20	0.2	Small	.003	.003	.003	.003	.003	.003
			Large	.002	.002	.002	.002	.002	.002
		0.5	Small	.003	.003	.003	.003	.003	.003
			Large	.003	.003	.003	.003	.003	.003
		0.8	Small	.002	.002	.002	.002	.002	.002
			Large	.003	.003	.003	.003	.003	.003
	40	0.2	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002
		0.5	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002
		0.8	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 12. *Mean RMSE of the Response Time Variance Parameter Estimates.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.004	.004	.004	.004	.004	.004
			Large	.004	.004	.004	.004	.004	.004
		0.5	Small	.004	.004	.004	.004	.004	.004
			Large	.005	.005	.005	.005	.005	.005
		0.8	Small	.004	.004	.004	.004	.004	.004
			Large	.004	.004	.004	.004	.004	.004
	40	0.2	Small	.002	.002	.002	.002	.002	.002
			Large	.003	.003	.003	.003	.003	.003
		0.5	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002
		0.8	Small	.002	.002	.002	.002	.002	.002
			Large	.003	.003	.003	.003	.003	.003
1000	20	0.2	Small	.003	.003	.003	.003	.003	.003
			Large	.002	.002	.002	.002	.002	.002
		0.5	Small	.003	.003	.003	.003	.003	.003
			Large	.003	.003	.003	.003	.003	.003
		0.8	Small	.002	.002	.002	.002	.002	.002
			Large	.003	.003	.003	.003	.003	.003
	40	0.2	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002
		0.5	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002
		0.8	Small	.002	.002	.002	.002	.002	.002
			Large	.002	.002	.002	.002	.002	.002

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 13. *Mean Bias of the Two Regression Coefficients for the Item Intercept Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.492	.527	.523	.492	.475	.514	.534	.544	.514	.476
			Large	.138	.202	.214	.137	.159	.664	.657	.693	.664	.911
		0.5	Small	.511	.541	.539	.514	.489	.538	.545	.556	.544	.493
			Large	.130	.192	.204	.131	.152	.677	.662	.702	.677	.928
		0.8	Small	.499	.527	.525	.500	.476	.548	.556	.568	.549	.509
			Large	.139	.199	.214	.140	.157	.689	.675	.717	.697	.935
	40	0.2	Small	.044	.070	.077	.044	.049	.278	.291	.306	.278	.354
			Large	-.103	-.081	-.079	-.103	-.059	-.152	-.103	-.103	-.152	.077
		0.5	Small	.047	.073	.082	.047	.052	.271	.286	.301	.270	.345
			Large	-.101	-.078	-.076	-.101	-.059	-.125	-.076	-.072	-.124	.106
		0.8	Small	.027	.053	.059	.026	.031	.259	.274	.293	.262	.338
			Large	-.093	-.069	-.068	-.093	-.047	-.126	-.077	-.077	-.124	.102
1000	20	0.2	Small	.514	.532	.527	.514	.493	.541	.540	.549	.543	.496
			Large	.159	.193	.200	.159	.174	.701	.689	.712	.701	.936
		0.5	Small	.506	.522	.517	.505	.483	.556	.551	.554	.551	.513
			Large	.146	.181	.184	.146	.165	.679	.668	.689	.680	.920
		0.8	Small	.507	.523	.519	.506	.485	.544	.543	.553	.545	.500
			Large	.137	.177	.184	.142	.166	.647	.641	.667	.654	.893
	40	0.2	Small	.036	.050	.055	.035	.041	.267	.274	.283	.267	.349
			Large	-.098	-.085	-.084	-.098	-.055	-.126	-.097	-.096	-.126	.098
		0.5	Small	.035	.050	.054	.035	.041	.256	.265	.275	.256	.339
			Large	-.101	-.088	-.087	-.101	-.058	-.130	-.104	-.102	-.130	.093
		0.8	Small	.033	.049	.053	.033	.040	.260	.267	.276	.259	.341
			Large	-.095	-.083	-.081	-.095	-.051	-.136	-.109	-.107	-.135	.086

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 14. *Mean Bias of the Two Regression Coefficients for the Item Interaction Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.200	.053	.059	.200	.046	.482	.308	.280	.480	.241
			Large	.196	.031	.012	.197	.111	-.265	-.335	-.400	-.262	-.614
		0.5	Small	.165	.042	.046	.165	.036	.362	.246	.215	.361	.137
			Large	.230	.079	.063	.231	.154	-.336	-.380	-.444	-.337	-.715
		0.8	Small	.165	.035	.041	.170	.022	.446	.297	.263	.446	.218
			Large	.198	.035	.018	.201	.123	-.333	-.386	-.455	-.336	-.692
	40	0.2	Small	-.191	-.239	-.248	-.191	-.212	.116	.050	.027	.115	.079
			Large	.044	.012	.003	.044	-.005	.281	.172	.183	.281	.065
		0.5	Small	-.182	-.227	-.237	-.182	-.204	.129	.074	.045	.130	.090
			Large	.056	.023	.015	.056	.009	.261	.163	.166	.263	.040
		0.8	Small	-.172	-.216	-.225	-.171	-.193	.151	.096	.065	.151	.112
			Large	.047	.011	.003	.047	-.004	.220	.122	.130	.222	-.008
1000	20	0.2	Small	.133	.067	.070	.133	.012	.345	.289	.264	.344	.140
			Large	.205	.113	.103	.204	.131	-.338	-.369	-.405	-.338	-.699
		0.5	Small	.151	.086	.086	.151	.031	.358	.303	.278	.361	.148
			Large	.203	.114	.106	.205	.130	-.346	-.369	-.407	-.343	-.711
		0.8	Small	.120	.058	.061	.123	-.002	.355	.299	.278	.355	.154
			Large	.207	.111	.101	.199	.124	-.327	-.355	-.394	-.333	-.697
	40	0.2	Small	-.180	-.205	-.211	-.179	-.202	.126	.096	.078	.126	.086
			Large	.028	.008	.004	.028	-.018	.213	.154	.158	.214	.006
		0.5	Small	-.176	-.202	-.208	-.175	-.198	.141	.108	.090	.140	.097
			Large	.041	.022	.017	.042	-.003	.196	.139	.141	.195	-.021
		0.8	Small	-.160	-.187	-.193	-.160	-.182	.121	.092	.077	.124	.076
			Large	.037	.018	.012	.037	-.011	.246	.184	.187	.244	.034

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 15. *Mean Bias of the Two Regression Coefficients for the Item Intensity Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	-.089	-.090	-.095	-.089	-.146	.172	.170	.170	.172	.131
			Large	-.032	-.034	-.038	-.032	-.090	-.026	-.028	-.028	-.027	-.162
		0.5	Small	-.089	-.090	-.095	-.089	-.146	.173	.172	.171	.173	.132
			Large	-.029	-.031	-.034	-.029	-.087	-.026	-.027	-.028	-.025	-.162
		0.8	Small	-.089	-.091	-.095	-.089	-.146	.172	.170	.170	.170	.131
			Large	-.028	-.029	-.034	-.028	-.085	-.030	-.031	-.032	-.030	-.166
	40	0.2	Small	-.038	-.039	-.040	-.038	-.054	.142	.142	.140	.143	.104
			Large	.017	.017	.015	.017	.002	-.068	-.067	-.069	-.067	-.166
		0.5	Small	-.038	-.039	-.040	-.038	-.054	.141	.140	.139	.141	.103
			Large	.016	.015	.014	.016	.001	-.066	-.066	-.068	-.066	-.164
		0.8	Small	-.037	-.038	-.039	-.037	-.053	.141	.141	.139	.142	.102
			Large	.017	.016	.015	.017	.002	-.068	-.067	-.069	-.067	-.165
1000	20	0.2	Small	-.092	-.094	-.098	-.091	-.150	.171	.172	.170	.171	.132
			Large	-.029	-.030	-.034	-.029	-.087	-.023	-.024	-.025	-.024	-.159
		0.5	Small	-.089	-.091	-.096	-.090	-.147	.171	.170	.169	.170	.130
			Large	-.028	-.029	-.033	-.028	-.086	-.025	-.026	-.026	-.025	-.162
		0.8	Small	-.090	-.092	-.097	-.090	-.148	.173	.172	.171	.172	.132
			Large	-.027	-.027	-.031	-.026	-.084	-.028	-.028	-.028	-.026	-.165
	40	0.2	Small	-.037	-.038	-.040	-.038	-.053	.143	.142	.142	.143	.105
			Large	.017	.016	.015	.017	.002	-.068	-.068	-.069	-.068	-.166
		0.5	Small	-.038	-.039	-.040	-.038	-.054	.143	.142	.141	.143	.105
			Large	.017	.016	.015	.016	.001	-.065	-.065	-.066	-.065	-.162
		0.8	Small	-.038	-.039	-.040	-.037	-.054	.143	.142	.141	.143	.104
			Large	.017	.016	.015	.017	.002	-.065	-.066	-.067	-.066	-.164

