ABSTRACT

Title of Dissertation:	MODELING THE SPEED-ACCURACY-DIFFICULTY INTERACTION IN JOINT MODELING OF RESPONSES AND RESPONSE TIME
	Dandan Liao, Doctor of Philosophy, 2018
Dissertation directed by:	Associate Professor, Hong Jiao Measurement, Statistics and Evaluation Department of Human Development and Quantitative Methodology

With the rapid development of information technology, computer-based tests have become more and more popular in large-scale assessments. Among all the auxiliary data collected during the test-taking process, response times (RTs) seem to be one of the most important and commonly utilized sources of information. A commonly adopted assumption in joint modeling of RTs and item responses is that item responses and RTs are conditionally independent given a person's speed and ability, and a person has constant speed and ability throughout the test (e.g., Thissen, 1983; van der Linden, 2007).

However, researchers have been investigating more complex scenarios where the conditional independence assumption between item responses and RTs is likely to be violated in various ways (e.g., De Boeck, Chen, & Davison, 2017; Meng, Tao, & Chang, 2015; Ranger & Ortner, 2012b). Empirical evidence suggests that the

direction of conditional dependence differs among items in a systematic way (Bolsinova, Tijmstra, & Molenaar, 2017). For difficult items, correct responses are associated with longer RTs; for easier items, however, correct responses are usually associated with shorter RTs (Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer, Naumann, & Greiff, 2015; Partchev & De Boeck, 2012). This phenomenon reflects a clear pattern that item difficulty affects the direction of conditional dependence between item responses and RTs. However, such an interaction has not been explicitly explored in jointly modeling of RT and response accuracy.

In the present study, various approaches for joint modeling of RT and response accuracy are proposed to account for the conditional dependence between responses and RTs due to the interaction among speed, accuracy, and item difficulty. Three simulation studies are carried out to compare the proposed models with van der Linden's (2007) hierarchical model that does not take into account the conditional dependence with respect to model fit and parameter recovery. The consequences of ignoring the conditional dependence between RT and item responses on parameter estimation is explored. Further, empirical data analyses are conducted to investigate the potential violations of the conditional independence assumption between item responses and RTs and obtain a more fundamental understanding of examinees' test-taking behaviors.

MODELING THE SPEED-ACCURACY-DIFFICULTY INTERACTION IN JOINT MODELING OF RESPONSES AND RESPONSE TIME

by

Dandan Liao

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
2018

Advisory Committee:
Professor Hong Jiao, Chair
Professor Robert W. Lissitz
Professor Yang Liu
Professor Matthias von Davier
Professor Yan Li, Dean’s Representative

Dedication

This dissertation is dedicated to my parents and my boyfriend. Much of the research I have done during my graduate study relates to conditional probability, yet there are things in life that are unconditional. I cannot be more grateful for the unconditional love and support from my parents that have always been with me along the journey. Thank you for everything you did to raise me into the person I am today. A special dedication to my boyfriend who carries me over the last, the most important, and the most difficult hurdle of my degree. I would not be able to accomplish this milestone without your love, understanding, and patience.

Acknowledgements

I would like to express my sincerest gratitude to my committee members. They all have been my mentors for different periods of my graduate study, and I believe the lessons I learned from them will benefit me for a lifetime.

First, I would like to thank Dr. Hong Jiao, my academic advisor and my supervisor at the Maryland Assessment Research Center (MARC). She has always been supportive and encouraging in my good and bad times. Since I joined the program as a master’s student, she has provided significant opportunities to practice and develop my professional development both on campus and in the industry. My accomplishments in this field, if any, are largely due to Dr. Jiao’s tremendous help and guidance. Without her full support, I would not be able to grow and bloom in such a nurturing atmosphere. She is not only my advisor, but also my role model.

I also would like to thank Dr. Robert W. Lissitz, my supervisor at MARC as well. It is a great luxury to work closely with Dr. Lissitz for four years. He has shared so much wisdom and humor with me in both academic and non-academic aspects. More importantly, he cares about the students wholeheartedly. Knowing that I tended to procrastinate important tasks because I tried to be perfect, he urged me to not seek perfection but to focus on a good idea for my dissertation. His suggestion to set up detailed deadlines for the steps of my dissertation enabled me to finish it in a timely manner.

My special gratitude also goes to Dr. Matthias von Davier, my internship mentor at National Board of Medical Examiners. Even though I did not work with Dr. von Davier on this dissertation during my internship, he provided a lot of inspiring

insights when I discussed possible topics with him. As I was struggled with the significance of this dissertation, he told me that science is incremental; every study has its unique value so that different pieces of research can complement each other as a big step forward. It was a great comfort for me as I was anxiously preparing for the proposal defense.

I also wish to acknowledge the valuable help from Dr. Yan Li, my supervisor at the Joint Program of Survey Methodology. She is the very first supervisor in my graduate study who led me into the door on research. With her guidance, my work as a first-year master's student won two important awards and was turned into two publications on flagship journals later. When I was young and ignorant (even though I still am), she was the one who taught me hard work takes me further than natural ability.

I also appreciate the help from Dr. Yang Liu for his patience and constant encouragement. He provided a lot of valuable suggestions to my dissertation work. Even after my proposal defense, he set up a meeting with me to go through the potential improvements with me line by line.

Additionally, I would like to thank all my colleagues, friends and family in the U.S. and China for their support and understanding. I am particularly grateful for sharing my sorrow and joy with Chen Li in four years of my graduate study. A recent study has shown that Ph.D. students are three times as likely to have mental health issues as undergraduate students (Levecque, Anseel, De Beuckelaer, Van der Heyden, & Gisle, 2017). Having Chen as my company along every step of my journey is a key factor to keep me away from these issues.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

With the rapid development of information technology, computer-based tests have gained increasing popularity in large-scale assessments. Among all the auxiliary data collected during the test-taking process, response times (RTs) is one of the most important and commonly utilized sources of information. To better understand examinees' test-taking behaviors, various modeling frameworks have been proposed to analyze RT and its relationship with response accuracy (RA). Most existing research has focused on the relationship between RT and RA assuming all examinees respond to items in the same manner. However, increasing empirical research suggests that examinees undertake different response styles or switch problem-solving strategies for items with different characteristics. The present study investigates the relationship among RT, RA, and one of the most important psychometric item characteristics, item difficulty, by proposing a series of modeling approaches. The potential impact of accounting for or ignoring the interaction among speed, accuracy and item difficulty is explored.

1.1 Statement of the Problem

RT has been playing a crucial role in experimental cognitive psychology since the 1950s (Luce, 1986). It is believed that RTs reflect the time needed for basic thinking processes, including interpreting a stimulus, retrieving information, processing information to respond to a stimulus, and synthesizing information from multiple sources in both psychological and educational tests. Depending on the characteristics of the items on the test and allotted time for the test, there are mainly two types of tests, namely speed tests and power tests, as Gulliksen (1950) first pointed out. In a pure speed test, examinees are asked to work on as many items as possible in limited time. Items on a speed test are relatively easy, and the total count of items completed with correct answers directly reflects how fast an examinee can respond to similar items. On the other hand, in a pure power test, the accuracy of an examinee's response is measured instead. Compared to those on a speed test, items on a power test vary in difficulty or complexity in problem-solving process. Another important feature of power tests is that examinees respond to items without a time constraint. Therefore, their abilities may be measured more accurately as they are given enough time to attempt all items.

However, most educational tests fall in the category of neither pure speed tests nor pure power tests. Items on educational tests are usually selected to cover a spectrum of item difficulty, so that the ability parameters could be estimated with adequate precision along the whole scale. Due to practical concerns in cost- and time-effectiveness, examinees are required to respond to items within a certain time frame, regardless of whether they can reach all items or not. Thereby, most educational tests are in fact power tests administered under time constraints. As such, Hambleton and Swaminathan (1985) asserted that latency should be studied in addition to correctness. Thissen (1983) also argued that RA and RT are two dimensions involved in analyzing data from timed tests, and that modeling either dimension and ignoring another may yield biased or misleading results.

In fact, the relationship between speed and accuracy has been of interest to psychologists for over a century. A choice behavior was found ubiquitous across multiple species – a subject tends to respond to a stimulus less accurately for faster responses and more accurately for slower responses. This is often referred to as the “speed-accuracy tradeoff” (e.g., Garrett, 1922; Henmon, 1911; Luce, 1986). In the cognitive psychology field, it is usually considered a within-subject phenomenon that reflects how speed fluctuates for a certain examinee with a certain ability level (van der Linden, 2009). On the other hand, psychometric researchers have been interested in both within-subject and between-subject variabilities between speed and accuracy. For instance, researchers have asked whether examinees respond to different items with different speed and ability? Do examinees with higher ability level tend to respond faster or slower? Is spending more time on the items associated with higher probabilities of correct responses?

The nature of educational tests necessitates the study of RT and RA, and the availability of RTs alongside with response data renders the possibility of modeling RT and RA simultaneously. Psychometric researchers have proposed various approaches to modeling both RT and RA, including the drift diffusion model (e.g., Ratcliff, 1978; Tuerlinckx & De Boeck, 2005) and the hierarchical modeling for speed and accuracy (van der Linden, 2007).

A common assumption in joint modeling of RT and RA is the conditional independence assumption, which is important from both substantive and statistical aspects (van der Linden, 2009). It is assumed that item responses depend solely on the latent ability and RTs depend only on the latent speed. In other words, item responses and RTs are conditionally independent given a person’s speed and ability. However, researchers have uncovered more complex real-world scenarios where the conditional independence assumption between item responses and RTs is likely to be violated. Further, an interesting scenario has been found in real data from multiple testing programs that the direction of conditional dependence differs among items depending on item difficulty in a systematic way (e.g., Bolsinova, Tijmstra, & Molenaar, 2017; Goldhammer, Naumann, & Greiff, 2015). To further improve estimation accuracy and to better understand examinees' test-taking behaviors at a fundamental level, it may be of practical and theoretical importance to investigate the relationship among speed, accuracy, and item characteristics.

1.2 Purpose of the Study

The purpose of the present study is to explore the relationship among speed, accuracy and item difficulty, one of the most important item characteristics. This study is motivated by examining empirical data from multiple testing programs and the results from recent studies on speed and accuracy. In particular, there seems to be a consistent pattern regarding the conditional dependence between speed and accuracy, and its interaction with item difficulty. For difficult items, correct responses are associated with longer RTs, which appears to follow the speed-accuracy tradeoff; for easier items, however, correct responses are usually associated with shorter RTs, indicating an opposite pattern of the speed-accuracy tradeoff (Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer et al., 2015; Goldhammer, Naumann, Stelter, Tóth, & Rölke, 2014; Partchev & De Boeck, 2012).

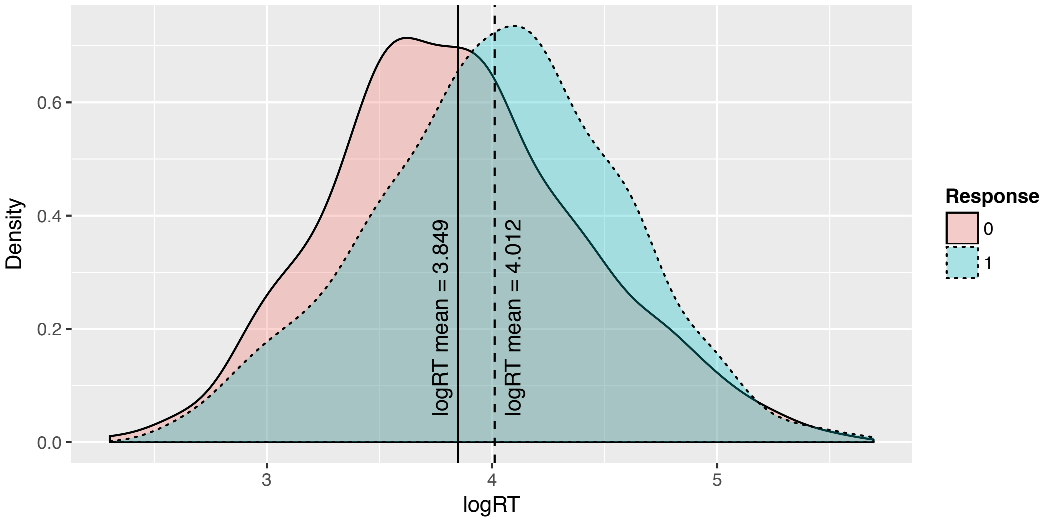


Figure 1. Logarithm of RT distributions for correct and incorrect responses of a difficult item with difficulty of 1.337.

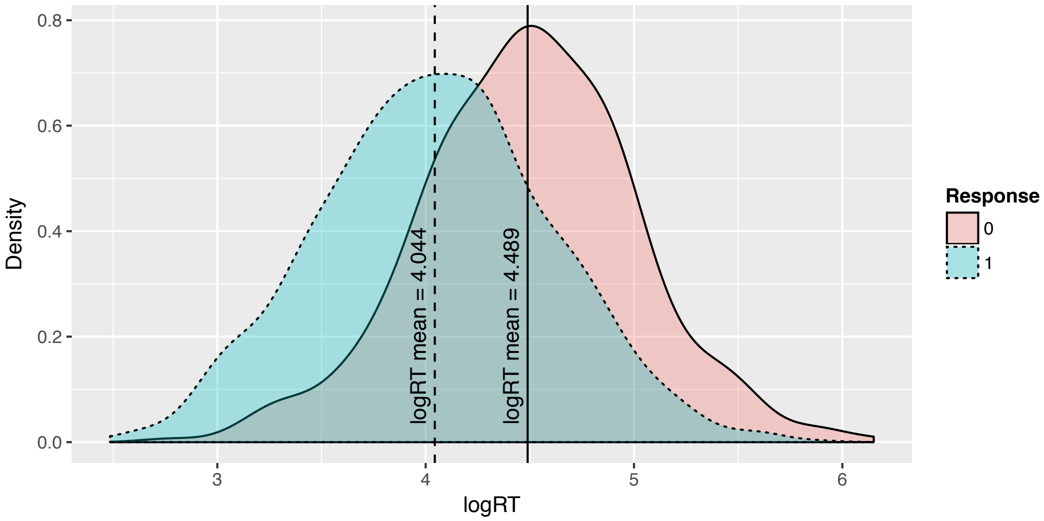


Figure 2. Logarithm of RT distributions for correct and incorrect responses of an easy item with difficulty of -.546.

To illustrate the relationship among speed, accuracy and item difficulty, Figures 1 and 2 demonstrate the phenomenon that the relationship between speed and accuracy tends to interact with item difficulty using response data and RTs from a large-scale credentialing testing program (Cizek & Wollack, 2017). The logarithm transformation is applied as a common approach to normalize RTs, denoted as logRT.

In both figures, the dashed density in blue indicates the distribution of logRT for correct responses, whereas the solid density in red indicates the distribution of logRT for incorrect responses. The dashed and solid vertical lines represent the mean logRT for correct and incorrect responses, respectively. Figure 1 shows the logRT distributions for correct and incorrect responses from a difficult item with a difficulty parameter of 1.337. The mean of logRT distribution for correct responses, as indicated by the dashed line on the right, is larger than that for incorrect responses. In Figure 2, for an easier item, the locations of the two distributions are reversed: incorrect responses are associated with longer RTs on average, whereas correct responses tend to be faster. On the untransformed scale, the difference between the means of RT is centered around 20 seconds, but could be larger than 40 seconds in the most extreme cases (see the left panel of Figure 3), which is not negligible given that the average RT among all examinees and items is about 65 seconds.

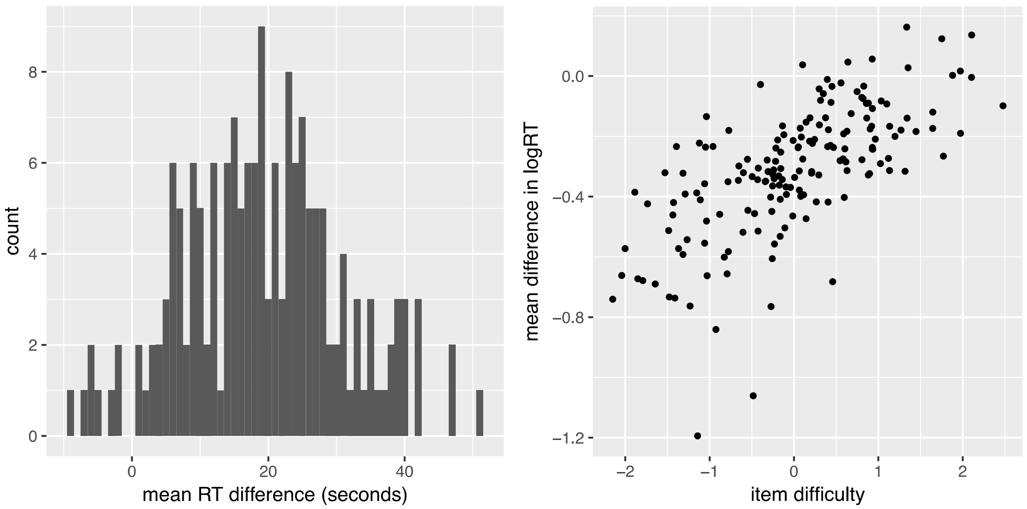


Figure 3. Histogram of the mean RT difference (left panel) and the scatterplot for the mean logRT difference and item difficulty (right panel).

Based on scenarios presented in Figures 1 and 2, some statistical tests and modeling approaches have been proposed to examine RT distributions conditioning on observed responses (e.g., Bolsinova & Maris, 2016; van der Linden & Glas, 2010). Findings from other studies also suggest that speed and accuracy interface in opposite directions for items with different difficulty levels (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer et al., 2014; Partchev & De Boeck, 2012). Yet the relationship between the magnitude of location shift in logRT and item difficulty has not been explicitly studied. A closer examination of the interaction reveals that the location shift of logRT between correct and incorrect responses for each item is strongly correlated with the item difficulty, as shown in the right panel of Figure 3. Patterns presented in Figures 1 to 3 have been cross-validated with data from testing programs in different fields.

Motivated by this phenomenon observed in empirical data and reported in previous studies (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer et al., 2014; Partchev & De Boeck, 2012), the present study aims at exploring the conditional dependence between responses and RTs due to the interaction among speed, accuracy, and item difficulty. Specifically, this study addresses the following questions:

1. What are the possible approaches to modeling RT and responses for speed-accuracy-difficulty interaction?

2. How are the item and person parameter estimates in the proposed models affected by manipulated factors in simulation studies, including sample size, the

number of items, the correlation between speed and ability, and the correlation between shift in time intensity parameter and item difficulty?

3. How do the proposed models perform compared to existing models for joint modeling of RT and RA, in simulation studies and real data analysis?

4. What is the impact of ignoring conditional dependence on parameter recovery?

5. Which model fit indices perform better on identifying the proposed models as the best fitting models under different simulation conditions when the proposed models are used for data generation?

These questions are answered in light of the findings from various simulation conditions and analyses of empirical data from several computer-based large-scale assessment programs.

1.3 Significance of the Study

There are three outcomes expected from this study. First, this study is motivated by a phenomenon that is common across different testing programs. Although previous studies have suggested that the direction of conditional dependence between RT and RA seems to be associated with item difficulty, the relationship between the magnitude of conditional dependence and item difficulty has not been explored in sufficient detail. In the present study, the interaction among speed, accuracy and item difficulty is investigated in detail by proposing a series of modeling approaches. In the proposed models, examinees could be classified by observed responses or latent variables, and the magnitude of conditional dependence is allowed to covary with item difficulty in different ways. By comparing the proposed models to existing models that either take into account or ignore the conditional dependence between RT and RA (van der Linden, 2007; van der Linden & Glas, 2010), a more fundamental understanding is obtained regarding the mechanisms of examinees' test-taking behaviors.

Second, the impact of ignoring conditional dependence between RT and RA is evaluated in the present study. The assumption of conditional independence has been frequently adopted as it facilitates the development of joint modeling of RT and RA based on separate modeling frameworks. However, investigation of empirical datasets indicates that this assumption does not usually hold under certain scenarios (e.g., Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016). As a result, parametric and non-parametric statistical tests have been developed to detect violations of this assumption (e.g., Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; van der Linden & Glas, 2010). In the present study, a detailed exploration is conducted for the potential impact on item and person parameter recovery under various simulation conditions when such dependence is ignored.

Third, accounting for the relationship among speed, accuracy and item difficulty may potentially improve the accuracy of item and person parameter estimates. On the one hand, classifying examinees by latent variables may be more accurate than observed responses, which may contain more undefined errors such as guessing and slipping. On the other hand, incorporating covariates in modeling RT could also contribute to higher precision in parameter estimation, as more information is utilized in the estimation process. As such, the proposed modeling approaches may yield more accurate parameter estimates.

1.4 Overview of the Chapters

In the following chapters, different approaches to modeling the interaction among speed, accuracy and difficulty are described and evaluated from both practical and theoretical perspectives.

In Chapter 2, the background and rationale for the study is established through a comprehensive literature review of existing methods related to modeling responses and RTs. First, common unidimensional IRT models and IRT models that incorporate RT as a covariate are reviewed. Second, various approaches and distributions for modeling RT are summarized, along with RT models that utilize information from RA-related latent variables. Moreover, findings are outlined based on the explorations of the relationship between RT and item characteristics, as well as the relationship between RT and the correctness of responses. Driven by the increasing need for measuring speed and accuracy simultaneously in timed tests, frameworks for joint modeling of RT and RA are elaborated. While most joint modeling methods are built on the assumption of conditional independence for simplicity and interpretability, it is often found that this assumption is violated in practice. Thus, recent development of different approaches to accounting for different types of violations are then surveyed. Lastly, the technical details of commonly used estimation methods are elaborated.

Chapter 3 describes the methods utilized in the present study in detail. The first section of this chapter elaborates the proposed models for speed-accuracy-difficulty interaction, which are extensions based on the current joint modeling framework that assumes conditional independence between RT and RA. In the second section, Bayesian estimation of model parameters via Markov chain Monte Carlo

(MCMC) is demonstrated. Technical details are provided regarding prior distribution, posterior distribution, and convergence criteria. The third section focuses on three simulation studies that evaluate the performance of the proposed models and existing models as well as the impact of ignoring conditional dependence under various conditions. Empirical data from two large-scale assessment programs are analyzed to demonstrate the application and utility of the proposed models in real testing scenarios.

Results from the simulation study and real data analysis are presented in Chapter 4. The recovery of model parameters and the impact of manipulated factors are reported for the simulation study. Several model fit indices are compared with respect to the detection rate of the data generating model. Moreover, different approaches to modeling the speed-accuracy-difficulty interaction are compared based on parameter recovery and model fit in empirical data.

Lastly, findings from the present study are summarized in Chapter 5. Interpretations and implications of the results are discussed regarding large-scale computer-based assessment where both responses and RTs are recorded. In addition, limitations of the study shed light on some future research directions for further understanding of test-taking behaviors and improving parameter estimation accuracy.

Chapter 2: Literature Review

This chapter reviews modeling approaches and estimation methods related to RT and RA in four sections, which serve as theoretical foundation for the proposed models in Chapter 3. In particular, the first two sections introduce IRT models and RT models for measuring latent ability and speed in separate frameworks. Building upon the two separate frameworks, researchers have proposed methods for joint modeling of RT and RA. In the following sections, the theoretical and practical implication of the conditional independence assumption is detailed, and then the reviewed methods are classified into three categories and summarized respectively: (a) joint modeling of RT and RA assuming conditional independence between RT and responses; (b) joint modeling of RT and RA distinguishing fast and slow responses; (c) joint modeling of RT and RA distinguishing correct and incorrect responses. Models that belong to the last two categories explicitly tackle with two types of violations of the conditional independence assumption. The estimation methods used for the proposal models are elaborated in the last section of this chapter.

2.1 Item Response Modeling

2.1.1 Standard IRT Models

IRT, also referred to as modern measurement theory, is a theory that concerns with latent ability on a psychological continuum and its relationship with item characteristics. The probability of a correct response for a specific item and a certain person is associated with the latent ability of the person and characteristics of the item via a logistic or probit link. Rather than estimating the true ability using summed score as in classical test theory, IRT focuses on modeling item-level responses that compose the total score. A major advantage of IRT is that it provides a more flexible framework for test users to put student ability and the attributes of the items (i.e., difficulty) on a common scale. Therefore, comparison among scores from different test forms is meaningful. Due to practical and theoretical advantages, IRT has received considerable attention in education, psychology (Embretson & Reise, 2000) and other fields, such as clinical research (e.g., Tractenberg, 2010), economics (e.g., Monica, 2008), political science (e.g., Clinton, Jackman, & Rivers, 2004; Matin & Quinn, 2002), and medical research (e.g., Cella et al., 2007).

The substantial benefits of IRT are built on a set of rigorous assumptions. Two most important fundamental assumptions are unidimensionality and local independence (Reckase, 2009). The unidimensionality assumption requires that the parameter that describes examinees only captures variance in one latent dimension (Lord & Novick, 1968; Rasch, 1960). However, this assumption is often violated in real testing scenarios, especially when tests have increasingly been developed for assessing skills from more than one dimensions. Extensive research has been conducted to determine the consequences of violating this assumption (e.g., Bolt, 1999; Camilli, Wang, & Fesq, 1995; Champlain, 1996; Jang & Roussos, 2007). To accommodate the multidimensional nature of more recent tests, researchers have proposed theory and estimation methods for multidimensional IRT models for simple and complex, compensatory and noncompensatory structures (e.g., Mulaik, 1972; Reckase, 1972, 2009; Sympson, 1978).

On the other hand, the local independence assumption entails two facets, local item independence and local person independence. That is, the probability of answering one item correctly does not increase or reduce the probability of a correct response to another item, and one person's probability of answering an item correctly does not influence another person's probability of a correct response. Possible causes of local item dependence (LID) might be related to additional factors that consistently affect the performance of some students on some items, such as speededness, practice, testlet dependence, item chaining etc (Yen, 1984, 1993). Such LID has been accounted for to improve estimation accuracy as random effects or interaction effects in modeling conditional distributions or log odds of possible response patterns (e.g., Bradlow, Wainer, & Wang, 1999; Hoskens & De Boeck, 1997; Ip, 2000; Wang & Wilson, 2005).

Since the present study aims at investigating the relationship between latent speed and ability as well as its interaction with item difficulty, only three common unidimensional IRT models for dichotomous items are reviewed in this section, including the Rasch (Rasch, 1960), the two-parameter logistic (2PL; Birnbaum, 1968), and the three-parameter logistic (3PL; Birnbaum, 1968) models.

The Rasch model is the most basic IRT model that characterizes the probability of a correct response with a person’s latent ability and an item difficulty parameter. It places ability and difficulty on the same scale, and assumes that higher latent ability or lower item difficulty leads to higher probability of obtaining a correct answer to the item. Such a monotonic relationship is described via a logistic function:

$$P(y_{ij} = 1|\theta_j, b_i) = \frac{1}{1 + \exp[-(\theta_j - b_i)]}, \quad (2.1)$$

where y_{ij} denotes the observed response for person j on item i, θ_j represents the latent ability for person j, and b_i indicates the difficulty parameter for item i, which is defined as the level on the latent continuum that yields a probability of .5 for a correct response.

This model assumes that all items discriminate among examinees equally, which means that any increase of the same distance from item difficulty would result in the same increase in the probability of a correct response for all items, and vice versa. Based on the Rasch model, the 2PL model was developed to allow differential discrimination parameters across items:

$$P(y_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1 + \exp[-a_i(\theta_j - b_i)]}. \quad (2.2)$$

Compared to (2.1), an additional parameter a_i is included in the formulation, reflecting the discrimination power specific for item i. a_i is constrained to be a positive value in most cases, which indicates that the monotonic increasing assumption between the latent ability and the probability of a correct response is maintained. Items that are more discriminating tend to have higher a_i values, where the probability of a correct response increases faster as the ability level increases.

The most generalized IRT model among the three is the 3PL model. This model accommodates a common testing scenario where examinees choose to make a guess on an item when they do not have the time or ability to solve the item. Thus a lower asymptote c_i is involved to tease out the effect of guessing or pseudo-guessing:

$$P\big(y_{ij} = 1\big|\theta_j, a_i, b_i, c_i\big) = c_i + \frac{1 - c_i}{1 + \exp\big[-a_i(\theta_j - b_i)\big]}. \tag{2.3}$$

In this model, the probability of a correct response no longer ranges from 0 to 1, but has c_i as the lower asymptote. The item difficulty b_i now represents the point where the probability of a correct response reaches $\frac{1+c_i}{2}$.

2.1.2 Incorporating RT for Modeling RA

Section 2.1.1 introduces three most common IRT models for measuring latent ability, the Rasch, the 2PL, and the 3PL IRT models. Built upon the standard IRT models, this section summarizes the models that incorporate RT as collateral information in the IRT models (Roskam, 1987, 1997; Verhelst et al., 1997; Wang & Hanson, 2005). A research question of interest in the studies reviewed in this section is how to model the speed-accuracy tradeoff with regard to the probability of a correct response. In all three studies, this relationship is modeled as linear combinations of latent ability, speed, and/or RT, reflecting the relative impact of speed and ability on the probability of a correct response. The specific assumptions and formulations in each study are elaborated respectively.