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 16. *Mean SE of the Two Regression Coefficients for the Item Intercept Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.048	.043	.043	.048	.051	.107	.085	.085	.107	.108
			Large	.049	.049	.050	.049	.049	.084	.078	.079	.084	.085
		0.5	Small	.061	.048	.048	.061	.058	.085	.071	.068	.083	.077
			Large	.066	.064	.066	.065	.063	.091	.085	.085	.090	.086
		0.8	Small	.066	.053	.054	.066	.071	.104	.091	.091	.107	.109
			Large	.054	.053	.057	.055	.052	.071	.066	.067	.069	.069
	40	0.2	Small	.041	.036	.037	.041	.041	.073	.066	.066	.073	.075
			Large	.031	.034	.038	.031	.032	.056	.058	.063	.056	.055
		0.5	Small	.026	.022	.022	.025	.025	.079	.069	.069	.079	.078
			Large	.033	.035	.037	.033	.033	.051	.050	.056	.051	.050
		0.8	Small	.032	.032	.032	.033	.031	.088	.081	.084	.088	.086
			Large	.026	.028	.031	.026	.028	.062	.060	.068	.062	.064
1000	20	0.2	Small	.047	.039	.040	.045	.043	.089	.078	.079	.090	.084
			Large	.049	.049	.050	.049	.049	.053	.051	.052	.053	.056
		0.5	Small	.050	.044	.045	.049	.051	.072	.066	.063	.071	.075
			Large	.044	.045	.046	.044	.041	.077	.074	.077	.078	.074
		0.8	Small	.048	.041	.042	.047	.046	.078	.071	.073	.082	.076
			Large	.054	.050	.052	.051	.048	.074	.074	.074	.075	.073
	40	0.2	Small	.021	.019	.020	.021	.021	.051	.048	.048	.052	.052
			Large	.022	.024	.025	.022	.023	.049	.048	.051	.049	.049
		0.5	Small	.028	.027	.028	.029	.029	.040	.039	.040	.039	.038
			Large	.020	.021	.022	.020	.021	.039	.040	.044	.039	.041
		0.8	Small	.023	.022	.022	.023	.024	.057	.054	.054	.056	.058
			Large	.018	.020	.021	.018	.019	.048	.049	.053	.048	.049

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 17. *Mean SE of the Two Regression Coefficients for the Item Interaction Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.097	.075	.077	.096	.100	.257	.193	.202	.254	.250
			Large	.128	.128	.133	.129	.130	.172	.130	.137	.172	.186
		0.5	Small	.164	.115	.126	.166	.152	.204	.162	.156	.200	.184
			Large	.103	.101	.104	.102	.105	.132	.102	.107	.131	.137
		0.8	Small	.172	.124	.130	.174	.163	.212	.163	.157	.212	.197
			Large	.114	.118	.120	.113	.111	.172	.136	.134	.172	.175
	40	0.2	Small	.071	.062	.064	.071	.071	.150	.134	.127	.152	.150
			Large	.058	.061	.063	.058	.061	.093	.084	.089	.092	.102
		0.5	Small	.043	.037	.038	.043	.042	.142	.116	.121	.143	.143
			Large	.065	.064	.066	.065	.065	.110	.101	.102	.111	.121
		0.8	Small	.049	.048	.049	.050	.048	.113	.091	.095	.112	.109
			Large	.051	.054	.055	.050	.053	.095	.084	.090	.098	.102
1000	20	0.2	Small	.113	.094	.099	.112	.102	.160	.137	.136	.159	.140
			Large	.089	.086	.088	.089	.082	.132	.113	.115	.132	.127
		0.5	Small	.099	.084	.087	.099	.097	.192	.162	.165	.193	.181
			Large	.078	.077	.080	.078	.074	.125	.110	.116	.126	.123
		0.8	Small	.090	.076	.076	.090	.086	.142	.121	.127	.148	.134
			Large	.057	.057	.060	.057	.055	.144	.129	.131	.145	.145
	40	0.2	Small	.034	.032	.032	.033	.033	.093	.084	.085	.096	.093
			Large	.033	.035	.036	.033	.033	.100	.091	.096	.099	.097
		0.5	Small	.039	.038	.039	.040	.039	.081	.075	.076	.080	.079
			Large	.032	.032	.033	.032	.034	.074	.064	.067	.073	.082
		0.8	Small	.035	.033	.034	.035	.034	.122	.112	.110	.123	.120
			Large	.033	.034	.035	.033	.036	.084	.079	.082	.085	.095

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 18. *Mean SE of the Two Regression Coefficients for the Item Intensity Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.008	.007	.008	.008	.008	.008	.008	.008	.008	.009
			Large	.006	.006	.006	.006	.005	.010	.011	.010	.010	.009
		0.5	Small	.006	.006	.005	.006	.006	.009	.009	.008	.009	.010
			Large	.006	.006	.006	.006	.006	.013	.014	.013	.013	.013
		0.8	Small	.005	.006	.005	.006	.005	.012	.012	.012	.012	.012
			Large	.006	.007	.006	.007	.006	.010	.011	.010	.011	.010
	40	0.2	Small	.003	.004	.003	.003	.003	.006	.006	.006	.007	.006
			Large	.003	.003	.003	.003	.003	.007	.007	.007	.007	.007
		0.5	Small	.003	.003	.003	.003	.003	.009	.009	.009	.009	.009
			Large	.003	.003	.003	.003	.003	.009	.009	.009	.009	.008
		0.8	Small	.003	.003	.003	.003	.003	.007	.007	.007	.007	.007
			Large	.004	.004	.004	.004	.004	.006	.007	.007	.006	.007
1000	20	0.2	Small	.005	.006	.005	.006	.005	.008	.009	.008	.007	.007
			Large	.005	.005	.005	.005	.005	.007	.007	.007	.008	.007
		0.5	Small	.005	.005	.005	.005	.005	.006	.006	.006	.007	.007
			Large	.004	.004	.004	.004	.004	.008	.008	.008	.008	.008
		0.8	Small	.006	.005	.005	.006	.005	.008	.009	.008	.008	.008
			Large	.005	.005	.005	.005	.004	.010	.011	.010	.010	.010
	40	0.2	Small	.002	.003	.002	.002	.002	.006	.006	.006	.006	.006
			Large	.002	.002	.002	.002	.002	.005	.005	.005	.005	.005
		0.5	Small	.002	.002	.002	.002	.002	.006	.005	.005	.005	.005
			Large	.002	.002	.002	.002	.002	.005	.005	.005	.005	.005
		0.8	Small	.002	.003	.002	.002	.002	.005	.006	.005	.005	.005
			Large	.003	.003	.003	.003	.003	.005	.005	.005	.004	.004