One of the first attempts to include RT information in IRT models is Roskam's Rasch response time model (Roskam, 1987, 1997):

$$P\big(y_{ij} = 1|\theta_j, t_{ij}, b_i\big) = \frac{\theta_j t_{ij}}{\theta_j t_{ij} + b_i} = \frac{\exp(\theta_j^* + t_{ij}^* - b_i^*)}{1 + \exp(\theta_j^* + t_{ij}^* - b_i^*)}. \tag{2.4}$$

In his model, θ_j is called mental speed, t_{ij} represents the RT that person *j* spends on item *i*, and b_i is the item difficulty, of which θ_j^*, t_{ij}^*, and b_i^* are the population logarithm analogues. The ability parameter in the standard Rasch model is replaced by an “effective ability parameter”, denoted by $\theta_j t_{ij}$, or equivalently, $\theta_j^* + t_{ij}^*$ on the logarithm scale. A speed-accuracy tradeoff is captured in this model as the probability of a correct response increases when t_{ij} increases for a given item. In a pure power test where in theory t_{ij} can increase to infinity, the probability of a correct response for any item is one.

Rather than including actual RT as a covariate, Verhelst et al. (1997) proposed a similar model that incorporates the effect of speed in a Rasch-like model. Moreover, the inclusion of a shape parameter for item *i* permits more flexible RT distributions:

$$P\left(y_{ij} = 1 | \theta_j, \tau_j, b_i\right) = \left\{\frac{1}{1 + \exp\left[-\left(\theta_j - \tau_j - b_i\right)\right]}\right\}^{\pi_i}, \qquad (2.5)$$

where θ_j and τ_j are the latent ability and speed parameters for person j, b_i is the item difficulty, and π_i is an item-specific parameter that allows the shapes of the RT distribution to be different across items. Their model is derived from the product of two distributions, a generalized extreme-value distribution for the latent ability conditioning on RT, and a gamma distribution for the marginal RT distribution. Similar to Roskam (1987, 1997), Verhelst et al. (1997) also assumed that faster responses are associated with lower probability of a correct response.

Both Roskam’s (1987, 1997) and Verhelst et al.’s (1997) models can be viewed as variations of the Rasch model. Wang and Hanson (2005), on the other hand, proposed a four-parameter logistic (4PL) model based on the 3PL model. Its formulation is expressed as follows:

$$P\left(y_{ij} = 1 | \theta_j, \rho_j, a_i, b_i, c_i, d_i\right) = c_i + \frac{(1 - c_i)}{1 + \exp\left[-a_i\left(\theta_j - \rho_j d_i / t_{ij} - b_i\right)\right]}, \qquad (2.6)$$

where all parameters are defined in the same manner as those for the 3PL model in (2.3), except one term in the logit function $\rho_j d_i / t_{ij}$. In this term, ρ_j and d_i are referred to as the slowness parameters of person j and item i. With this term incorporated, the authors assume that item and person slowness parameters have the same effect on the probability of a correct response. Moreover, a speed-accuracy tradeoff is imposed in that as t_{ij} increases, the probability of a correct response also increases and the effect of t_{ij} gradually fades out. As t_{ij} approaches infinity, the $\rho_j d_i / t_{ij}$ term drops out and the functional form in (2.6) approximates the probability of the 3PL model. In other words, an examinee who spends more time on an item is more likely to respond correctly to it, but the effect of RT on RA becomes negligible when the examinee is allowed to use as much time as needed on the item. Such a formulation may be more realistic than the assumption applied in Roskam's (1987, 1997) and Verhelst et al.'s (1997) models, that the probability of a correct response approaches one as RT approximates infinity or an examinee responds extremely slowly to an item, regardless of item and person characteristics.

2.2 RT Modeling

RTs on test items are a reliable and potentially valuable source of information for modeling speed as well as serving as collateral information for modeling latent ability. It has been shown that incorporating RT can improve ability estimation accuracy (e.g., Ferrando & Lorenzo-Seva, 2007; Meng et al., 2015), detect aberrant response behaviors (e.g., Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; van der Linden & Guo, 2008), control for differential speededness in computerized adaptive testing (e.g., van der Linden, Scrams, & Schnipke, 1999; van der Linden & Xiong, 2013). This section introduces some models and the distributions that are frequently utilized in RT modeling.

2.2.1 Standard RT Models

In most testing scenarios, RT distributions are non-negative and positively skewed, which motivates the choice of log-normal distributions for RT modeling. In fact, several studies have reported good fit in modeling actual RTs using log-normal distributions (e.g., Schnipke & Scarms, 1999; Thissen, 1983; van der Linden, Scrams, & Schnipke, 1999). Moreover, the use of the nice statistical properties of a normal model is permitted by adopting the log-normal transformation (Klein Entink, van der Linden, & Fox, 2009). van der Linden (2006) proposed a log-normal model for RT analogous to the 2PL IRT model, which is specified as follows:

$$\log\left(t_{ij}\right) = \beta_i - \tau_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.7}$$

where β_i is the item time intensity parameter for item i, τ_j is the speed parameter for person j, and the error term ε_{ij} is assumed to be normally distributed with mean of 0 and inverse variance of α_i. α_i is also referred to as the item-specific time discrimination parameter, which quantifies the variability of the logRT distribution. As one can see, α_i and β_i are the counterparts of a_i and b_i in the 2PL IRT model. Similar formulations are used for controlling differential speededness (e.g., van der Linden et al, 1999) and detecting aberrant response behaviors (e.g., van der Linden & van Krimpen-Stoop, 2003). Fox, Klein Entink, and van der Linden (2007) and Klein Entink, Fox, and van der Linden (2009) also included a slope parameter for the person speed parameter to characterize differential effects of items on the examinees.

Finding that the log transformation cannot always remove the skewness of the RT distributions, Klein Entink, van der Linden, and Fox (2009) developed the Box-Cox normal model. While the log-normal model (van der Linden, 2006) only works for normalizing positively skewed distributions, the Box-Cox transformation (Box & Cox, 1964) is widely applied to convert symmetric or (positively or negatively) skewed distributions into normal distributions:

$$T^{(v)} = \begin{cases} \dfrac{T^{v} - 1}{v}, & v \neq 0 \\ \log T, & v = 0, \end{cases} \tag{2.8}$$

where $v \in \mathbb{R}$ is a parameter that controls the degree to which the untransformed distribution is compressed. As one can see, the log transformation is included as a special case when $v = 0$. When $v \neq 0$, a different nonlinear transformation is applied to the variable with smaller vs yielding higher degrees of compression. Applying such a transformation, a Box-Cox normal model for RT can be expressed as follows (Klein Entink, van der Linden, & Fox, 2009):

$$t_{ij}^{(v)} = \beta_i - \tau_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \alpha_i^{-2}). \tag{2.9}$$

The mathematical form of this model is quite similar to the log-normal model in (2.7), except that the left-hand side of the model is replaced by RT, t_{ij} after the Box-Cox transformation. The other parameters, β_i, τ_j, ε_{ij}, and α_i, are also interpreted in a similar way but on the transformed scale.

There are two ways to apply the Box-Cox normal model in practice. One can estimate either a single v for all items or item-specific v for each item. It is expected that the item-specific vs would fit the data better than the single v, although both of them can improve model fit from the log-normal model. However, applying item-specific vs to the data results in different scales among items, thus the item and person parameters may not be directly comparable. Another drawback of the Box-Cox normal model is that it may not be appropriate for distributions with unusual shapes, such as bimodal or mixture distributions.

The semiparametric Cox proportional hazards (PH) model (Ranger & Ortner, 2012a; Wang, Fan, Chang, & Douglas, 2013) is an appealing alternative that avoids both drawbacks of the Box-Cox normal model. The Cox PH model is a widely-used modeling approach in survival analysis that concerns with the change in hazard rate over time. A specific time point of interest, named “time-to-event”, refers to the time “until an event occurs”. In the Cox PH model, hazard rate of a specific time point is modeled as the product of two components, a baseline hazard function, and an exponential function of the effect parameters that reflect how hazard rate varies with explanatory variables. In RT modeling, responding to an item is usually the event of interest, and RT is considered as the “time-to-event”. Hazard rate represents the probability of responding to an item in the next moment, which can also be conceived as the rate at which an examinee works at a specific time point. Therefore, an examinee works more intensively when his or her hazard rate is higher (Ranger & Ortner, 2012a; Wenger & Gibson, 2004). Mathematically, the Cox PH model can be expressed as follows:

$$h_{ij}(t_{ij}|\tau_j) = h_{0j}(t_{ij}) \exp(\gamma_i \tau_j), \tag{2.10}$$

where t_{ij} denotes RT, $h_{0j}(\cdot)$ is the baseline hazard function, τ_j is the speed parameter for person j, γ_i is an item-specific slope parameter that determines the increase in hazard rate. A larger γ_i indicates faster increase in hazard rate due to increase in speed.

There are at least two reasons why one would prefer the Cox PH model over the Box-Cox normal model. First, the Cox PH model is able to fit a wider variety of RT distributions due to its semiparametric nature. Second, because it employs nonparametric transformations, comparison across items is possible in the Cox PH model even though item-specific transformations are applied.

An even more generalized and flexible RT model that unifies the log-normal model, the Box-Cox model, and the Cox PH model is the linear transformation model (Ranger & Kuhn, 2013; Wang, Chang, Douglas, 2013). No specific transformation or specific error distribution is assumed in this model; instead, it only requires the transformation to be order-preserving, and the transformed RT is modeled by the weighted sum of covariates and a random error term. Assuming that the latent speed is the only covariate, this model can be expressed as follows:

$$H_i(t_{ij}) = \gamma_i \tau_j + \varepsilon_{ij}, \tag{2.11}$$

where γ_i and τ_j are defined the same as in (2.10), $H_i(\cdot)$ is an order-preserving transformation of RT distribution for item i, and ε_{ij} denotes the errors. Under this model, it can be algebraically proven that applying the Box-Cox transformation and with normal errors would yield the Box-Cox normal model. Similarly, substituting $H_i(\cdot)$ by a nonparametric transformation and a Gumbel (Gumbel, 1935, 1941) distribution for the errors results in the Cox PH model.

In addition to the Cox PH model, researchers have adopted several common parametric survival time distributions for RT modeling, including exponential, gamma, and Weibull distributions. Specifically, Scheiblechner (1979) developed an exponential distribution that models RT density using a linear combination of item and person effects:

$$f(t_{ij}) = (\tau_j + \gamma_i) \exp[-(\tau_j + \gamma_i)t_{ij}], \tag{2.12}$$

where τ_j and γ_i are the person and item speed parameters, respectively. Note that these two parameters are defined differently from those in (2.10) and (2.11). The item speed parameter can be decomposed into a weighted sum of the time intensity required by each component process. A comparable formulation with a gamma distribution is presented by Maris (1993), which is a two-parameter generalization of the exponential distribution. Verhelst, Verstralen, and Jansen's (1997) also suggested modeling RT with a gamma distribution, where speed and accuracy are considered as two complementary aspects that compose a more basic concept called mental power.

Another survival time distribution that is commonly used for RT modeling is the Weibull distribution. Roskam (1997) applied a one-parameter Weibull distribution for the marginal distribution of RT:

$$f(t) = \lambda t \exp\left(-\frac{\lambda}{2}t^2\right), \tag{2.13}$$

where $\lambda = \frac{\theta_j}{b_i\delta_j}$, θ_j and δ_j are interpreted as the mental speed and persistence of person *j*, and b_i is the item difficulty for item *i*. The definition of δ_j indicates that the probability of a correct response increases as an examinee's persistence to answer the item increases. One big difference between the Roskam's (1997) approach and the other approaches reviewed in this section is that it models the test completion time, rather than individual item RTs. Rouder, Sun, Speckman, Lu, and Zhou (2003) also introduced a model utilizing a three-parameter Weibull distribution, which is mainly applicable to the experimental paradigm (Rouder et al., 2003). In this type of test, almost the same cognitive process is required by the stimuli in each item, thus it is reasonable to assume that item characteristics do not affect the time spent on items.

To summarize, researchers have tried various distributions and models to improve model fit of the RT distributions, including normal and log-normal distributions (e.g., Thissen, 1983, van der Linden, 2006), Box-Cox normal model (Klein Entink, van der Linden, & Fox, 2009), the Cox PH model (Ranger & Ortner, 2012a; Wang, Fan, et al., 2013), linear transformation model (Ranger & Kuhn, 2013; Wang, Chang, et al., 2013), exponential (Scheiblechner, 1979), gamma (Maris, 1993; Verhelst et al., 1997), and Weibull (Roskam, 1997; Rouder et al., 2003; Tatsuoka & Tatsuoka, 1980) distributions. Comparing normal, log-normal, gamma, and Weibull distributions, Schnipke and Scrams (1999, 2002) concluded that log-normal model provides the best fit of RT distributions from both exploratory and confirmatory samples. Nonetheless, all models reviewed in this section provide meaningful interpretations of the data, though they may vary in terms of model assumptions, interpretability and flexibility. Readers are referred to Schnipke and Scrams (2002) for a more comprehensive review of other alternatives.

2.2.2 Incorporating RA for modeling RT

Similar to section 2.1.2, this section introduces models that incorporate RA-related variables for RT modeling. The relationship between speed and accuracy is still one of the most important research questions; however, instead of estimating the probability of a correct response, models in this section focus on modeling RT distributions. A recent example of this type of models is developed by Gaviria (2005), where the author specifies a double log-normal distribution for a rescaled time for a correct response:

$$\ln\left[\frac{t_{ij}-T_0}{A}\right] = -a_i(\theta_j - b_i) + \varepsilon_{ij}, \qquad (2.14)$$

where *A* is a scaling constant, T_0 represents the time required to answer an extremely easy item, and the $a_i(\theta_j - b_i)$ follows the structure of the 2PL IRT model. A log-normal distribution is chosen to model the residuals with a mean of zero and item-dependent variance σ_i^2.

Another popular model of this type is proposed by Thissen (1983). His model introduces person and item effects on RTs in a similar way as in analysis of variance:

$$\log(t_{ij}) = \mu + \tau_j + \beta_i - \rho(a_i\theta_j - b_i) + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim N(0, \sigma^2). \qquad (2.15)$$

In Thissen's (1983) model, μ indicates the average level of the population and item domain, τ_j and β_i are the slowness parameters for person *j* and item *i* respectively. Notice that although the item and person slowness parameters are termed the same as Wang and Hanson's (2005) 4PL RT model in (2.6), the interpretations are quite different. The item and person slowness parameters in (2.6) reflect how the item- and person-specific coefficients of RT affect the probability of a correct response, whereas in (2.15) they are interpreted as the main effects on $\log(t_{ij})$. The fourth term on the right-hand side of the equation regresses the log odds of a correct response on $\log(t_{ij})$ following the 2PL model, where a coefficient ρ controls the magnitude of association between the two. Essentially, this modeling approach incorporates the impact of IRT model parameters on the RT modeling.

Ferrando and Lorenzo-Seva (2007) extend this model to a different response parameter structure, where $a_i\theta_j - b_i$ is replaced by $\sqrt{a_i^2\left(\theta_j - b_i\right)^2}$. Both Thissen's (1983) and Ferrando and Lorenzo-Seva's (2007) models imply that when $\rho > 0$, examinees with higher ability tend to work fast than lower ability examinees, while the opposite relationship is implied when $\rho < 0$. Therefore, these two models are more flexible than Wang and Hanson's (2005) in that the association between ability and speed could be either positive or negative. Ranger and Kuhn (2012) proposed an extension of Ferrando and Lorenzo-Seva's (2007) study with the absolute distance between ability and difficulty, denoted by $\left|\theta_j - b_i\right|$.

2.3 Joint Modeling of RT and RA

In the first two sections of this chapter, the theoretical foundations of joint modeling of RT and RA are reviewed, in terms of IRT models and their variations to include RT as collateral information, as well as standard and extended RT models incorporating parameters from IRT models. This section describes several modeling frameworks for joint modeling of RT and RA that have been commonly utilized to analyze test data with a time limit.

Before detailed introduction to each framework, it is worth noticing that some researchers from the cognitive psychology field have adopted a different strategy for separate analysis of RT and RA (Klein Entink, Kuhn, Hornke, & Fox, 2009). For instance, researchers have examined the impact of item characteristics on item difficulty and RT separately (e.g., Embretson, 1998; Gorin, 2005; Primi, 2001). Although such strategy provides insights about how RT and RA vary independently, it does not permit modeling the relationship between them. On the contrary, joint modeling of RT and RA facilitates the simultaneous estimation of IRT and RT model parameters, as well as the investigation of the relationship between RT and RA. Therefore, this section focuses on reviewing methods for joint modeling of RT and RA.

Studies reviewed in this section mainly aim at modeling two sources of variabilities that cause examinees to respond in different manners: between-subject and within-subject differences. Between-subject differences are of interest to a lot of models introduced in section 2.3.1, where the conditional independence between RT and RA is assumed. In these models, differences in examinees' response behaviors are attributed to differences in their ability and speed, as well as item parameters. It is assumed that examinees respond to the items with a constant ability and a constant speed across the test (e.g., Goldhammer & Kroehne, 2014; Meng et al., 2015; van der Linden, 2009). In other words, there is no within-subject difference as of how an examinee interacts with the items. While the assumption of a constant ability is more widely acceptable, assuming a constant speed might be less viable in real testing scenarios where examinees can switch problem-solving strategies between items. As such, studying within-subject differences provides the opportunity to analyze underlying psychological processes of an examinee's test-taking behaviors. For instance, examinees may use different cognitive strategies to solve the items (van der Maas & Jansen, 2003), fake on some items (Holden & Kroner, 1992), or demonstrate item pre-knowledge (McLeod, Lewis, & Thissen, 2003). Other effects, including learning and practice (Carpenter, Just, & Shell, 1990), fatigue and motivation

(Mollenkopf, 1950), can also be examined by modeling within-subject differences. Most approaches to modeling conditional dependence between RT and RA in sections 2.3.2 and 2.3.3 are proposed to analyze within-subject differences explicitly.

In the following sections, methods for joint modeling of RT and RA are categorized by different assumptions of conditional dependence between responses and RTs, as suggested in Ranger and Ortner (2012b). Mathematically, the joint distribution of RT and RA can be expressed as follows:

$$f(y_{ij}, t_{ij} | \theta_j, \tau_j, \delta_i, \gamma_i), \tag{2.16}$$

where y_{ij} represents an item response from person *j* to item *i*, t_{ij} is the RT associated with response y_{ij}, θ_j and τ_j denote the latent ability and speed parameters for person *j*, δ_i and γ_i indicate the item parameters for item *i* in IRT and RT models respectively. Based on this expression, one can choose to model the joint distribution of RT and RA directly if they are assumed to covary following a certain functional form. Along these lines, models have been developed with specific scoring rules that reward or penalize certain responses made within certain time (e.g., Dennis & Evans, 1996; Maris & van der Maas, 2012; van der Maas & Wagenmakers, 2005). These models present another distinctive line of research, therefore are not reviewed in detail in this literature review.

Other than applying the scoring rules for modeling the joint distribution of responses and RTs directly, one can choose among three different approaches to modeling the joint distribution of RT and RA. First, equation (2.16) can be decomposed into two marginal distributions for RT and RA, assuming that responses and RTs are conditionally independent given the two latent traits and respective item parameters:

$$f(y_{ij}, t_{ij} | \theta_j, \tau_j, \delta_i, \gamma_i) = f(y_{ij} | \theta_j, \tau_j, \delta_i) f(t_{ij} | \theta_j, \tau_j, \gamma_i). \qquad (2.17)$$

Further simplifications have been advocated by Thissen (1983) and van der Linden (2007):

$$f(y_{ij}, t_{ij} | \theta_j, \tau_j, \delta_i, \gamma_i) = f(y_{ij} | \theta_j, \delta_i) f(t_{ij} | \tau_j, \gamma_i), \qquad (2.18)$$

which is the common definition of conditional independence assumption. Specifically, the conditional independence assumption states that responses solely depend on the latent ability and IRT model parameters whereas RTs solely depend on the latent speed and RT model parameters, and y_{ij} and t_{ij} are conditionally independent of each other given the associated parameters. Such an assumption is often applied when modeling item responses and RTs (e.g., Klein Entink, van der Linden, & Fox, 2009; Thissen, 1983; van der Linden, 2007; Wang, Fan, et al., 2013), which is also supported by empirical evidence in psychological research (e.g., Kennedy, 1930; Tate, 1948). In the present study, this definition of conditional independence assumption is adopted due to its popularity in this line of research, despite the existence of other possible alternative definitions.

Second, the joint distribution of RT and RA can be factored as a conditional distribution for one source of information and a marginal distribution for another (Bloxom, 1985), when the conditional independence assumption is violated. Note that the conditional dependence of interest in the present study can be considered as within-item dependence that exists between the item response and RT for the same item. This should be distinguished from other types of dependence, such as the conditional dependence between item responses or between RTs for different items, which are indeed between-item dependence. In particular, the following factorization has been advocated in some studies (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2017; Goldhammer, Steinwascher, Kroehne, & Naumann, 2017; Ingrisone, 2008):

$$f\left(y_{ij}, t_{ij} \middle| \theta_j, \tau_j, \delta_i, \gamma_i\right) = f\left(y_{ij} \middle| t_{ij}, \theta_j, \tau_j, \delta_i, \gamma_i\right) f\left(t_{ij} \middle| \theta_j, \tau_j, \delta_i, \gamma_i\right). \qquad (2.19)$$

This formulation implies that an observed response to an item relies on the RT spent on this specific item. (2.19) can also be simplified as follows assuming responses and RTs are dependent on different latent traits only:

$$f\left(y_{ij}, t_{ij} \middle| \theta_j, \tau_j, \delta_i, \gamma_i\right) = f\left(y_{ij} \middle| t_{ij}, \theta_j, \delta_i\right) f\left(t_{ij} \middle| \tau_j, \gamma_i\right). \qquad (2.20)$$

Alternatively, one can factor (2.16) into the marginal distribution of responses and the conditional distribution of RTs given the associated responses (e.g., Bolsinova & Tijmstra, 2016; van der Linden & Glas, 2010). This approach has an opposite assumption that the RT on an item depends on the response made to this item:

$$f\left(y_{ij}, t_{ij} \middle| \theta_j, \tau_j, \delta_i, \gamma_i\right) = f\left(y_{ij} \middle| \theta_j, \tau_j, \delta_i, \gamma_i\right) f\left(t_{ij} \middle| y_{ij}, \theta_j, \tau_j, \delta_i, \gamma_i\right), \qquad (2.21)$$

which, again, can be simplified as

$$f\left(y_{ij}, t_{ij} \middle| \theta_j, \tau_j, \delta_i, \gamma_i\right) = f\left(y_{ij} \middle| \theta_j, \delta_i\right) f\left(t_{ij} \middle| y_{ij}, \tau_j, \gamma_i\right). \qquad (2.22)$$

Based on the three approaches described above, models that jointly harness the benefits from RTs and responses are classified and elaborated for each approach respectively with respect to model structure, parameter estimation, and how the speed-accuracy tradeoff is represented in the framework.

2.3.1 Conditionally Independent RTs and responses

One straightforward factorization of the joint distribution of RT and RA is realized via the conditional independence assumption as demonstrated in (2.18). A model that adopts the conditional independence assumption is the drift diffusion model (Ratcliff, 1978). This model focuses on the underlying response processes based on a diffusion process. It is assumed that when two alternative options are presented to a subject, the evidence of both options accumulates over time by a Wiener process. A decision is made when the information accumulates to a certain boundary. In this model, responses and RTs are conditionally independent given the latent ability and speed. Although such a model has mostly been applied in experimental psychology for within-individual data, it has been employed for analyzing cross-sectional data composed of responses and RTs (Molenaar, Tuerlinckx, & van der Maas, 2015c; Tuerlinckx & De Boeck, 2005; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011; Vandekerckhove, 2009; Wagenmakers, 2009).

In addition to the drift diffusion model, researchers have developed some other approaches that also advocate the conditional independence assumption between RTs and responses (e.g., Ranger & Kuhn, 2014a; Van Breukelen, 2005; van der Linden, 2007), with van der Linden’s (2007) hierarchical framework as the most prominent framework (Ranger & Kuhn, 2014b). His model describes two sources of information in a two-level model, where an IRT model and a RT model are specified for responses and RTs respectively on the first level. On the second level, rather than imposing a direct mathematical function between RT and RA, the item and person parameters for two models on the first level are assumed to covary. For item responses, van der Linden (2007) employed the 3PL model as presented in (2.3); in fact, any IRT model can be used in modeling item responses. For the RT model, van der Linden's (2006) log-normal model as shown in (2.7) is employed. Again, this framework is flexible for other RT models reviewed in section 2.2 as well. On top of the two first-level models for RT and RA, two second-level multivariate normal distributions are further specified for item and person parameters respectively. Assuming a 3PL model for item response modeling, the mean vector and covariance matrix for item parameters are

$$\boldsymbol{\mu}_I = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta), \tag{2.23}$$

and

$$\boldsymbol{\Sigma}_I = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix}. \tag{2.24}$$

Similarly, the mean vector for person parameters is defined as

$$\boldsymbol{\mu}_P = (\mu_\theta, \mu_\tau), \tag{2.25}$$

and the covariance matrix is

$$\boldsymbol{\Sigma}_P = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\tau\theta} & \sigma_\tau^2 \end{pmatrix}. \tag{2.26}$$

In the above mean vectors and covariance matrices, the subscript *I* refers to item-related parameters and subscript *P* refers to person-related parameters.

This hierarchical modeling framework is similar to Thissen's (1983) model in (2.15) because they both assume RTs follow a log-normal distribution and focus on modeling RT with item and person effects. Yet they are different from at least four aspects. First, as indicated by the name, time intensity parameter β_i in (2.7) is an estimate of the average time used on the item, whereas the item slowness parameter β_i in (2.15) represents a deviation on the average RT of a specific item from μ, the overall average RT of the population and item domain. Second, the person effect τ_j in (2.7) is termed a person speed parameter rather than person slowness parameter as the symbol precedes it changes from positive to negative. Third, no direct functional relationship is imposed between the log odds of a correct response and the logRT in the RT model. Rather, the relationship between speed and accuracy is taken care of at the second level. Lastly, the error term in (2.7) follows a normal distribution with item-specific variance term, instead of a constant variance across all items.

van der Linden’s (2007) hierarchical framework provides a flexible and readily interpretable modeling framework for joint modeling of speed and accuracy. Built on this framework, IRT and RT models introduced in sections 2.1 and 2.2 can be utilized for more modeling options. For instance, Klein Entink, van der Linden, and Fox (2009) proposed a straightforward extension with the Box-Cox normal model for RT modeling. Some other RT models have also been embedded as a first level model in the hierarchical framework, such as the Cox PH model (Ranger & Kuhn, 2014a; Wang, Fan, et al., 2013) and the linear transformation model (Wang, Chang, et al., 2013). Other approaches assume a more complex underlying responding mechanism, such as the race model (Ranger, Kuhn, & Gaviria, 2015), where actual responses are determined by competing stochastic processes for possible response options.

In addition to the modifications of the RT model, extensions have been made to accommodate more effects from covariates, examinee clustering, multiple sources of responses and RTs, as well as non-normal logRT distribution. Specifically, covariates, such as item characteristics, are included to model variability in item parameters (Klein Entink, Kuhn, et al., 2009; Loeys, Rosseel, & Baten, 2011). Another study from Klein Entink, Fox, and van der Linden (2009) incorporated a multilevel structure for groups of examinees and included covariates at both person and group levels. To model examinees' test performance as well as feedback behaviors, a multivariate hierarchical model was developed to model four latent traits from both sources (Fox, Klein Entink, & Timmers, 2014). When the normal assumption of logRT distribution is violated, Molenaar and Bolsinova (2017) proposed a model for non-normal logRT distribution to distinguish non-normality due to heteroscedastic residual variances and skewed latent speed. Moreover, a generalized linear factor model (Molenaar, Tuerlinckx, & van der Maas, 2015b) was proposed to unify several common modeling approaches for responses and RTs, including van der Linden (2007), Fox, Klein Entink, and van der Linden (2007), Klein Entink, Fox, and van der Linden (2009), and Glas and van der Linden (2010). The only mild restriction is that the item model at the second level is omitted.

Most studies introduced in this section focus on modeling test data from multiple sources to analyze the relationship between speed and accuracy via simulation studies and empirical data analyses. In particular, simulation studies have been conducted to evaluate the sensitivity of parameter recovery to various manipulated factors, such as sample size (Fox et al., 2014; Kang, 2016; Ranger &

Kuhn, 2014b; Suh, 2010; Wang, Fan, et al., 2013), test length (Fox et al., 2014; Kang, 2016; Molenaar et al., 2015b; Suh, 2010; Wang, Fan, et al., 2013), the correlation between latent speed and ability (Klein Entink, 2009; Patton, 2015; Suh, 2010), the distribution for modeling RT (Kang, 2016; Patton, 2015; Molenaar & Bolsinova, 2017; Wang, Fan, et al., 2013). Common evaluation criteria include bias, empirical standard error (SE), mean square error (MSE), root mean square error (RMSE), correlation and 95% confidence interval or credible interval.