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 19. *Mean RMSE of the Two Regression Coefficients for the Item Intercept Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.494	.528	.525	.494	.477	.525	.541	.551	.525	.489
			Large	.146	.207	.220	.145	.166	.670	.661	.697	.669	.915
		0.5	Small	.514	.543	.541	.517	.492	.545	.550	.560	.550	.499
			Large	.146	.202	.215	.146	.165	.683	.668	.708	.683	.932
		0.8	Small	.503	.530	.528	.504	.481	.558	.563	.575	.560	.521
			Large	.149	.206	.221	.150	.165	.693	.679	.720	.701	.937
	40	0.2	Small	.060	.079	.085	.060	.064	.288	.298	.313	.287	.362
			Large	.107	.088	.087	.108	.067	.162	.118	.121	.162	.095
		0.5	Small	.054	.076	.085	.054	.058	.282	.295	.309	.281	.353
			Large	.106	.086	.085	.106	.068	.135	.091	.091	.135	.118
		0.8	Small	.042	.062	.068	.042	.044	.274	.286	.304	.276	.349
			Large	.096	.075	.075	.097	.055	.140	.098	.102	.139	.121
1000	20	0.2	Small	.517	.534	.528	.516	.495	.548	.546	.554	.550	.503
			Large	.167	.200	.206	.166	.181	.703	.691	.714	.703	.937
		0.5	Small	.509	.524	.519	.507	.486	.560	.555	.558	.556	.518
			Large	.153	.186	.190	.152	.170	.683	.672	.693	.684	.923
		0.8	Small	.509	.524	.521	.508	.487	.550	.548	.558	.551	.506
			Large	.147	.184	.191	.151	.173	.651	.645	.671	.659	.896
	40	0.2	Small	.041	.054	.058	.041	.046	.272	.278	.287	.272	.353
			Large	.101	.088	.088	.100	.059	.135	.109	.109	.135	.110
		0.5	Small	.045	.057	.061	.046	.050	.259	.268	.278	.259	.341
			Large	.103	.091	.090	.103	.061	.135	.111	.111	.135	.102
		0.8	Small	.040	.054	.058	.040	.046	.266	.273	.281	.265	.346
			Large	.097	.085	.084	.097	.055	.144	.120	.119	.143	.099

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 20. *Mean RMSE of the Two Regression Coefficients for the Item Interaction Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.222	.092	.097	.222	.110	.547	.364	.345	.543	.347
			Large	.234	.132	.133	.236	.171	.316	.359	.423	.314	.642
		0.5	Small	.233	.122	.134	.234	.156	.416	.295	.266	.413	.229
			Large	.252	.128	.121	.252	.187	.361	.393	.457	.362	.728
		0.8	Small	.239	.129	.136	.244	.165	.494	.339	.306	.494	.294
			Large	.229	.123	.122	.231	.166	.375	.410	.474	.377	.714
	40	0.2	Small	.204	.247	.256	.204	.224	.190	.143	.130	.191	.170
			Large	.072	.062	.063	.073	.062	.296	.191	.204	.296	.121
		0.5	Small	.187	.230	.240	.187	.208	.191	.138	.129	.193	.169
			Large	.086	.068	.068	.085	.065	.284	.192	.195	.285	.128
		0.8	Small	.179	.221	.231	.178	.199	.189	.132	.115	.188	.157
			Large	.069	.055	.055	.068	.053	.240	.149	.158	.243	.102
1000	20	0.2	Small	.175	.115	.121	.174	.103	.380	.319	.297	.379	.198
			Large	.223	.143	.135	.223	.155	.363	.386	.421	.363	.710
		0.5	Small	.181	.120	.122	.180	.101	.406	.343	.324	.410	.234
			Large	.218	.137	.133	.219	.150	.368	.385	.423	.365	.722
		0.8	Small	.150	.095	.097	.152	.086	.382	.323	.306	.384	.204
			Large	.215	.125	.117	.207	.136	.357	.377	.415	.363	.712
	40	0.2	Small	.183	.208	.214	.182	.205	.157	.127	.115	.159	.127
			Large	.043	.036	.037	.043	.037	.236	.179	.185	.236	.098
		0.5	Small	.180	.206	.211	.180	.202	.163	.132	.117	.162	.125
			Large	.052	.039	.037	.053	.034	.210	.153	.156	.208	.084
		0.8	Small	.164	.190	.196	.164	.185	.172	.145	.134	.175	.142
			Large	.050	.039	.037	.050	.037	.260	.200	.204	.259	.100

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 21. *Mean RMSE of the Two Regression Coefficients for the Item Intensity Parameter.*

				Continuous Covariate					Dichotomous Covariate				
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_D
500	20	0.2	Small	.089	.090	.095	.089	.146	.172	.170	.170	.172	.131
			Large	.033	.035	.038	.033	.091	.028	.030	.030	.029	.163
		0.5	Small	.089	.091	.095	.089	.147	.174	.172	.171	.173	.132
			Large	.029	.031	.035	.029	.087	.029	.030	.030	.028	.163
		0.8	Small	.089	.091	.095	.089	.146	.173	.171	.170	.171	.131
			Large	.029	.030	.034	.029	.086	.032	.033	.033	.031	.167
	40	0.2	Small	.038	.039	.041	.039	.054	.143	.142	.141	.143	.104
			Large	.017	.017	.016	.018	.003	.068	.068	.069	.068	.166
		0.5	Small	.038	.039	.041	.038	.054	.141	.140	.139	.141	.103
			Large	.017	.016	.015	.016	.003	.067	.067	.068	.066	.164
		0.8	Small	.037	.038	.039	.037	.053	.141	.141	.139	.142	.103
			Large	.017	.017	.015	.017	.004	.068	.068	.069	.067	.166
1000	20	0.2	Small	.092	.095	.098	.091	.150	.171	.172	.170	.171	.132
			Large	.029	.031	.034	.029	.087	.024	.025	.026	.026	.160
		0.5	Small	.089	.091	.096	.090	.147	.171	.170	.169	.170	.131
			Large	.028	.029	.033	.028	.086	.027	.027	.028	.027	.162
		0.8	Small	.090	.092	.097	.090	.148	.173	.172	.171	.172	.132
			Large	.027	.028	.032	.027	.084	.029	.030	.030	.028	.166
	40	0.2	Small	.038	.039	.040	.038	.054	.143	.142	.142	.143	.105
			Large	.017	.017	.016	.017	.002	.068	.068	.069	.068	.166
		0.5	Small	.038	.039	.040	.038	.054	.143	.142	.141	.143	.105
			Large	.017	.016	.015	.017	.002	.065	.065	.066	.065	.163
		0.8	Small	.038	.039	.040	.038	.054	.143	.142	.141	.143	.105
			Large	.017	.017	.016	.017	.003	.066	.066	.067	.066	.164

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 22. *Mean Bias, SE of the Regression Intercept for the Item Intercept Parameter.*

				Bias						SE					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	-.234	-.226	-.234	-.235	-.187	-.228	.069	.057	.058	.069	.044	.069
			Large	-.321	-.306	-.324	-.320	-.378	-.327	.062	.060	.060	.062	.046	.062
		0.5	Small	-.236	-.222	-.231	-.240	-.178	-.229	.053	.048	.050	.053	.041	.053
			Large	-.323	-.306	-.327	-.323	-.374	-.332	.069	.067	.068	.069	.058	.068
		0.8	Small	-.252	-.236	-.247	-.255	-.190	-.247	.065	.059	.062	.068	.053	.066
			Large	-.315	-.296	-.317	-.316	-.362	-.321	.069	.069	.068	.066	.054	.070
	40	0.2	Small	-.326	-.319	-.327	-.325	-.394	-.306	.057	.051	.052	.057	.033	.057
			Large	.174	.156	.150	.173	-.360	.213	.039	.041	.045	.039	.028	.041
		0.5	Small	-.328	-.324	-.330	-.326	-.400	-.307	.063	.058	.057	.063	.032	.062
			Large	.159	.142	.134	.159	-.359	.199	.045	.045	.049	.046	.041	.045
		0.8	Small	-.332	-.328	-.335	-.331	-.411	-.310	.065	.060	.062	.064	.041	.063
			Large	.177	.162	.153	.175	-.339	.218	.050	.052	.058	.050	.041	.051
1000	20	0.2	Small	-.263	-.251	-.256	-.262	-.204	-.258	.063	.058	.058	.064	.039	.063
			Large	-.316	-.305	-.317	-.315	-.358	-.327	.048	.049	.049	.048	.038	.049
		0.5	Small	-.265	-.253	-.259	-.265	-.200	-.260	.070	.065	.065	.070	.051	.070
			Large	-.298	-.287	-.299	-.299	-.351	-.309	.050	.049	.052	.051	.036	.050
		0.8	Small	-.244	-.233	-.239	-.243	-.183	-.239	.054	.049	.051	.056	.040	.054
			Large	-.285	-.275	-.292	-.291	-.349	-.296	.051	.049	.052	.053	.040	.051
	40	0.2	Small	-.328	-.325	-.328	-.328	-.402	-.313	.052	.050	.051	.051	.032	.050
			Large	.165	.155	.152	.166	-.356	.196	.031	.031	.033	.032	.023	.032
		0.5	Small	-.326	-.323	-.329	-.327	-.407	-.311	.036	.034	.036	.036	.031	.034
			Large	.178	.169	.166	.177	-.345	.210	.038	.039	.042	.037	.028	.039
		0.8	Small	-.330	-.325	-.329	-.329	-.408	-.313	.042	.042	.041	.042	.029	.042
			Large	.173	.165	.162	.173	-.353	.204	.033	.033	.035	.034	.028	.033