In general, increasing sample size and test length both yield more accurate item and person parameter estimates since more information can be borrowed from RTs as expected (e.g., Kang, 2016; Marianti, 2015; Suh, 2010; Wang, Fan, et al., 2013). Similarly, higher correlation between latent ability and speed also results in higher recovery accuracy for both person and item parameters. With respect to person parameters, the precision of ability parameter estimates increases as the correlation between RTs and responses increases (Klein Entink, 2009; Patton, 2015), though the improvement may not be practically meaningful for correlation less than .5 (Ranger, 2013; van der Linden, Klein Entink, & Fox, 2010). In other words, the more collateral information contained in RTs, the more accurate the ability estimates. Further, the effect of incorporating RTs on ability estimates seems to vary systematically along the ability scale. That is, although the effect of incorporating RTs is relatively small for examinees located near the population mean, the estimation accuracy of ability estimates improves considerably for examinees near the two ends of the latent continuum, especially when correlation is high (Klein Entink, 2009; Molenaar & Bolsinova, 2017; Patton, 2015).

On the other hand, more accurate item parameters from IRT model can be obtained by joint modeling of RT and RA within the hierarchical framework (Ranger & Kuhn, 2012; van der Linden et al., 2010). Kang (2016) found that increasing sample size and the correlation between ability and speed reduces the bias and MSE of item parameters. Klein Entink (2009) specifically analyzed the impact of incorporating RT on item discrimination parameter and concluded that the MSE of item discrimination decreases as the correlation between RTs and responses increases, and, interestingly, as item discrimination increases.

In addition to test length, sample size, and the correlation between speed and ability, researchers have also compare different RT models in terms of the sensitivity of item and person parameters to RT model misspecification. Patton (2015) compared the log-normal, the Weibull and the Box-Cox normal model and found that ability, speed, and correlation estimates are robust to misspecification of the RT model under the hierarchical modeling framework. Utilizing the Cox PH model for RT modeling, Wang, Fan, et al., (2013) examined parameter recovery for exponential, Weibull, and nonmonotone baseline hazard function, and concluded that the model can always be accurately recovered. When comparing different modeling frameworks, van der Linden’s (2009) hierarchical framework outperforms Thissen’s (1983) and Wang and Hanson’s (2005) models (Suh, 2010), but the speed-accuracy response model (SARM; Maris & van der Maas, 2012) seems to provide higher model-based reliability than the hierarchical framework (van Rijn & Ali, 2017).

Simulation studies offer a means to examine model parameter recovery under certain conditions, whereas analyzing empirical datasets enables researchers to explore true model parameters in real testing scenarios. One parameter of particular interest in joint modeling of RT and RA is the correlation between speed and ability. Surprisingly, researcher have reported both strong and weak correlations in both positive and negative directions. For instance, Klein Entink, Fox, and van der Linden (2009) analyzed data from National World Assessment Test (NAW-8) and found a correlation of -.76. Similarly, Klein Entink, Kuhn, et al. (2009) used data from a large-scale figural reasoning ability test and reported a strongly negative (-.61) correlation. Other examples of negative correlations include Roberts and Stankov (1999) and van der Linden and Fox (2015). On the contrary, researchers have also found a correlation of .65 from Amsterdam Chess Test Data (Fox & Marianti, 2016), .3 from American Institute of Certified Public Accountants (AICPA) certification program (van der Linden, 2007), among others (e.g., Klein Entink, 2009; Marianti, 2015; Wang & Xu, 2015). van der Linden et al. (1999) even reported .035 correlation using data from Arithmetic Reasoning Test in the Armed Services Vocational Aptitude Battery (ASVAB) item bank.

In fact, these seemingly contradictory findings often inform the nature of the tests. A negative correlation between speed and ability usually indicates that the test is non-speeded, such that high ability examinees have better time management during the test (Klein Entink, Fox, & van der Linden, 2009), whereas a positive correlation may suggest a speeded test. With respect to the correlations between time intensity and other item parameters, the correlations are generally negative for the discrimination parameter in both IRT and RT models (e.g., Fox & Marianti, 2016; Klein Entink, Kuhn, et al., 2009), and positive for item difficulty (e.g., Fox &

Marianti, 2016; Klein Entink, Kuhn, et al., 2009; Marianti, 2015; van der Linden, 2007).

Thus far, models that are based on van der Linden's (2007) hierarchical framework have focused on modeling between-subject differences. That is, examinees are assumed to adopt the same ability and the same speed for answering all items on the test. Moreover, van der Linden and Glas (2010) noted that the conditional independence assumption between responses and RTs only holds when the speed and ability of an examinee keep constant throughout the entire test. As a result, any fluctuations on speed, and therefore fluctuations on ability, would lead to violations of conditional independence assumption between the response and RT for a specific item. To identify sources of misfit and capture within-subject variations, methods have been proposed for evaluating model fit (Marianti, 2015; Ranger & Kuhn, 2014b; Ranger, Kuhn, & Szardenings, 2017), person fit (Fox & Marianti, 2017; Marianti et al., 2014), as well as conditional independence assumption (van der Linden & Glas, 2010; Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016). Potential misfit due to within-subject fluctuations can be modeled using methods introduced in sections 2.3.2 and 2.3.3 by relaxing the assumption of conditional independence between responses and RTs.

2.3.2 Distinguishing Fast and Slow Responses

Conditional independence between responses and RTs can be violated in different ways. A most straightforward way to account for this dependence is to add residual correlations between responses and the associated RTs as in Ranger and Ortner (2012b) and Meng et al. (2015). However, violations of conditional independence between responses and RTs may not always appear as residual correlations among all examinees. In fact, the residual correlation between responses and RTs for an item may cancel out at the population level if a negative residual correlation exists between RTs and responses for one group of examinees and a positive one for another group depending on their ability level (Bolsinova & Tijmstra, 2016). Moreover, this violation might not only arise from different ability levels, but also heterogeneous response processes. Such response processes may be due to variable ability and speed (e.g., Partchev & De Boeck, 2012), as well as different item characteristics (e.g., Bolsinova, De Boeck, & Tijmstra, 2017).

One specific type of violation that researchers have been interested in is how an individual examinee's pace can be different on items throughout the test. These differences may reflect different test-taking behaviors, such as rapid-guessing behaviors (Wang & Xu, 2015), item pre-knowledge (Lee & Wollack, 2017), or dual response processes (Goldhammer et al., 2014, 2015, 2017). One modeling option is to utilize an IRT model with a binomial tree structure to distinguish fast and slow responses, namely the IRTree model (De Boeck & Partchev, 2012; DiTrapani, Jeon, De Boeck, & Partchev, 2016; Partchev & De Boeck, 2012). This approach disentangles the fast and slow classes by splitting RTs based on median of RTs within-person or within-item, and models RA depending on which class a specific response falls into. As such, class sizes are arbitrarily chosen by researchers for two classes. Moreover, dichotomizing continuous RTs reduces the information that could have been used in joint modeling of responses and RTs.

An appealing alternative to the IRTree model is mixture modeling with different latent classes representing item and person properties for different speed (Lee & Wollack, 2017; Marianti, 2015; Molenaar, Bolsinova, Rozsa, & De Boeck, 2016; Molenaar, Bolsinova, & Vermunt, 2016; Molenaar, Oberski, Vermunt & De Boeck, 2016; Wang & Xu, 2015). Moreover, within-subject differences are accounted for by incorporating a person- and item-specific class membership, such that examinees' speed varies from item to item. Researchers have proposed several parametric mixture modeling approaches for differential latent ability in fast and slow modes (Molenaar, Bolsinova, Rozsa, & De Boeck, 2016), rapid guessing behavior (Wang & Xu, 2015), and examinees with item pre-knowledge (Lee & Wollack, 2017). Specifically, Molenaar, Bolsinova, Rozsa, and De Boeck (2016) model examinees in different modes with the same functional forms for responses and RTs, but the item and person parameters are different for fast and slow modes. Whereas in Wang and Xu's (2015) and Lee and Wollack's (2017) studies, examinees are assumed to follow different IRT and RT models in different classes. Marianti (2015) further developed a generalized mixture dynamic speed model for examinees with stationary and non-stationary speed, accounting for both between-subject and within-subject differences by dividing a test into blocks of items. To reduce parameter estimation bias and avoid detecting spurious classes when RT distributions are not correctly specified, a semi-parametric remedy was proposed where RTs are categorized into an arbitrary number of categories (Molenaar, Bolsinova, & Vermunt, 2016). Rather than allowing latent ability to be modeled separately for different modes or classes, Molenaar, Oberski, et al. (2016) employed a hidden Markov modeling framework for modeling variations in speed and ability. It also assumes item parameters to be different across states, but a constant speed and a constant ability are assumed throughout the test.

Both IRTree modeling and mixture modeling classify examinees into discrete latent classes based on information from responses and RTs. However, some researchers argued that the impact of RTs on the measurement properties of the IRT model is likely to be continuous (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2016; Fox & Marianti, 2016). Following this logic, researchers have proposed models that include RT effects into IRT modeling and decompose the joint distribution of responses and RTs as demonstrated in (2.18) (Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2017; De Boeck, Chen, & Davison, 2017; Goldhammer et al., 2017; Ingrisone, 2008; Wang, 2006).

Most of these models are based on IRT and RT models introduced in the previous sections of this chapter. For instance, Wang (2006) proposed a joint model of responses and RTs by employing an IRT model similar to Wang and Hanson’s (2005) 4PL-RT model and one-parameter Weibull distribution as shown in (2.13). Wang’s (2006) formulation reflects a pacing strategy discussed in Wang and Zhang (2006), that examinees tend to spend more time on items with similar difficulty levels as their ability levels. Based on Wang’s (2006) model, Ingrisone (2008) applied a two-parameter Weibull distribution for RT modeling, allowing not only the scale, but also the shape of the RT distribution to vary. Employing a log-normal RT model (van der Linden, 2006), De Boeck et al. (2017) incorporated logRT into response modeling explored the effects of spontaneous speed vs. imposed speed in a test with item-specific time constraints.

In addition to incorporating RT or logRT directly, the effects of standardized residual RT have also been investigated thoroughly on the item difficulty and discrimination parameters (Bolsinova, De Boeck, & Tijmstra, 2017), and to be item-specific, person-specific, or both (Bolsinova, Tijmstra, & Molenaar, 2017). More generalized linear modeling frameworks have been developed to incorporate random and fixed effects (Goldhammer et al., 2017; Klein Entink, 2009) and flexible cross-relation function that specifies the relationship between speed and ability (Molenaar, Tuerlinckx, & van der Maas, 2015a).

Results from these studies suggest that the conditional independence assumption between responses and RTs is often violated (e.g., Bolsinova & Tijmstra, 2016; Goldhammer et al., 2017; Molenaar, Bolsinova, Rozsa, & De Boeck, 2016), which indicates the lack of measurement invariance of latent ability and speed (De Boeck et al., 2017). The residual correlations between responses and RTs are found to be negative for most easy items, less negative or positive for difficult items (Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer et al., 2014; Molenaar, Bolsinova, Rozsa, & De Boeck, 2016; Partchev & De Boeck, 2012). In other words, for difficult items, spending more time increases the probability of a correct response, while for easy items the probability of a correct response decreases.

Most studies conclude that incorporating RT effects in IRT models or employing mixture models for distinguishing response processes with variable speed improves model fit (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Molenaar, Oberski, et al., 2016), yet the choice of best-fitting model is subject to the source of conditional dependence (Bolsinova, Tijmstra, & Molenaar, 2017). For the mixture modeling approach with fast and slow classes, researchers have reported the percentage of fast latent class of 38%-44% based on the semiparametric mixture model (Molenaar, Bolsinova, & Vermunt, 2016), and 23% at the initial state in the hidden Markov model (Molenaar, Oberski, et al., 2016), indicating that most examinees only produce fast responses on less than half of the items.

2.3.3 Distinguishing Correct and Incorrect Responses

The relationship between RTs and RA has long been of interest to researchers in both psychometrics and cognitive psychology fields. Descriptive studies of RTs generally found that examinees tend to spend more time on items they miss than those they answer correctly (e.g., Bergstrom, Gershon, & Lunz, 1994; Chang, 2007; Hornke, 2000; Lee, 2007; Swanson, Featherman, Case, Luecht, & Nungester, 1999), and RTs for correct and incorrect responses do not seem to follow the same distribution (Lee, 2007).

A few studies have been proposed recently to detect the violations of conditional independence between responses and RTs due to different RT distributions for correct and incorrect responses (Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; Glas & van der Linden, 2010; van der Linden & Glas, 2010). Specifically, Glas and van der Linden (2010) and van der Linden and Glas (2010) proposed a Lagrange multiplier test by including a location shift parameter λ_i for item i, where all other parameters are defined the same as in (2.7):

$$\log(t_{ij}) = \beta_i + y_{ij}\lambda_i - \tau_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.27}$$

By testing the null hypothesis of $\lambda_i = 0$, the conditional independence assumption can be examined. However, the statistical properties of this test are dependent on whether the same conditional independence assumption holds for all other items. Moreover, this test is appropriate when the conditional dependence only arises from the difference in the location of the logRT distributions. Bolsinova and Maris (2016) noticed these drawbacks and developed non-parametric Kolmogorov-Smirnov tests that are applicable when the summed score of items on the test is a sufficient statistic for latent ability. They mainly considered three types of violations, the mean of the logRT distribution varies for correct and incorrect responses and for different ability levels, and the variance of the logRT distribution varies for correct and incorrect responses. The first two types of violations can be expressed by replacing $y_{ij}\lambda_i$ in (2.27) by $\frac{y_{ij}\lambda_i}{\alpha_i}$ and $\frac{\kappa\theta_j(2y_{ij}-1)}{\alpha_i}$:

$$\log(t_{ij}) = \beta_i + \frac{y_{ij}\lambda_i}{\alpha_i} - \tau_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.28}$$

and

$$\log(t_{ij}) = \beta_i + \frac{\kappa\theta_j(2y_{ij} - 1)}{\alpha_i} - \tau_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_i^{-2}), \tag{2.29}$$

where κ reflects how the location of the logRT distribution shifts between correct and incorrect responses for examinees with different ability levels. For instance, $\kappa > 0$ indicates that correct responses are faster for high-ability students, whereas incorrect responses are faster for low-ability students. Relaxing the assumption of a sufficient statistic, Bolsinova and Tijmstra (2016) proposed three posterior predictive checks for the same two types of violations as shown in (2.28) and (2.29), and further considered variations of the logRT distributions which are item-dependent and person-dependent. van Rijn and Ali (2017) also mentioned a possible extension of (2.27) where both time intensity and time discrimination parameters vary depending on item responses. It is concluded, however, that most items only demonstrate location shift between response correctness in the illustration examples (Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; van der Linden & Glas, 2010).

Moreover, Bolsinova and Tijmstra (2017) extended van der Linden’s (2007) hierarchical model to contain cross-loadings between ability and RTs, which enables one to utilize more collateral information from ability for RT modeling. The authors manipulated test length, the correlation between speed and ability, and standard deviation (SD) of the cross-loadings and compared the performance of the two-parameter normal ogive model, van der Linden’s (2007) simple structure hierarchical model, and the proposed model. Findings from this study suggest that adding cross-loadings between ability and RTs further improves the estimation of ability based on the simple structure hierarchical model, and that the cross-loadings indicate the differences among examinees with different ability levels with respect to their speed. Magnus, Willoughby, Blair, and Kuhn (2017) also conducted a study to analyze empirical data with similar model structures. They concluded that the inclusion of RT information improved measurement precision of the ability estimates, particularly at the extreme levels. Bolsinova and Tijmstra (2017) further put forward a future research question to investigate specific item characteristics as a substantive explanation for different patterns between ability and speed.

In the next section, possible model estimation methods are summarized for joint modeling of RT and RA, with a focus on the technical details of Bayesian estimation used in the present study.

2.4 Model Estimation

There are two main model estimation frameworks in statistics, namely the frequentist inference and the Bayesian inference. From a frequentist point of view, data are a random sample that can be replicated with unknown but fixed parameters. In other words, the parameters remain constant in the repeatable data generation process. On the other hand, the Bayesians tend to think that the observed data are fixed and considered a realized sample of an underlying population. Parameters are random, instead of fixed, and are described probabilistically. Another major difference between the two is that Bayesian statisticians use a prior distribution to express the probability of the model parameters, reflecting the belief or hypothesis of the distribution of model parameters before collecting any data. Frequentists, however, do not rely on a prior distribution and only use probability to describe observed and unobserved data. With respect to estimation, Bayesian inference may be more computationally intensive compared to frequentist inference due to complex posterior distributions.

Most studies reviewed in this chapter apply the two estimation frameworks for model parameter estimation. Several popular software packages have been used for maximum likelihood estimation, a common method from the frequentist perspective, including M*plus* (Muthén & Muthén, 2007) and LatentGOLD (Vermunt & Magidson, 2013). Whereas for the Bayesians, researchers have employed Markov chain Monte

Carlo (MCMC) methods implemented in JAGS (Plummer, 2015), WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003), OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009), and Stan (Gelman, Lee, & Guo, 2015).

In the proposed study, the Bayesian inference is chosen for four reasons. First, Bayesian inference via MCMC methods is more flexible than the frequentist estimation methods implemented in the currently available software packages in terms of model structure (e.g., Bustamante, Nielsen, & Hartl, 2003). Second, when dealing with low-information data, the Bayesian approach seems to be able to achieve better accuracy and coverage compared to the maximum likelihood approach (e.g., Beerli, 2006). Third, the maximum likelihood estimation method is subject to different types of convergence issues, such as singularity of the information matrix and local maxima, which can be avoided in Bayesian estimation by using different priors and drawing samples from the posterior distributions. Fourth, even if diffuse priors are used, the Bayesian inference has practical advantages in that the person and item parameter estimates in the IRT model are restricted to a reasonable range (e.g., Lord, 1986). In the following sections, common sampling methods and model convergence diagnosis in Bayesian estimation are summarized.

2.4.1 Introduction to Bayesian Inference

Bayesian statistical inference uses Bayes' theorem to update the prior belief about parameters when more data becomes available. The Bayes' theorem is expressed as follows:

$$P(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{X})}, \tag{2.30}$$

where $\boldsymbol{\theta}$ represents the parameters of interest and $\boldsymbol{X}$ indicates the data. On the right-hand side of (2.30), $P(\boldsymbol{X}|\boldsymbol{\theta})$ in the numerator is defined as the likelihood of observing the available data given the parameters, $P(\boldsymbol{\theta})$ is the prior probability distribution of the parameters, and $P(\boldsymbol{X})$ is the marginal probability of the data. The Bayes' theorem states that some mathematical operations of the three terms yield the posterior distribution of the parameters given currently available data and prior distributions, as shown on the left-hand side of (2.30). As Bayesians treat data as fixed, $P(\boldsymbol{X})$ is in fact a constant. Another nice property of any probability density function is that it integrates to one over the entire space. Taking these two properties into account, the Bayes' theorem can be simplified as follows:

$$Posterior \propto Likelihood \times Prior. \tag{2.31}$$

Therefore, the posterior distribution of the parameters given the data are proportional to the product of likelihood function and the prior distribution of the parameters.

2.4.2 Markov Chain Monte Carlo Methods

The key object of Bayesian inference is the posterior distribution of the model parameters. For simple statistical models where a closed form solution exists for the posterior distribution, parameter estimates can be solved analytically. However, for complex models with non-closed form solution and high-dimensional parameter space, sampling based estimation procedures can be applied to obtain parameter estimates. A class of common sampling based estimation procedures is the MCMC methods, which is quite flexible in terms of the shape of the posterior distributions as well as the number of parameters.

The MCMC methods construct a Markov chain based on samples from a probability distribution and approximate the target posterior distribution better as the number of iterations increases (Gelman, Carlin, Stern, & Rubin, 2003). This stochastic process converges to an equilibrium distribution, which is considered approximately equal to the target posterior distribution. In fact, it has been demonstrated that the target distribution can be approximated with any accuracy as the number of iterations approaches infinity (Robert & Casella, 1999). The Markov property ensures that the parameter estimates at the next iteration are independent of any previous iterations and only dependent on the current iteration.

There are several common MCMC sampling methods, including the Gibbs sampler (Geman & Geman, 1984), the Metropolis sampler (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), and the Metropolis-Hastings (M-H) sampler (Hastings, 1970). The Gibbs sampler is the simplest MCMC method that requires conditionally conjugate models, while the Metropolis sampler is applicable to models that are not conditionally conjugate, which is further generalized to the M-H sampler for asymmetric proposal distribution.

An important assumption of the Gibbs sampling algorithm is that the conditional distributions of all parameters can be specified. Based on this assumption, a complex multivariate posterior distribution from which it is hard to draw samples can then be decomposed into simpler univariate distributions, conditioning on other model parameters, which is easier to sample from. Assuming three parameters of interest $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, the full conditional distributions can be specified as $p(\theta_1|\theta_2, \theta_3, \boldsymbol{X}), p(\theta_2|\theta_1, \theta_3, \boldsymbol{X}), p(\theta_3|\theta_1, \theta_2, \boldsymbol{X})$. After providing the algorithm with some arbitrary starting values $\boldsymbol{\theta}^{\mathbf{0}} = (\theta_1^0, \theta_2^0, \theta_3^0)$, the Gibbs sampler proceeds with drawing samples from each full conditional distribution at every iteration i ($i \geq 1$) as follows:

(1) Sample θ_1^i from $p\left(\theta_1 | \theta_2^{i-1}, \theta_3^{i-1}, \boldsymbol{X}\right)$.

(2) Sample θ_2^i from $p\left(\theta_2 | \theta_1^i, \theta_3^{i-1}, \boldsymbol{X}\right)$.

(3) Sample θ_3^i from $p\left(\theta_3 | \theta_1^i, \theta_2^i, \boldsymbol{X}\right)$.

Steps (1) to (3) are repeated until the chains converge. While the Gibbs sampler requires that all parameters have closed form full conditional distributions, this is not always the case. A more generalized MCMC algorithm is needed when one or more parameters do not have closed form full conditional distributions, such as the M-H sampler. Rather than drawing samples from the full conditional distributions sequentially, the M-H algorithm utilizes a proposal distribution of the parameters to determine whether to accept or reject the proposed new state. Suppose that there are three parameters as before, the following steps are carried out at every iteration i ($i \geq$ 1):

(1) Sample a proposal draw θ_1^* from a proposal distribution $g\left(\theta_1^* \middle| \theta_1^{i-1}\right)$.

(2) Substitute θ_1^* in $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^{i-1}, \theta_3^{i-1})$ and calculate the ratio

$$r = \frac{p(\boldsymbol{\theta}^* | \boldsymbol{X}) g(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{\boldsymbol{i-1}})}{p\left(\boldsymbol{\theta}^{\boldsymbol{i-1}} \middle| \boldsymbol{X}\right) g(\boldsymbol{\theta}^{\boldsymbol{i-1}} | \boldsymbol{\theta}^*)}.$$

(3) Accept the proposal draw θ_1^* with probability of min(r, 1) and update $\boldsymbol{\theta}^{\boldsymbol{i}} = (\theta_1^*, \theta_2^{i-1}, \theta_3^{i-1})$, otherwise reject and retain $\boldsymbol{\theta}^{\boldsymbol{i}} = (\theta_1^{i-1}, \theta_2^{i-1}, \theta_3^{i-1})$.

(4) Repeat steps (1) to (3) for θ_2 and θ_3.

As a special case of the M-H algorithm, the Metropolis algorithm essentially applies the same sampling algorithm to symmetrical proposal distributions, so that the ratio could be simplified as $r = \frac{p(\boldsymbol{\theta}^{*}|\boldsymbol{X})}{p\left(\boldsymbol{\theta}^{\boldsymbol{i}-\mathbf{1}}\middle|\boldsymbol{X}\right)}$. However, the use of asymmetric proposal distributions usually speeds up model convergence and thus the M-H algorithm outperforms the Metropolis algorithm in terms of computation efficiency (Gelman et al., 2003).

2.4.3 Convergence Diagnosis

Evaluating chain convergence in monitoring the simulated states of the Markov chains is a critical issue in model estimation. The Markov chain should theoretically converge to a stationary distribution from which the samples from the posterior distribution are drawn. There are several factors that can impact the convergence rate (Kim & Bolt, 2007). First, high autocorrelations in the Markov chains result in a slow convergence rate where the samples cannot be considered as independent draws from the posterior distribution. Therefore, a large number of iterations are needed before a valid sample from the posterior distribution can be obtained. Second, the choice of sampling algorithms can affect the convergence rate. For instance, as stated earlier, the M-H algorithm is more efficient than the Metropolis algorithm due to asymmetric proposal distributions (Gelman et al., 2003). Lastly, non-convergence issue could also relate to model identification problems. In such cases, the model identification constraints are not sufficient to estimate the parameters of interest.

Lack of convergence can be detected from two aspects, visual inspection of plots of the Markov chains as well as diagnostic indices. The first aspect includes plots for history, running mean, density, quantiles of the chains, among others. Figure 4 shows two examples of history plots demonstrating convergent and non-convergent evidence. As one can see, the convergent chain on the upper panel is rather stable across iterations, whereas the non-convergent chain on the lower panel demonstrates much more variability at different phases of the chain. Similarly, stable running mean and quantiles indicate the convergence of a chain. In addition, the density plot of the samples should be smooth and unimodal when a chain is converged.

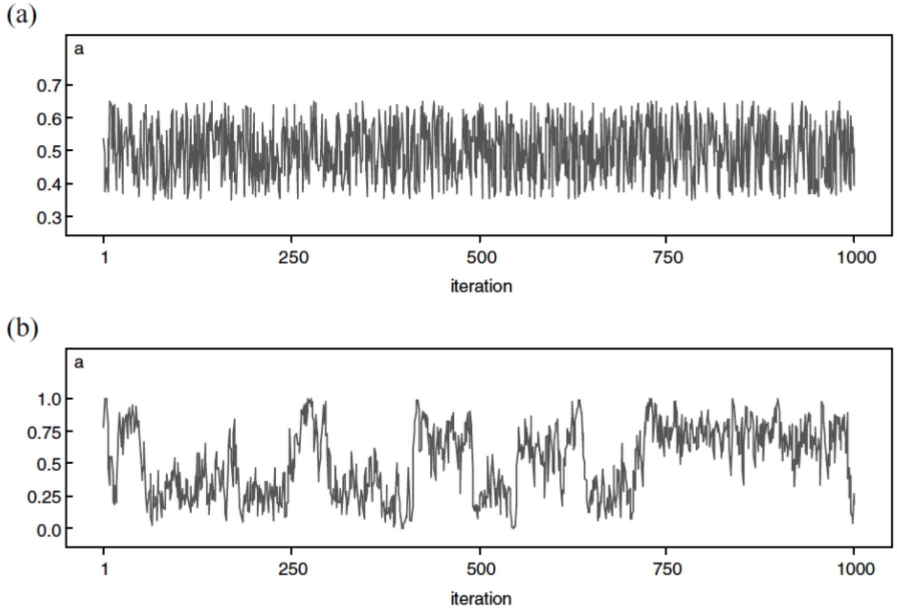


Figure 4. Examples of sampling history plots displaying evidence of (a) convergence and (b) non-convergence (Adapted from Kim and Bolt, 2007, p. 43).

Non-convergence in the Markov chains is sometimes apparent through visual inspections. However, there are other scenarios where it is not as easy to detect non-convergence through examining the plots, for example when the number of parameters is large. In such cases, diagnostic indices can be calculated to provide a numerical gauge of model convergence. Two commonly utilized diagnostic indices among them are Geweke's (1992) z-score and Gelman and Rubin's (1992) potential scale reduction factor, also called $\hat{R}$. Based on Geweke's (1992) approach, a z-score is computed as the standardized difference between the first 10% and last 50% of the chain for each model parameter. The significance of this z-score is tested against zero as it is assumed to follow the standard normal distribution. Falling in the non-significance range (i.e., $-1.96 \leq z \leq 1.96$) is considered as evidence of convergence. Another approach to evaluating convergence numerically when multiple chains are simulated is Gelman and Rubin's (1992) $\hat{R}$. The idea of this statistic is to compare the between-chain variance and within-chain variance for each parameter. Convergence is achieved for a parameter when $\hat{R}$ approximates 1.0 (Gelman & Rubin, 1992). In the proposed study, a combination of diagnostic plots and indices are applied to examine the convergence of the model parameters.

Chapter 3: Methods

The first chapter introduces the motivation of the joint modeling of RT and RA accounting for the interaction among speed, accuracy, and item difficulty in the context of timed tests and the potential contributions of the present study. In the second chapter, the theoretical foundations of the joint modeling approach are summarized for modeling item responses and RTs respectively, as well as several joint modeling frameworks, allowing RT and RA to be conditionally independent or dependent. The estimation methods utilized to obtain model parameter estimates are also reviewed in the second chapter, with a focus on the Bayesian inference. Built on the first two chapters, this chapter first illustrates the proposed models for violations of conditional independence due to interactions among speed, accuracy, and item difficulty, and then demonstrates the implementation of estimation methods in the Bayesian framework. The proposed models are evaluated via simulation studies and empirical data analyses, as presented in the last two sections of this chapter.