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 23. *Mean RMSE of the Regression Intercept for the Item Intercept Parameter.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.244	.233	.241	.245	.192	.239
			Large	.327	.312	.329	.326	.381	.333
		0.5	Small	.242	.227	.236	.245	.183	.235
			Large	.330	.314	.335	.330	.379	.339
		0.8	Small	.260	.244	.255	.264	.197	.255
			Large	.323	.304	.324	.323	.366	.329
	40	0.2	Small	.331	.324	.331	.330	.396	.311
			Large	.178	.161	.157	.177	.361	.217
		0.5	Small	.334	.329	.335	.332	.401	.313
			Large	.166	.149	.142	.166	.361	.204
		0.8	Small	.338	.333	.340	.337	.413	.316
			Large	.184	.170	.164	.182	.341	.224
1000	20	0.2	Small	.270	.258	.262	.270	.207	.265
			Large	.319	.309	.321	.319	.360	.330
		0.5	Small	.274	.261	.267	.275	.207	.270
			Large	.303	.291	.303	.303	.353	.313
		0.8	Small	.250	.238	.244	.250	.187	.245
			Large	.289	.280	.297	.295	.351	.301
	40	0.2	Small	.332	.329	.332	.332	.404	.317
			Large	.168	.158	.156	.169	.357	.198
		0.5	Small	.328	.325	.331	.329	.408	.313
			Large	.182	.174	.171	.181	.347	.213
		0.8	Small	.332	.328	.331	.331	.409	.315
			Large	.177	.168	.166	.177	.354	.207

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 24. *Mean Bias, SE of the Regression Intercept for the Item Interaction Parameter.*

				Bias						SE					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	-.270	-.245	-.222	-.269	.093	-.258	.160	.142	.143	.158	.107	.160
			Large	.179	.145	.174	.179	.393	.184	.085	.072	.074	.084	.057	.086
		0.5	Small	-.212	-.206	-.178	-.209	.106	-.198	.102	.093	.094	.102	.112	.102
			Large	.233	.193	.223	.234	.420	.245	.109	.100	.103	.109	.090	.107
		0.8	Small	-.243	-.229	-.195	-.234	.108	-.231	.100	.096	.100	.102	.109	.103
			Large	.221	.182	.213	.223	.407	.229	.087	.078	.080	.087	.075	.090
	40	0.2	Small	-.069	-.051	-.036	-.069	.223	-.083	.106	.098	.096	.108	.068	.106
			Large	-.262	-.218	-.221	-.262	.366	-.248	.068	.063	.066	.068	.058	.079
		0.5	Small	-.070	-.059	-.039	-.070	.229	-.084	.109	.098	.100	.110	.064	.112
			Large	-.253	-.217	-.215	-.254	.361	-.245	.073	.069	.071	.074	.062	.085
		0.8	Small	-.073	-.062	-.046	-.074	.237	-.091	.090	.083	.086	.089	.062	.089
			Large	-.247	-.211	-.210	-.248	.340	-.233	.078	.070	.075	.080	.059	.086
1000	20	0.2	Small	-.232	-.231	-.217	-.233	.081	-.219	.098	.085	.088	.097	.087	.097
			Large	.185	.161	.176	.185	.374	.197	.092	.087	.088	.093	.069	.093
		0.5	Small	-.196	-.196	-.179	-.196	.125	-.182	.101	.094	.093	.102	.068	.101
			Large	.169	.143	.158	.167	.354	.183	.067	.063	.067	.067	.066	.068
		0.8	Small	-.252	-.250	-.231	-.248	.068	-.237	.095	.087	.093	.100	.075	.095
			Large	.169	.144	.161	.170	.361	.183	.090	.086	.088	.091	.067	.090
	40	0.2	Small	-.074	-.069	-.056	-.074	.225	-.087	.061	.057	.059	.061	.050	.060
			Large	-.230	-.207	-.209	-.231	.353	-.231	.069	.066	.069	.069	.045	.074
		0.5	Small	-.072	-.065	-.051	-.070	.235	-.084	.058	.058	.057	.058	.045	.059
			Large	-.235	-.212	-.213	-.234	.339	-.233	.055	.050	.053	.054	.038	.059
		0.8	Small	-.060	-.057	-.047	-.063	.235	-.074	.076	.073	.071	.076	.048	.076
			Large	-.250	-.225	-.226	-.250	.352	-.250	.050	.046	.047	.049	.037	.052

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 25. *Mean RMSE of the Regression Intercept for the Item Interaction Parameter.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.313	.283	.264	.312	.142	.304
			Large	.198	.162	.189	.198	.397	.203
		0.5	Small	.235	.226	.202	.232	.154	.223
			Large	.258	.218	.245	.258	.430	.267
		0.8	Small	.262	.248	.219	.255	.154	.253
			Large	.238	.198	.227	.240	.414	.246
	40	0.2	Small	.126	.110	.102	.128	.233	.134
			Large	.270	.227	.231	.270	.370	.260
		0.5	Small	.130	.114	.107	.131	.238	.140
			Large	.264	.228	.227	.265	.366	.259
		0.8	Small	.116	.104	.097	.116	.245	.127
			Large	.260	.222	.223	.260	.345	.248
1000	20	0.2	Small	.252	.246	.234	.253	.119	.239
			Large	.206	.183	.196	.207	.380	.218
		0.5	Small	.220	.217	.202	.221	.143	.208
			Large	.182	.156	.171	.180	.361	.195
		0.8	Small	.269	.264	.249	.267	.101	.255
			Large	.191	.168	.184	.193	.367	.204
	40	0.2	Small	.096	.089	.082	.096	.231	.106
			Large	.241	.217	.220	.241	.356	.243
		0.5	Small	.092	.087	.077	.091	.239	.103
			Large	.242	.218	.219	.241	.341	.241
		0.8	Small	.096	.092	.085	.099	.240	.106
			Large	.255	.230	.231	.255	.354	.256

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 26. *Mean Bias, SE of the Regression Intercept for the Item Intensity Parameter.*

				Bias						SD					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	-.098	-.098	-.094	-.097	.062	-.095	.006	.007	.006	.006	.006	.007
			Large	.036	.036	.039	.037	.196	.046	.006	.007	.006	.006	.004	.006
		0.5	Small	-.147	-.147	-.143	-.146	.013	-.143	.005	.005	.005	.006	.005	.006
			Large	-.012	-.012	-.010	-.013	.148	-.003	.008	.009	.008	.008	.005	.008
		0.8	Small	-.090	-.089	-.086	-.089	.070	-.086	.009	.009	.008	.008	.005	.008
			Large	.048	.048	.051	.048	.207	.058	.006	.006	.006	.006	.004	.006
	40	0.2	Small	-.052	-.052	-.050	-.053	.142	-.056	.005	.004	.004	.005	.004	.005
			Large	.055	.054	.056	.054	.243	.045	.005	.005	.005	.005	.004	.005
		0.5	Small	-.098	-.098	-.096	-.099	.095	-.102	.007	.006	.006	.006	.003	.007
			Large	.007	.006	.009	.006	.196	-.003	.006	.006	.006	.006	.003	.006
		0.8	Small	-.041	-.042	-.040	-.042	.152	-.045	.004	.004	.004	.004	.004	.004
			Large	.065	.065	.067	.065	.254	.056	.005	.006	.005	.006	.004	.006
1000	20	0.2	Small	-.101	-.102	-.098	-.101	.058	-.098	.005	.005	.004	.004	.004	.004
			Large	.032	.031	.034	.032	.193	.041	.004	.005	.004	.004	.003	.004
		0.5	Small	-.099	-.099	-.096	-.099	.060	-.096	.005	.005	.004	.004	.004	.005
			Large	.034	.035	.037	.035	.195	.044	.005	.004	.004	.005	.003	.004
		0.8	Small	-.090	-.090	-.087	-.090	.069	-.087	.006	.006	.005	.005	.005	.006
			Large	.046	.045	.049	.046	.206	.055	.006	.006	.006	.006	.004	.006
	40	0.2	Small	-.055	-.055	-.054	-.055	.140	-.058	.003	.004	.003	.003	.002	.004
			Large	.052	.052	.053	.053	.241	.043	.004	.004	.004	.004	.002	.004
		0.5	Small	-.053	-.053	-.052	-.053	.141	-.057	.004	.004	.004	.004	.002	.004
			Large	.052	.052	.053	.052	.242	.043	.004	.004	.004	.004	.003	.004
		0.8	Small	-.044	-.044	-.042	-.043	.151	-.047	.004	.005	.004	.004	.003	.004
			Large	.061	.061	.062	.062	.251	.052	.004	.004	.004	.004	.002	.004