3.1 Joint Modeling for the Speed-Accuracy-Difficulty Interaction

In this section, models for the conditional dependence between responses and RTs are proposed to account for the speed-accuracy-difficulty interaction based on van der Linden’s (2007) hierarchical modeling framework. Such a modeling framework is chosen for three reasons. First, it has been shown that the log-normal distribution fits RT distributions better than other alternative distributions, such as Weibull and gamma distributions (e.g., Schnipke & Scrams, 1999). Second, even though models such as the Box-Cox normal model (Klein Entink, van der Linden, &

Fox, 2009) and the linear transformation model (Ranger & Kuhn, 2013; Wang, Chang, et al., 2013) can accommodate more generalized RT distributions and less stringent assumptions, the log-normal model is more parsimonious with acceptable model fit and easily interpretable model parameters. Third, rather than imposing a functional relationship between speed and accuracy, the hierarchical modeling framework offers an appealing approach to describe the relationship between speed and accuracy by allowing them to covary.

As discussed in section 2.3, violations of the conditional independence assumption have mainly been investigated from two aspects. On the one hand, researchers have wondered whether fast and slow responses are associated with distinguishable latent traits or different test-taking processes. On the other hand, a less explored approach is to assume that correct and incorrect responses give rise to different RT distributions. A number of studies suggest that item difficulty seems to interact with the relationship between speed and accuracy systematically (Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer et al., 2014, 2015; Partchev & De Boeck, 2012) and advocate the investigation of specific item characteristics (Bolsinova & Tijmstra, 2017). The present study follows the second approach of modeling different RT distributions for correct and incorrect responses because the variable speed assumption may be more viable than the variable ability assumption.

In the present study, different approaches are developed to explore the relationship between item difficulty and time intensity parameters depending on correct and incorrect responses based on van der Linden and Glas' (2010) model as demonstrated in (2.27). In such an approach, the shift in time intensity parameter can be considered a direct measure of the conditional dependence between responses and RTs (van der Linden & Glas, 2010).

The present study only focuses on the shift in time intensity parameter for the following three reasons. First, explorations of data from various testing programs indicate a linear relationship between item difficulty and the difference between the average logRT for correct and incorrect responses. However, it remains unclear how the variance of the logRT for correct and incorrect responses changes in relation to item difficulty. Second, though this interaction may also affect the variance of the RT distributions, empirical research has shown that in real testing scenarios more items demonstrate shift in the location of the RT distribution rather than the variance when conditional dependence between responses and RTs is present (Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016). Third, the effects of shift in both time intensity and time discrimination parameters could be confounded with each other, thus the impact on parameter estimates may be less evident.

Mathematically, a general form of the probability density function for the proposed models can be expressed as follows:

$$f\left(t_{ij}\middle|u_{ij},\tau_j,\alpha_i,\beta_i,\lambda_i\right)=\frac{\alpha_i}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i\left[\ln t_{ij}-\left(\beta_i+u_{ij}\lambda_i-\tau_j\right)\right]\right]^2\right\}, \qquad (3.1)$$

where u_{ij} represents a binary indicator that classifies person j's RT to item i into one of the two RT distributions, and λ_i indicates the shift in time intensity parameter triggered by $u_{ij}=1$. Notice that the sign proceeding $u_{ij}\lambda_i$ is changed from negative as in van der Linden and Glas' (2010) to positive for easier interpretations. In the proposed model, a positive λ_i indicates that item responses with $u_{ij}=1$ are more time-intensive than those with $u_{ij} = 0$, whereas in van der Linden and Glas' (2010) model representation, a positive λ_i represents a decrease in time intensity for $u_{ij} = 1$. The binary indicator can be based on observed item responses or latent response processes.

Given the strong linear correlation between item difficulty and the shift magnitude λ_i demonstrated in Figure 3, λ_i can be modeled with a linear link to item difficulty:

$$\lambda_i = \omega_0 + \omega_1 b_i + \phi_i, \tag{3.2}$$

where ω_0 and ω_1 are the intercept and slope parameters that determine the linear association between the shift in time intensity parameter and item difficulty, and ϕ_i is an item-specific random effect that follows a normal distribution with mean of zero and variance of σ_ϕ^2. Substituting (3.2) in (3.1), the probability density function of the joint model of responses and RTs conditioning on item difficulty and observed item responses can be expressed as:

$$\begin{aligned} &f\left(t_{ij} \middle| y_{ij}, \tau_j, b_i, \alpha_i, \beta_i, \phi_i, \omega_0, \omega_1\right) \\ &\quad = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left[\ln t_{ij} - \left(\beta_i + y_{ij}(\omega_0 + \omega_1 b_i + \phi_i) - \tau_j\right)\right]\right]^2\right\}. \end{aligned} \tag{3.3}$$

van der Linden and Glas (2010) treated λ_i as a fixed effect in RT modeling, whereas in the present study, λ_i is decomposed into fixed effects associated with item difficulty, and a random effect for each specific item. Compared to van der Linden and Glas' (2010) model, (3.3) allows one to examine the magnitude of the relationship between item difficulty and the shift in parameters for RT distribution, indicated by ω_1. ω_0 represents the shift in time intensity parameter when item difficulty is zero. When $y_{ij} = 0$, the time intensity parameter of item i is β_i, which can be viewed as the baseline time intensity for a specific item. When $y_{ij} = 1$, the item time intensity parameter of item *i* shifts to $\beta_i + (\omega_0 + \omega_1 b_i + \phi_i)$. If ω_1 is positive, item difficulty is positively associated with λ_i, and vice versa. Based on the results from previous studies that correct responses are often less time-intensive than incorrect responses for easier items, whereas correct responses are more time-intensive than incorrect responses for difficult items (Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer et al., 2014, 2015; Partchev & De Boeck, 2012), it is therefore expected that ω_1 would be greater than zero. Solving $\lambda_i = 0$ yields $-\omega_0/\omega_1$, which represents the item difficulty or ability level at which a correct response and an incorrect response are equally time-intensive.

The full linear model as demonstrated in (3.2) allows the shift λ_i to be fully explained by a linear transformation of the item difficulty and a random effect. However, one might adopt different assumptions and choose among the constrained models. One option is to assume perfect correlation between item difficulty and the shift and drop the random effect as follows:

$$\lambda_i = \omega_0 + \omega_1 b_i. \tag{3.4}$$

Moreover, one may ignore the relationship between item difficulty and the shift and assume $\omega_1 = 0$:

$$\lambda_i = \omega_0 + \phi_i, \tag{3.5}$$

then the model is similar to van der Linden and Glas' (2010) model, except that λ_i is considered random and the sign preceding λ_i is reversed.

In addition to the observed item responses, the pattern of RT distribution shift based on observed responses may also reflect a switch in latent response process, such as problem-solving strategy. A different problem-solving strategy may be provoked when an examinee's ability is greater than the item difficulty. The indicator u_{ij} that triggers the shift in RT location can thus be determined by $I(\theta_j - b_i > 0)$, where $I(\cdot)$ is an indicator function that equals 1 if the condition in parenthesis holds and 0 otherwise. Combining the indicator function and (3.2), the joint model of responses and RTs conditioning on the item-person distance and item difficulty can be expressed as follows:

$$\begin{aligned} & f\left(t_{ij} \middle| \theta_j, \tau_j, b_i, \alpha_i, \beta_i, \phi_i, \omega_0, \omega_1\right) \\ & = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left[\ln t_{ij} - \left(\beta_i + \ I(\theta_j - b_i > 0)(\omega_0 + \omega_1 b_i + \phi_i) - \tau_j\right)\right]\right]^2\right\}. \end{aligned} \tag{3.6}$$

This model shares the same idea as Bolsinova and Tijmstra (2017), which incorporates the cross-loadings between ability and RTs. It is expected that more information from RTs could be "borrowed" to improve the estimation accuracy of ability and item difficulty parameters, even though the item-person distance is dichotomized in the proposed model.

Another reason why the indicator function might be preferred over the observed item responses in modeling different RT distributions is that guessing and slipping effects are inevitable in most real testing scenarios with multiple-choice questions. Assuming that the RT for an item reflects the response processes associated with an item response, low-ability examinees who make a lucky guess may not have gone through the processes required for a correct response, whereas highly competent examinees may mistakenly choose the wrong answers with all required response processes. Therefore, the responses in such scenarios are inconsistent with the information provided in RTs. Compared to item responses, it is expected that classifying examinees' RTs based on the indicator function would yield more refined RT distributions that reflect the actual response processes taken. A possible alternative to address the slipping and guessing effects is to replace the indicator function $I(\theta_j - b_i > 0)$ by the probability of the IRT model, thus shift in RT distributions is gradual depending on the IRT probability rather than being triggered by a binary indicator. This perspective is not included in the present study since it does not perform better than the proposed models in pilot simulations.

The proposed models and the alternative models are summarized in Table 1 with respect to the model structure of the shift on RT distribution. Specifically, the two full proposed models (see (3.3) and (3.6)) and their constrained versions are compared. Of the six proposed models, four take into account the effect of item difficulty on the location shift between the two RT distributions. Another alternative model is the original hierarchical model (van der Linden, 2007) assuming conditional independence between responses and RTs. All models in Table 1 utilize the Rasch model for modeling item responses and a log-normal model for RT modeling, but they differ in how the conditional dependence between responses and RTs is modeled.

The Rasch model is chosen for modeling item responses for several reasons. On the one hand, Rasch model is the simplest IRT model, which is widely used in licensure and certification tests (e.g., O'neill, Marks, & Reynolds, 2005; Swanson, Case, Ripkey, Clauser, & Holtman, 2001), as well as some large-scale educational assessments (e.g., Adams & Wu, 2002). On the other hand, if the 2PL or the 3PL IRT model is used, the estimation error in the discrimination and guessing parameter estimates may be absorbed in the item difficulty or ability estimates, which is likely to introduce more error to RT parameter estimates. In the following sections of this chapter, the estimation procedure of the proposed model within the Bayesian framework is demonstrated, and the research plan to evaluate the proposed model through simulation studies and empirical data analyses is described.

Table 1. *The proposed and the alternative models.*

Model	Abbreviation	Shift Indicator u_{ij}	Shift Magnitude λ_i
Joint model conditioning on item response and difficulty with random effects	JM-RD1	y_{ij}	$\omega_0 + \omega_1 b_i + \phi_i$
Joint model conditioning on item response and difficulty without random effects	JM-RD2	y_{ij}	$\omega_0 + \omega_1 b_i$
Joint model conditioning on item response	JM-R	y_{ij}	$\omega_0 + \phi_i$
Joint model conditioning on item-person distance and difficulty with random effects	JM-DD1	$I(\theta_j - b_i > 0)$	$\omega_0 + \omega_1 b_i + \phi_i$
Joint model conditioning on item-person distance and difficulty without random effects	JM-DD2	$I(\theta_j - b_i > 0)$	$\omega_0 + \omega_1 b_i$
Joint model conditioning on item-person distance	JM-D	$I(\theta_j - b_i > 0)$	$\omega_0 + \phi_i$
Hierarchical model (van der Linden, 2007)	HM	NA	NA

3.2 Model Parameter Estimation

In the present study, Bayesian estimation of model parameters is carried out in R2jags package (Su & Yajima, 2015) in R to interface with JAGS (Version 4.2.0; Plummer, 2015). Parameters of interest include item difficulty, time intensity and time discrimination, parameters related to the magnitude of the shift, ability and speed, as well as the correlation between ability and speed. Following van der Linden's (2007) hierarchical modeling framework for RT and RA, the posterior distribution of the parameters for the proposed models can be generally expressed as follows:

$$\begin{aligned}&f\left(\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{b},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\phi},\omega_0,\omega_1,\sigma_\phi^2,\boldsymbol{\mu}_P,\boldsymbol{\mu}_I,\boldsymbol{\Sigma}_P,\boldsymbol{\Sigma}_I\middle|\boldsymbol{y},\boldsymbol{t}\right)\\ &\propto f(\boldsymbol{y},\boldsymbol{t}|\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{b},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\phi},\omega_0,\omega_1)f(\boldsymbol{\theta},\boldsymbol{\tau}|\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)f(\boldsymbol{b},\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{\mu}_I,\boldsymbol{\Sigma}_I)f(\boldsymbol{\phi}|\sigma_\phi^2)\\ &\times f(\omega_0)f(\omega_1)f\left(\sigma_\phi^2\right)f(\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)f(\boldsymbol{\mu}_I,\boldsymbol{\Sigma}_I),\end{aligned} \tag{3.7}$$

where parameters in bold represent the vectors of item and person parameters or matrices of observed data. In this expression, the likelihood of observed item responses and RTs can be expanded into the following for JM-RD1:

$$\begin{aligned}&f(\boldsymbol{y},\boldsymbol{t}|\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{b},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\phi},\omega_0,\omega_1)\\ &=\prod_{j=1}^{J}\prod_{i=1}^{I}f\left(y_{ij}\middle|\theta_j,b_i\right)f\left(t_{ij}\middle|y_{ij},\tau_j,b_i,\alpha_i,\beta_i,\phi_i,\omega_0,\omega_1\right),\end{aligned} \tag{3.8}$$

where *I* and *J* represent the test length and sample size respectively. Employing the Rasch model for item responses, the probability density function of an observed item response is expressed as

$$f\left(y_{ij}\middle|\theta_j,b_i\right)=\left\{\frac{1}{1+\exp\left[-\left(\theta_j-b_i\right)\right]}\right\}^{y_{ij}}\left\{\frac{1}{1+\exp\left[\left(\theta_j-b_i\right)\right]}\right\}^{1-y_{ij}}, \tag{3.9}$$

which is the same for all proposed and alternative models presented in Table 1. For JM-RD1, the probability density function of RT is presented in (3.3). For JM-RD2 and JM-R, the probability density functions of RT are constrained versions of the full model expressed as

$$
\begin{aligned}
&f\left(t_{ij}\middle|y_{ij}, \tau_j, b_i, \alpha_i, \beta_i, \omega_0, \omega_1\right) \\
&\qquad = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left[\ln t_{ij} - \left(\beta_i + y_{ij}(\omega_0 + \omega_1 b_i) - \tau_j\right)\right]\right]^2\right\},
\end{aligned}
\tag{3.10}
$$

and

$$
\begin{aligned}
&f\left(t_{ij}\middle|y_{ij}, \tau_j, \alpha_i, \beta_i, \phi_i, \omega_0\right) \\
&\qquad = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left[\ln t_{ij} - \left(\beta_i + y_{ij}(\omega_0 + \phi_i) - \tau_j\right)\right]\right]^2\right\},
\end{aligned}
\tag{3.11}
$$

respectively. Similarly, for the three models that classify examinees based on item-person distance, the probability density functions of RT are replaced by $f\left(t_{ij}\middle|\theta_j, \tau_j, b_i, \alpha_i, \beta_i, \phi_i, \omega_0, \omega_1\right)$ as presented in (3.6) and its constrained versions.