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 27. *Mean RMSE of the Regression Intercept for the Item Intensity Parameter.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.098	.098	.094	.097	.062	.095
			Large	.037	.037	.039	.037	.197	.047
		0.5	Small	.147	.147	.143	.146	.014	.143
			Large	.015	.015	.013	.015	.148	.009
		0.8	Small	.090	.089	.086	.089	.070	.087
			Large	.049	.049	.051	.049	.207	.058
	40	0.2	Small	.053	.053	.050	.053	.142	.056
			Large	.055	.055	.056	.055	.243	.046
		0.5	Small	.099	.098	.096	.099	.095	.102
			Large	.009	.009	.010	.009	.196	.007
		0.8	Small	.042	.042	.040	.042	.152	.045
			Large	.065	.065	.067	.065	.254	.056
1000	20	0.2	Small	.101	.102	.098	.101	.058	.098
			Large	.032	.032	.034	.033	.193	.041
		0.5	Small	.099	.099	.096	.099	.060	.096
			Large	.035	.035	.037	.035	.195	.044
		0.8	Small	.090	.090	.087	.090	.069	.088
			Large	.046	.046	.049	.046	.206	.056
	40	0.2	Small	.055	.055	.054	.055	.140	.058
			Large	.053	.052	.053	.053	.241	.043
		0.5	Small	.053	.053	.052	.053	.141	.057
			Large	.052	.052	.053	.052	.242	.043
		0.8	Small	.044	.044	.042	.043	.151	.047
			Large	.062	.061	.063	.062	.251	.052

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 28. *Mean Standard Error Bias of the Regression Coefficient for the Item Intercept Parameter.*

				Continuous Covariate						Dichotomous Covariate					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.223	.204	.205	.223	.207	.223	.312	.296	.297	.311	.288	.312
			Large	.155	.143	.141	.155	.145	.155	.229	.218	.215	.229	.279	.229
		0.5	Small	.212	.202	.203	.212	.203	.212	.340	.315	.316	.342	.324	.340
			Large	.140	.127	.126	.140	.132	.140	.225	.212	.211	.226	.280	.225
		0.8	Small	.210	.198	.199	.210	.192	.210	.323	.299	.299	.320	.294	.323
			Large	.149	.136	.132	.147	.140	.149	.240	.228	.223	.241	.292	.240
	40	0.2	Small	.104	.100	.099	.104	.101	.104	.253	.243	.240	.252	.265	.253
			Large	.049	.036	.036	.049	.068	.049	.119	.102	.103	.119	.298	.119
		0.5	Small	.121	.114	.115	.121	.118	.121	.249	.242	.238	.247	.264	.249
			Large	.047	.035	.037	.047	.065	.047	.124	.111	.111	.124	.303	.124
		0.8	Small	.115	.106	.105	.114	.113	.115	.244	.233	.225	.242	.260	.244
			Large	.054	.042	.043	.054	.071	.054	.114	.100	.099	.115	.286	.114
1000	20	0.2	Small	.222	.215	.216	.223	.212	.222	.325	.316	.312	.325	.309	.325
			Large	.147	.140	.140	.147	.137	.147	.249	.244	.239	.249	.300	.249
		0.5	Small	.215	.207	.208	.216	.202	.215	.338	.324	.325	.338	.314	.338
			Large	.150	.142	.141	.150	.143	.150	.222	.216	.210	.219	.281	.222
		0.8	Small	.218	.212	.210	.219	.206	.218	.332	.323	.318	.329	.312	.332
			Large	.141	.139	.136	.144	.138	.141	.227	.219	.215	.226	.285	.227
	40	0.2	Small	.124	.120	.121	.124	.121	.124	.274	.269	.265	.274	.290	.274
			Large	.053	.046	.047	.053	.073	.053	.118	.110	.110	.118	.306	.118
		0.5	Small	.116	.112	.112	.115	.114	.116	.284	.277	.274	.287	.305	.284
			Large	.055	.049	.049	.055	.075	.055	.127	.117	.116	.128	.314	.127
		0.8	Small	.121	.117	.117	.121	.118	.121	.268	.261	.260	.269	.284	.268
			Large	.057	.050	.051	.057	.078	.057	.119	.110	.109	.119	.305	.119

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 29. *Mean Standard Error Bias of the Regression Coefficient for the Item Interaction Parameter.*

				Continuous Covariate						Dichotomous Covariate					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.240	.195	.198	.241	.244	.240	.262	.219	.222	.268	.280	.262
			Large	.104	.046	.048	.102	.097	.104	.187	.140	.142	.186	.280	.187
		0.5	Small	.174	.159	.157	.173	.189	.174	.320	.263	.281	.326	.353	.320
			Large	.132	.084	.088	.135	.124	.132	.233	.184	.190	.235	.343	.233
		0.8	Small	.173	.153	.154	.174	.187	.173	.323	.264	.283	.327	.346	.323
			Large	.116	.058	.061	.118	.114	.116	.186	.135	.148	.187	.292	.186
	40	0.2	Small	.076	.058	.058	.076	.078	.076	.183	.136	.147	.182	.187	.183
			Large	.040	.012	.015	.040	.059	.040	.128	.081	.086	.128	.266	.128
		0.5	Small	.103	.084	.086	.103	.106	.103	.189	.157	.154	.189	.190	.189
			Large	.031	.007	.010	.032	.053	.031	.106	.059	.069	.104	.244	.106
		0.8	Small	.097	.073	.074	.097	.100	.097	.217	.182	.181	.219	.226	.217
			Large	.047	.019	.022	.047	.064	.047	.122	.080	.086	.120	.264	.122
1000	20	0.2	Small	.200	.188	.185	.201	.215	.200	.325	.294	.304	.324	.358	.325
			Large	.124	.098	.099	.123	.128	.124	.198	.172	.173	.198	.335	.198
		0.5	Small	.225	.207	.208	.226	.230	.225	.309	.283	.289	.309	.331	.309
			Large	.131	.104	.104	.131	.132	.131	.199	.168	.168	.198	.333	.199
		0.8	Small	.224	.208	.214	.226	.233	.224	.345	.317	.318	.340	.365	.345
			Large	.148	.120	.121	.149	.148	.148	.173	.144	.148	.173	.305	.173
	40	0.2	Small	.107	.095	.095	.108	.109	.107	.226	.201	.202	.223	.231	.226
			Large	.055	.039	.040	.055	.078	.055	.097	.074	.076	.097	.255	.097
		0.5	Small	.099	.086	.086	.098	.101	.099	.232	.202	.206	.232	.239	.232
			Large	.056	.042	.043	.056	.076	.056	.123	.102	.105	.123	.274	.123
		0.8	Small	.108	.095	.096	.108	.111	.108	.200	.176	.181	.198	.209	.200
			Large	.054	.039	.041	.054	.076	.054	.111	.086	.089	.110	.260	.111

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 30. *Mean Standard Error Bias of the Regression Coefficient for the Item Intensity Parameter.*

				Continuous Covariate						Dichotomous Covariate					
J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.165	.166	.168	.166	.165	.165	.260	.260	.260	.260	.247	.260
			Large	.124	.123	.126	.123	.129	.124	.190	.189	.190	.191	.220	.190
		0.5	Small	.168	.166	.170	.167	.167	.168	.258	.258	.258	.258	.246	.258
			Large	.124	.123	.126	.124	.128	.124	.187	.186	.187	.187	.217	.187
		0.8	Small	.168	.166	.170	.167	.168	.168	.255	.255	.255	.255	.244	.255
			Large	.123	.122	.126	.123	.128	.123	.190	.189	.190	.190	.220	.190
	40	0.2	Small	.067	.067	.069	.067	.070	.067	.153	.153	.156	.153	.161	.153
			Large	.041	.041	.043	.041	.046	.041	.092	.092	.094	.092	.173	.092
		0.5	Small	.067	.067	.069	.068	.070	.067	.150	.151	.153	.150	.158	.150
			Large	.041	.041	.043	.041	.046	.041	.091	.091	.092	.090	.172	.091
		0.8	Small	.067	.067	.069	.068	.070	.067	.153	.152	.155	.152	.161	.153
			Large	.040	.040	.042	.040	.045	.040	.093	.092	.094	.093	.173	.093
1000	20	0.2	Small	.168	.167	.170	.167	.168	.168	.260	.258	.260	.260	.248	.260
			Large	.125	.124	.127	.124	.129	.125	.193	.193	.193	.192	.223	.193
		0.5	Small	.169	.167	.171	.168	.168	.169	.261	.262	.261	.260	.248	.261
			Large	.126	.125	.128	.125	.130	.126	.191	.191	.191	.192	.221	.191
		0.8	Small	.167	.168	.170	.167	.167	.167	.259	.259	.259	.258	.247	.259
			Large	.124	.123	.127	.124	.130	.124	.189	.188	.190	.188	.220	.189
	40	0.2	Small	.068	.068	.070	.068	.071	.068	.153	.153	.156	.153	.161	.153
			Large	.042	.042	.044	.042	.048	.042	.094	.094	.096	.094	.175	.094
		0.5	Small	.069	.068	.071	.068	.071	.069	.154	.154	.157	.154	.162	.154
			Large	.042	.041	.044	.042	.047	.042	.094	.094	.096	.094	.174	.094
		0.8	Small	.068	.068	.070	.068	.071	.068	.154	.154	.157	.155	.162	.154
			Large	.041	.041	.044	.041	.047	.041	.095	.094	.096	.094	.176	.095

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 31. *Mean Standard Error Bias of the Regression Intercept for the Item Intercept and Item Interaction Parameters.*

	Item Intercept	Item Interaction

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D	JRT_DINA_LLTM	JRT_DINA	DINA	DINA-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.206	.191	.191	.206	.158	.199	.169	.128	.131	.173	.160	.191
			Large	.144	.133	.132	.144	.107	.184	.141	.104	.107	.141	.121	.220
		0.5	Small	.226	.203	.199	.226	.164	.217	.235	.184	.190	.237	.153	.255
			Large	.138	.126	.124	.138	.095	.179	.122	.086	.089	.122	.088	.209
		0.8	Small	.216	.194	.192	.213	.153	.207	.241	.184	.187	.242	.163	.258
			Large	.134	.121	.119	.137	.097	.172	.139	.099	.102	.140	.101	.218
	40	0.2	Small	.190	.180	.178	.189	.127	.202	.145	.106	.111	.142	.103	.148
			Large	.092	.080	.080	.092	.084	.226	.097	.062	.065	.097	.078	.199
		0.5	Small	.185	.175	.174	.185	.129	.197	.138	.109	.107	.138	.105	.140
			Large	.086	.075	.076	.085	.069	.222	.088	.052	.058	.087	.071	.192
		0.8	Small	.186	.176	.171	.186	.122	.200	.159	.123	.121	.159	.107	.164
			Large	.082	.069	.067	.082	.070	.213	.085	.053	.057	.083	.074	.191
1000	20	0.2	Small	.209	.199	.197	.208	.162	.202	.214	.197	.198	.216	.160	.235
			Large	.150	.143	.140	.149	.107	.190	.118	.099	.100	.116	.095	.213
		0.5	Small	.198	.188	.187	.198	.146	.191	.223	.197	.202	.222	.187	.241
			Large	.145	.139	.135	.144	.108	.189	.140	.119	.118	.140	.094	.235
		0.8	Small	.215	.205	.203	.213	.158	.208	.221	.197	.196	.215	.174	.238
			Large	.146	.141	.136	.145	.105	.190	.112	.092	.094	.112	.092	.210
	40	0.2	Small	.193	.187	.184	.194	.128	.209	.179	.158	.156	.179	.111	.185
			Large	.094	.089	.088	.094	.085	.238	.079	.059	.060	.079	.081	.192
		0.5	Small	.209	.204	.201	.210	.128	.226	.177	.151	.155	.178	.114	.181
			Large	.086	.079	.078	.087	.080	.230	.093	.075	.076	.093	.086	.211
		0.8	Small	.202	.196	.194	.203	.131	.218	.166	.146	.148	.165	.116	.173
			Large	.092	.086	.086	.091	.081	.236	.097	.078	.081	.098	.088	.216

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

Table B. 32. *Mean Standard Error Bias of the Regression Intercept for the Item Intensity Parameter.*

J	I	$\rho_{\theta\tau}$	γ	JRT_DINA_LLTM	JRT_DINA	Lognormal	Lognormal-LLTM	JRT_DINA_LLTM_C	JRT_DINA_LLTM_D
500	20	0.2	Small	.170	.168	.168	.169	.131	.166
			Large	.127	.124	.125	.127	.103	.150
		0.5	Small	.170	.169	.168	.170	.132	.167
			Large	.124	.122	.123	.124	.102	.148
		0.8	Small	.167	.166	.165	.167	.131	.165
			Large	.127	.124	.125	.126	.104	.150
	40	0.2	Small	.118	.116	.117	.117	.082	.124
			Large	.073	.069	.071	.073	.056	.133
		0.5	Small	.116	.114	.115	.117	.083	.122
			Large	.072	.069	.070	.072	.057	.132
		0.8	Small	.119	.116	.117	.118	.082	.125
			Large	.073	.069	.071	.073	.057	.133
1000	20	0.2	Small	.170	.169	.170	.171	.132	.168
			Large	.127	.125	.127	.127	.103	.151
		0.5	Small	.170	.169	.169	.171	.132	.167
			Large	.127	.126	.126	.126	.103	.151
		0.8	Small	.169	.169	.169	.170	.131	.167
			Large	.125	.124	.125	.125	.102	.149
	40	0.2	Small	.118	.117	.118	.118	.082	.124
			Large	.072	.071	.072	.072	.055	.134
		0.5	Small	.117	.117	.117	.118	.083	.124
			Large	.072	.071	.072	.072	.055	.133
		0.8	Small	.118	.116	.117	.117	.081	.123
			Large	.073	.071	.073	.072	.055	.134

Note. J = sample size; I = test length; $\rho_{\theta\tau}$ = correlation between ability and speed; γ = regression coefficients.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence.* New York, NY: Springer.

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating Student Covariates in Cognitive Diagnosis Models. *Journal of Classification*, 30(2), 195–224.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement*, *15*(2), 129-137.

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika*, *50*, 383-397.

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 62-79.

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, 1-23. doi: 1.1007/s11336-016-9537-6.

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 62-79.

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology*, *9*, 1525.

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, *41*(2), 123-145.

Bolsinova, M., & Tijmstra, J. (2019). Modeling differences between response times of correct and incorrect responses. *Psychometrika*, *84*(4), 1018-1046.

Bolsinova, M., & Tijmstra, J. (2017). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13-38.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 257-279.

Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. *Frontiers in Psychology*, *8*, 202.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 13-38.

Box, G. E. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*(2), 290–302.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal

Statistical Society, Series B (Methodological), 211-252.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*, 403–425. https://doi.org/10.1007/s11336-013-9350-4

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434-455.

Butter, R., De Boeck, P., & Verhelst, N. (1998). An item response model with internal restrictions on item difficulty. *Psychometrika*, *63*(1), 47–63.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618. https://doi.org/10.1177/0146621613488436

Cohen, J. (1965). Some statistical issues in psychological research. *Handbook of Clinical Psychology*, 95–121.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Congdon, P. (2003). *Applied Bayesian modelling*. Wiley.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *40*, 454–476. https://doi.org/10.3102/1076998615595403

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, 37–54.

De Boeck, P. (2008). Random Item IRT Models. *Psychometrika*, *73*, 533-559. https://doi.org/10.1007/s11336-008-9092-x

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 225-237.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1-28.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26. https://doi.org/10.1177/0146621610377081

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*(6), 447–468. https://doi.org/10.1177/0146621612449069

De Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model. *Journal of Consumer Research*, 34, 260–278. https://doi.org/10.1086/518532

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. Psychometrika, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J. (2007, April). Evaluation *of model fit in a large-scale assessment application of cognitive diagnosis*. Presentation at the annual meeting of the national council on measurement in education, Chicago, IL.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model:

Development and applications. *Journal of Educational Measurement*, *45*, 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130.

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

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

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*, 227–249. https://doi.org/10.1111/j.1745-3984.2010.00110.x

DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence*, *56*, 82-92.

Donders, F. (1869). Over de snelheid van psychische processen. *Nederlands Archief voor Genees- en Natuurkunde*, *4*, 117-145. [Translated and reproduced in 1969 as "On the speed of mental processes, in Acta Psychologica, 30, 412-431).

Embretson, S. E. (1985). Multicomponent latent trait models for test design. *Test design: Developments in psychology and psychometrics*, 195-218.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433. https://doi.org/10.1007/BF02294564

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economic and Statistics*, 92–107.

Ferrando, P. J.,& Lorenzo-Seva, U. (2007a). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543.

Ferrando, P. J., & Lorenzo-Seva, U. (2007b). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research*, *42*, 675–706.

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice*, *30*(4), 3–15. doi:10.1111/j.1745-3992.2011.00218.x

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications.* New York: Springer. (p.227)

Fox, J. P., Klein Entink, R. K., & Timmers, C. (2014). The joint multivariate modeling of multiple mixed response sources: Relating student performances with feedback behavior. *Multivariate Behavioral Research*, *49*(1), 54-66.

Fox, J. P., Klein Entink, R. K., & van der Linden, W. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software*, *20*, 1-14.

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4),

540-553.

Fox, J. P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*(2), 243-262.

Gaviria, J. L. (2005). Increase in precision when estimating parameters in computer assisted testing using response times. *Quality & Quantity*, *39*, 45–69.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515-534.

Gelman, S., & Gelman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721-741.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530-543.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, *12*(4), 369-381.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized Adaptive Testing With Item Cloning. *Applied Psychological Measurement*, *27*, 247–261. https://doi.org/10.1177/0146621603027004001

Glas, C. A., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 603-626.

Goldhammer, F., & Kroehne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: Experimental time limits, posterior time limits, and response time modeling. *Applied Psychological Measurement*, *38*(4), 255-267.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608-626.

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, *3*(1), 21-40.

Goldhammer, F., Steinwascher, M. A., Kroehne, U., & Naumann, J. (2017). Modelling individual response time effects between and within experimental speed conditions: A GLMM approach for speeded tests. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 238-256.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*(4), 351-373.

Gu, Y., & Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*, *84*(2), 468-483.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*, 665-686.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210.

Horsnell, G. (1953). The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika*, *40*(1/2), 128–136.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge University Press.

Ingrisone II, J. N. (2008). Modeling the joint distribution of response accuracy and response time (Doctoral dissertation). Retrieved from http://diginole.lib.fsu.edu/islandora/object/fsu%3A182101.

Janssen, R., Tuerlinckx, F., Meulders, M., & de Boeck, P. (2000). A Hierarchical IRT Model for Criterion-Referenced Measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285. https://doi.org/10.2307/1165207

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and

connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272. https://doi.org/10.1177/01466210122032064

Kang, H. A. (2016). Likelihood estimation for jointly analyzing item responses and response times (Doctoral dissertation). Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/92803/KANGDISSERTATION-2016.pdf?sequence=1&isAllowed=y.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*(3), 422.

Klein Entink, R. H. (2009). Statistical Models for Responses and Response Times (Doctoral dissertation). Retrieved from http://www.kleinentink.eu/download/ThesisKE.pdf.

Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21-48.

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

Klein Entink, R. H., van der Linden, W. J., & Fox, J. P. (2009). A Box-Cox normal model for response times. *The British Journal of Mathematical and Statistical Psychology*, *62*(Pt 3), 621–640. https://doi.org/10.1348/000711008X374126

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *The Journal of Experimental Education*, *43*(1), 61–69.

Lao, H. (2016). Estimation of diagnostic classification models without constraints: Issues with class label switching [PhD Thesis]. University of Kansas.

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359-379.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin & H. Hotelling (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). Stanford University Press.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: CRC Press.

Li, F. (2008). A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning (Unpublished doctoral dissertation). University of Georgia, Athens, GA.

Liao, D. (2018). *Modeling the speed-accuracy-difficulty interaction in joint modeling of responses and response time* (Doctoral dissertation, University of Maryland, College Park).

Liao, M., & Jiao, H. (2018). *Incorporating Item Features into Diagnostic Classification Models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New York City, NY.

Liu, R., & Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In L. A. van der Ark, D. M. Bolt, W.-

C. Wang, & J. A. Douglas (Eds.), Quantitative psychology research (pp. 243–254). Springer.

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, *76*(3), 487-503.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*(2), 204–209.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, *2*, 99–120. https://doi.org/10.2307/1164802

Magnus, B., Willoughby, M., Blair, C., & Kuhn, L. (2017). Integrating item accuracy and reaction time to improve the measurement of inhibitory control abilities in early childhood. *Assessment*, *26*(7), 1296-1306. doi:1.1177/1073191117740953

Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, *43*(8), 639-654.

Marianti, S. (2015). Contributions to the joint modeling of responses and response times (Doctoral dissertation). Retrieved from https://ris.utwente.nl/ws/portalfiles/portal/6052052.

Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*, 426-451.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*(3), 445-469.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77(4), 615-633.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Wadsworth.

McLeod, J., Butterbaugh, D., Masters, J., & Schaper, E. (2015, April). *Predicting item difficulty by analysis of language features.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Meng, X. B., Tao, J., & Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*(1), 1-27.

Minchen, N. P. (2017). Continuous Response in Cognitive Diagnosis Models: Response Time Modeling, Computerized Adaptive Testing, and Q-Matrix Validation (Doctoral Dissertation). Retrieved at https://rucore.libraries.rutgers.edu/rutgers-lib/55588/PDF/1/play/.

Mislevy, R. J. (1987). Exploiting Auxiliary Information About Examinees in the Estimation of Item Parameters. *Applied Psychological Measurement*, *11*(1), 81-91. https://doi.org/10.1177/014662168701100106

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, *12*(3), 281-296.

Mislevy, R. J., & Sheehan, K. M. (1988). The Information Matrix in Latent Variable Models. *ETS Research Report Series*, *1988*(1), i–34. https://doi.org/10.1002/j.2330-8516.1988.tb00280.x

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*(4), 661-679.

Molenaar, D., & Bolsinova, M. (2017). A heteroscedastic generalized linear model with a non-normal speed factor for responses and response times. *British Journal of Mathematical and Statistical Psychology*, 70(2), 297-316.

Molenaar, D., Bolsinova, M., Rozsa, S., & De Boeck, P. (2016). Response mixture modeling of intraindividual differences in responses and response times to the Hungarian WISC-IV Block Design test. *Journal of Intelligence*, *4*(3), 10-29.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2016). A semi-parametric within subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205-228. Retrieved from http://members.home.nl/jeroenvermunt/molenaar2016.pdf.

Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279-297.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606-626.

Molenaar, D., Rózsa, S., & Bolsinova, M. (2019). A heteroscedastic hidden Markov mixture model for responses and categorized response times. *Behavior Research Methods*, *51*(2), 676-696.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015a). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56-74.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015b). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197-219.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015c). Fitting diffusion item response theory models for responses and response times using the R package diffIRT. *Journal of Statistical Software*, *66*(4), 1-34.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*(2), 252-284.

OECD. (2014). *PISA 2012 Results: Creative problem Solving: Students' skills in tackling real-life problems (Volume V)*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264208070-en.

Park, J. Y., Johnson, M. S., & Lee, Y. S. (2015). Posterior predictive model checks for cognitive diagnostic models. *International Journal of Quantitative Research in Education*, *2*(3-4), 244-264.

Park, Y. S., & Lee, Y.-S. (2014). An Extension of the DINA Model Using Covariates: Examining Factors Affecting Response Probability and Latent Classification. *Applied Psychological Measurement*, *38*, 376–390. https://doi.org/10.1177/0146621614523830

Park, Y. S., Xing, K., & Lee, Y.-S. (2018). Explanatory Cognitive Diagnostic Models: Incorporating Latent and Observed Predictors. *Applied Psychological Measurement*, *42*, 376–392. https://doi.org/10.1177/0146621617738012

Park, Y. S., & Lee, Y.-S. (2019). Explanatory cognitive diagnostic models. In *Handbook of Diagnostic Classification Models* (pp. 207–222). Springer.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated?. *Intelligence*, *40*(1), 23-32.

Patton, J. M. (2015). Some consequences of response time model misspecification in educational measurement (Doctoral dissertation). Retrieved from https://curate.nd.edu/downloads/n583xs57z25.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, 114–133.

Plummer, M. (2015). JAGS Version 4.0.0 user manual. Lyon, France. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*(1), 41-70.

R Core Team (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL https://www.Rproject.org/.

Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*, *55*, 361–382.

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, *78*(3), 538-544.

Ranger, J., & Kuhn, J. T. (2012a). A flexible latent trait model for response times in tests. *Psychometrika*, *77*, 31–47.

Ranger, J., & Kuhn, J. T. (2012b). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*, *36*(3), 214-231.

Ranger, J., & Kuhn, J. T. (2013). Analyzing response times in tests with rank correlation approaches. *Journal of Educational and Behavioral Statistics*, *38*, 61–80. https://doi.org/10.3102/1076998611431086

Ranger, J., & Kuhn, J. T. (2014a). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 388-407.

Ranger, J., & Kuhn, J. T. (2014b). Testing fit of latent trait models for responses and response times in tests. *Psychological Test and Assessment Modeling*, *56*(4), 382-404.

Ranger, J., Kuhn, J. T., & Gaviria, J. L. (2015). A race model for responses and

response times in tests. *Psychometrika*, *80*(3), 791-810.

Ranger, J., Kuhn, J. T., & Szardenings, C. (2017). Analyzing model fit of psychometric process models: An overview, a new test and an application to the diffusion model. *British Journal of Mathematical and Statistical Psychology*, 70(2), 209-224.

Ranger, J., & Ortner, T. (2011). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *65*, 334–349. DOI:10.1111/j.2044-8317.2011.02032.x

Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*(2), 128-148.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogische Institut.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). Amsterdam: North-Holland.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 187–208). https://doi.org/10.1007/978-1-4757-2691-6_11

Rounder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions, *Psychometrika*, *68*, 589–606.

Rowe, M., Ozuru, Y., & McNamara, D. (2006). *An analysis of standardized reading ability tests: What do* questions *actually measure?* In ICLS 2006 - International Conference of the Learning Sciences, Proceedings (Vol. 2, pp. 627-633).

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications*. New York: Guilford.

Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*(1), 18-38.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist*, *6*(2), 461–464.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Schnipke, D. L., & Scrams, D. J. (1999). Representing response time information in item banks. *Law School Admission Council, Report 97-09.*

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Hillsdale, NJ: Lawrence Erlbaum

Associates.

Sinharay, S., & Johnson, M. S. (2019). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*.

Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement*, *44*, 376-392.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual* (Version 1.4.3). Cambridge, UK: MRC Biostatistics Unit.

Sternberg, R. J. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychology*, *30*, 276–315.

Sternberg, R. J. (1977a). *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1977b). Component processes in analogical reasoning. *Psychological Review*, *84*, 353–378. doi: 10.1037/0033-295X.84.4.353

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology*, *109*, 119–159. doi: 10.1037/0096-3445.109.2.119

Sternberg, R. J. (1985). *Beyond IQ. A Triarchic Theory of Human Intelligence*. Cambridge, MA: University Press.

Sternberg, R. J. (1986). Toward a unified theory of human reasoning. *Intelligence*, *10*, 281–314. doi: 10.1016/0160-2896(86)90001-2

Su, Y. S., & Yajima, M. (2015). R2jags: Using R to run 'JAGS'. R package version

0.5–7. *Available: CRAN. R-project. org/package= R2jags. (September 2015).*

Suh, H. (2010). A study of Bayesian estimation and comparison of response time models in item response theory (Doctoral dissertation). Retrieved from https://kuscholarworks.ku.edu/bitstream/handle/1808/6788/Suh_ku_0099D_10821_DATA_1.pdf?sequence=1&isAllowed=y.

Tang, F., & Zhan, P. (2020). The development of an instrument for longitudinal learning diagnosis of rational number operations based on parallel tests. *Frontiers in Psychology*, *11*, 2246.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, *10*(1), 55–73.

Templin, J. (2004). Generalized linear mixed proficiency models (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287.

Thissen, D. (1983). Timed Testing: An Approach Using Item Response Theory. In D. J. Weiss (Ed.), *New Horizons in Testing* (pp. 179–203). https://doi.org/10.1016/B978-0-12-742780-5.50019-6

Thomas, A., O'Hara, R., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*., *6*(1), 12-17. http://cran.r-project.org/doc/Rnews

Tiku, M. L. (1964). Approximating the general non-normal variance-ratio sampling distributions. *Biometrika*, *51*(1–2), 83–95.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629-65.

Van Breukelen, G. J. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, *70*(2), 359-376.

van der Linden, W. J. (2006). A Lognormal Model for Response Times on Test Items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. https://doi.org/10.3102/10769986031002181

van der Linden, W. J. (2009). Predictive Control of Speededness in Adaptive Testing. *Applied Psychological Measurement*, *33*(1), 25–41. https://doi.org/10.1177/0146621607314042

van der Linden, W. J. (2011a). Setting Time Limits on Tests. *Applied Psychological Measurement*, *35*(3), 183–199. https://doi.org/10.1177/0146621610391648

van der Linden, W. J. (2011b). Test Design and Speededness. *Journal of Educational Measurement*, *48*(1), 44–60. https://doi.org/10.1111/j.1745-3984.2010.00130.x

van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120-139.

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327-347.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive

Testing. *Applied Psychological Measurement*, *23*(3), 195–210. https://doi.org/10.1177/01466219922031329

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339-356.

Vandekerckhove, J. (2009). Extensions and applications of the diffusion model for two-choice response times (Doctoral dissertation). Retrieved from https://lirias.kuleuven.be/bitstream/1979/2658/2/Thesis.pdf.

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 317-345.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A Logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 169–185). https://doi.org/10.1007/978-1-4757-2691-6_10

Vermunt, J., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, 2005(2), i-35.

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*, 2014(2), 1–13.

Wagenmakers, E. J. (2009). Methodological and empirical developments for the

Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641-671.

Wang, T. (2006). A model for the joint distribution of item response and response time using a one-parameter Weibull distribution. (Center for Advanced Studies in Measurement and Assessment Research Report, no. 20). Iowa City, IA: University of Iowa.

Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 144-168.

Wang, S., & Chen, Y. (2020). Using Response Times and Response Accuracy to Measure Fluency Within Cognitive Diagnosis Models. *Psychometrika*, 85, 600-629. https://doi.org/10.1007/s11336-020-09717-2

Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*, 381-417.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wang, S., Hu, Y., Wang, Q., Wu, B., Shen, Y., & Carr, M. (2020). The development of a multidimensional diagnostic assessment with learning tools to improve 3-D mental rotation skills. *Frontiers in Psychology*, *11*, 305.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223-254.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456-477.

Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using Response Times to Assess Learning Progress: A Joint Model for Responses and Response Times. *Measurement: Interdisciplinary Research and Perspectives*, *16*, 45–58. https://doi.org/10.1080/15366367.2018.1435105

Wang, S., Zhang, S., & Shen, Y. (2019). A joint modeling framework of responses and response times to assess learning outcomes. *Multivariate Behavioral Research*, *55*(1), 49-68.

Wenger, M. J., & Gibson, B. S. (2004). Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 708-719.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. *Assessment of competencies in educational contexts*, 91-120.

Wood, R., Wilson, D. T., Muraki, E., Schilling, S. G., Gibbons, R., & Bock, R. D. (2002). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software International Inc.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (ETS Research Report Series, RR-03-32). Princeton, NJ: ETS.

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 262–286. https://doi.org/10.1111/bmsp.12114

Zhan, P., Jiao, H., Wang, W. C., & Man, K. (2018). A multidimensional hierarchical framework for modeling speed and ability in computer-based multidimensional tests. *arXiv preprint arXiv:1807.04003*.

Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in psychology*, *9*, 607.

Zhang, S., & Wang, S. (2018). Modeling Learner Heterogeneity: A Mixture Learning Model With Responses and Response Times. *Frontiers in Psychology*, *9*, 2339.

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 646-661.