Further, $f(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ and $f(\boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ represent the multivariate normal distributions of person parameters and item parameters given the mean and covariance matrix for person and item respectively. The random effects $\boldsymbol{\phi}$ are assumed to be drawn from a normal distribution with mean of 0 and variance of σ_ϕ^2. In addition, $f(\omega_0)$ and $f(\omega_1)$ are the distributions of the intercept and slope for the effect of item difficulty on RT location shift. The last three terms, $f\left(\sigma_\phi^2\right)$, $f(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$, and $f(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ denote the distributions of the variance of random effect, mean and covariance of person and item parameters. Finally, the posterior distribution of the parameter space is obtained by multiplying the likelihood and all the prior distributions on the right-hand side of (3.7).

The posterior distribution is derived by drawing samples from the prior distributions and updating the likelihood of the observed responses and RTs sequentially. As such, setting appropriate prior distributions is important for facilitating model convergence. In the present study, the prior distributions are chosen based on Meng et al.'s (2015) study. Specifically, the following prior distributions are adopted for person and item parameters:

$$\begin{pmatrix}\theta_j\\ \tau_j\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\theta\\ \mu_\tau\end{pmatrix}, \begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau}\\ \sigma_{\theta\tau} & \sigma_\tau^2\end{pmatrix}\right), \qquad \begin{pmatrix}b_i\\ \beta_i\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_b\\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_b^2 & \sigma_{b\beta}\\ \sigma_{b\beta} & \sigma_\beta^2\end{pmatrix}\right). \tag{3.12}$$

For model identification purposes, μ_θ and μ_τ are set to be zero. Notice that only the item difficulty parameters and the time intensity parameters are drawn from a bivariate normal distribution, which are independent of the time discrimination parameters. Since the inverse of squared time discrimination is the variance of the RT distribution, the following prior distribution is chosen for the time discrimination parameters:

$$\frac{1}{\alpha_i^2} \sim InvGamma(1,1). \tag{3.13}$$

This is different from the prior for item parameters used in van der Linden's (2007) study, where all item parameters are assumed to follow a multivariate normal distribution. There are two reasons why the bivariate normal distribution and the inverse-gamma distribution are utilized in the present study. First, van der Linden (2007) demonstrated that among all the correlations between item parameters (see (2.24)), only the correlation between item difficulty and time intensity is significantly different from zero. Second, pilot simulation runs show that both prior settings yield accurate parameter estimates, yet using the bivariate normal and the inverse-gamma priors is much more computationally efficient than using the multivariate normal prior. Using the multivariate normal distribution in van der Linden (2007), 10 iterations take 10 seconds for item response and RT data from 500 examinees and 20 items in JAGS, whereas the bivariate normal and inverse-gamma priors only take 1 second to finish 10 iterations for the same dataset. This finding is consistent with the study by Molenaar, Tuerlinckx, and van der Maas (2015b) that ignoring the covariances among item parameters does not negatively affect the parameter recovery. Therefore, to expedite the simulation studies with similar parameter recovery accuracy, the bivariate normal and the inverse-gamma priors are chosen for the item parameters.

The additional parameters that link the effect of item difficulty to the location shift are drawn from the following prior distributions:

$$\omega_0 \sim N(0,1), \omega_1 \sim N(0,1), \phi_i \sim N\left(0, \sigma_\phi^2\right), \tag{3.14}$$

where the constraint $\sum_{i=1}^{I} \phi_i = 0$ is applied for model identification as well. Specifically, ω_0 and ω_1 are assumed to follow a standard normal distribution. The random effects, ϕ_is, are normally distributed with a mean of 0 and unknown variance. Further, hyper priors are an extra set of priors from which the hyper-parameters are drawn, which are the parameters of the prior distributions specified above. Specifically, the hyper priors for $\boldsymbol{\mu}_I$, $\boldsymbol{\Sigma}_I$, $\boldsymbol{\Sigma}_P$, and σ_ϕ^2 are specified as follows:

$$\begin{gathered} \mu_b \sim N(0,2), \mu_\beta \sim N(4,2), \\ \boldsymbol{\Sigma}_I \sim InvWishart(\boldsymbol{I}_2, 2), \boldsymbol{\Sigma}_P \sim InvWishart(\boldsymbol{I}_2, 2), \\ \sigma_\phi^2 \sim InvGamma(1,1), \end{gathered} \tag{3.15}$$

where $\boldsymbol{I}_2$ is the 2-dimensional identity matrix. The shapes of the priors and hyper priors are chosen based on the literature (e.g., Klein Entink, Fox, & van der Linden, 2009; Klein Entink, van der Linden, & Fox, 2009; Meng et al., 2015; van der Linden,

2007) and preliminary analysis of empirical data. For the item difficulty parameter, the mean of the hyper prior is fixed at 0 such that the range of latent ability and item difficulty are approximately the same. For the time intensity parameter, the mean of the hyper prior is set at 4 to resemble the mean logRT in the first empirical dataset used in the present study (i.e., 3.98). Further, the inverse-Wishart distribution is often chosen as a hyper prior for the multivariate normal distribution due to its conjugacy to the multivariate normal distribution. Similarly, the inverse-gamma family is conditionally conjugate for the variance of random effects in that if the variance follows an inverse-gamma prior distribution, the conditional posterior distribution of the variance is also inverse-gamma (Gelman, 2006).

Bayesian estimation requires the starting values for each parameter be provided as the first state of each Markov chain (Gelman et al., 2003). In the present study, JAGS randomly generates the starting values for all parameters. After generating starting values, two chains of 30,000 iterations are run for each dataset and the first 20,000 are discarded as burn-in iterations. The numbers of total and burn-in iterations are determined by Gelman and Rubin's (1992) $\hat{R}$ and visual examination of the history, density, and quantile plots. In the pilot simulation runs with 30,000 iterations, the proposed and alternative models all have $\hat{R} < 1.1$ for all parameters, and the plots also demonstrate evidence that the two chains are stable and well-mixed. Then a thinning of 2 is applied to reduce the autocorrelation in the Markov chains, yielding a total of 10,000 for the final sample. Parameter estimates are summarized based on the 10,000 final sample.

3.3 Simulation Design

The proposed modeling framework and the methods for model parameter estimation are introduced in the previous two sections. To examine the performance of the proposed models, three simulation studies are carried out. Simulation studies 1 and 2 have the same simulation design but different data generating models and fitting models. Simulation study 3 compares the performance of the seven models summarized in Table 1 under two simulation conditions with small sample size and weakest correlations among parameters. The fixed and manipulated factors in the simulation studies are illustrated first, then the criteria for evaluating parameter recovery and overall model fit are demonstrated.

3.3.1 Manipulated Factors

The manipulated factors in simulation studies 1 and 2 include sample size (500, 1,000), test length (20, 40), the correlation between speed and ability (.2, .5, .8), and the correlation between item difficulty and location shift in RT distribution (.3, .7). Difference between simulation studies 1 and 2 lies in the data generating models and fitting models. Simulation study 1 generates data based on the JM-RD1 as demonstrated in (3.3) and fit the data to the three models conditioning on responses and the HM, whereas simulation study 2 utilizes the JM-DD1 in (3.6) as the data generating model and fits the data to the three models conditioning on item-person distance and the HM. Specific levels of manipulated factors are summarized in Table 2. Fully crossing the four manipulated factors results in a total of 24 simulation conditions in simulation studies 1 and 2, as displayed in Table 3. In simulation study 3, data generated from JM-RD1 and JM-DD1 are fit to all seven models summarized in Table 1 under conditions 1 and 7, where sample size is 500, test length is 20 and 40, and the last two factors are at the lowest correlation levels. These two conditions are chosen to evaluate the recovery of model parameters and the performance of model fit indices under less favorable circumstances.

Table 2. *Summary of manipulated factors.*

	Manipulated Factors			
Levels	Sample Size	Test Length	$\rho_{\theta\tau}$	$\rho_{b\lambda}$
1	500	20	.2	.3
2	1,000	40	.5	.7
3			.8	

Table 3. *Summary of simulation conditions.*

	Manipulated Factors			
Condition No.	Sample Size	Test Length	$\rho_{\theta\tau}$	$\rho_{b\lambda}$
1	500	20	.2	.3
2	500	20	.2	.7
3	500	20	.5	.3
4	500	20	.5	.7
5	500	20	.8	.3
6	500	20	.8	.7
7	500	40	.2	.3
8	500	40	.2	.7
9	500	40	.5	.3
10	500	40	.5	.7
11	500	40	.8	.3
12	500	40	.8	.7
13	1,000	20	.2	.3
14	1,000	20	.2	.7
15	1,000	20	.5	.3
16	1,000	20	.5	.7
17	1,000	20	.8	.3
18	1,000	20	.8	.7
19	1,000	40	.2	.3
20	1,000	40	.2	.7
21	1,000	40	.5	.3
22	1,000	40	.5	.7
23	1,000	40	.8	.3
24	1,000	40	.8	.7

The four manipulated factors and their levels were chosen based on previous literature in this line of research and pilot simulation runs. The first three factors have been frequently manipulated in recent studies. In terms of sample size, for example, Fox et al. (2014) and Marianti (2015) found good model parameter recovery when sample size is 300 and 500. Also with two levels, Lee (2007) and Molenaar, Oberski, et al. (2016) used 500 and 1,000 to represent smaller and larger sample sizes. Ranger and Kuhn (2014b) manipulated three levels, 250, 500, and 1,000, to evaluate the Type I error rate and power of the proposed model fit index. Suh (2010) chose four sample sizes, 100, 500, 1,000, and 2,000 in comparing Thissen's (1983), Wang and Hanson's (2005) and van der Linden's (2007) models. Most studies that manipulated sample size concluded that increasing sample size improves ability estimates. When sample size is treated as a fixed factor, researchers often simulate 500 (Fox & Marianti, 2017; Molenaar & Bolsinova, 2017; Molenaar, Bolsinova, & Vermunt, 2016) or 1,000 examinees (Fox & Marianti, 2016; Klein Entink, 2009; Klein Entink, Fox, & van der Linden, 2009; Patton, 2015; Wang & Xu, 2015) in simulation studies. Therefore, 500 and 1,000 are selected as they are common levels when sample size is treated as both manipulated and fixed factors.

Similarly, model parameter estimates are more accurate with longer test length since more information can be borrowed from RTs (e.g., Kang, 2016; Marianti, 2015; Suh, 2010; Wang, Fan, et al., 2013). Specifically, Suh (2010) set the number of items as 30 and 60, whereas Bolsinova, De Boeck, and Tijmstra (2017) considered test length at two levels, 25 and 49. Kang (2016) chose test length to be 20 and 30 to examine the recovery of item parameters from the proposed likelihood-based methods. A number of other studies also manipulated 20 and 40 as two levels of sample size (Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; Fox & Marianti, 2016; Ingrisone, 2008; Lee, 2007; Molenaar et al., 2014; Wang, Fan, et al., 2013). As most studies choose to manipulate sample size as 20 and 40, these two levels are adopted in the present study as well.

With respect to the correlation between ability and speed, researchers have been interested in exploring the impact of different levels of correlation on parameter estimation, ranging from -1 to 1. Kang (2016) chose person parameters to be 0, .3, and .6 because van der Linden (2009) revealed that empirical estimates of the correlation were found to fall between -.65 and .30. Klein Entink (2009) compared person and item estimates for correlations of 0, .25, .75, 1 and concluded that higher correlation between ability and speed yielded more accurate parameter estimates. Patton (2015) considered a wider range of correlations from 0, .3, .6, to .9. Suh (2010) used correlations with the same magnitude as Patton (2015) but with both positive and negative relations. A few studies only manipulated two levels of correlations, 0 and .5 (Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; Bolsinova & Tijmstra, 2017). Further, when correlation is treated as fixed, levels such as -.3 (Ranger & Kuhn, 2012), .37 (Molenaar & Bolsinova, 2017), .4 (Molenaar, Bolsinova, & Vermunt, 2016), .5 (Klein Entink, Fox, & van der Linden, 2009; Klein Entink, van der Linden, & Fox, 2009; Molenaar et al., 2014), and .75 (Fox & Marianti, 2017) have been used. As such, .2, .5 and .8 are chosen to represent weak, moderate, and strong correlation between ability and speed in the present study. Only positive correlations are considered since the sign of correlation does not affect the amount of information shared between the two latent traits.

The last manipulated factor in the present study is the correlation between item difficulty and location shift on the RT distribution, which has not been explicitly explored in the literature. Based on preliminary results of empirical data from several large-scale assessment programs, a strong positive linear relationship between item difficulty and location shift is found with correlation around .6 to .7. Thus, .7 is selected to mimic patterns found in real testing scenarios. .3 represents a situation where the correlation between item difficulty and location shift is rather weak.

3.3.2 Fixed Factors

In addition to the manipulated factors, certain factors are fixed in the current simulation design, including the distributions of the latent ability and speed, the distributions of item difficulty and time intensity, and a number of fixed parameters. The data generation models for item responses and RT are also fixed to the Rasch model and the log-normal model respectively. Table 4 details the fixed factors and their corresponding levels in the present study.

Table 4. *Summary of fixed factors.*

Factor	Fixed Value
Distribution of ability	*N*(0, 1)
Distribution of speed	*N*(0, .25)
Distribution of item difficulty	*N*(0, 1)
Distribution of time intensity	*N*(4, .25)
Correlation between item difficulty and time intensity	.30
Time discrimination	2
SD of the location shift	.20
ω_0	-.30
IRT model	Rasch
RT model	log-normal model

Both latent ability and item difficulty are generated from the standard normal distribution, following the convention of numerous studies in IRT literature. The latent speed is also drawn from a normal distribution with mean of 0 and variance of .25, which is chosen based on the estimated variance of speed from empirical data analyses (Klein Entink, Fox, & van der Linden, 2009; Molenaar & Bolsinova, 2017). Varying the correlation between ability and speed at .20, .50, .80 results in the covariance of .10, .25, and .40 between the two latent traits respectively. The distribution of time intensity mimics the actual RT distribution in real testing scenarios where items require 1-2 minutes to finish, and the individual RT differences could be large (Lee, 2007).

Moreover, the correlation between item difficulty and time intensity is fixed at .30, indicating that more difficult items are also more time consuming. Researchers have chosen to fix the time discrimination parameters (e.g., Molenaar, Bolsinova, Rozsa, & De Boeck, 2016; Molenaar, Bolsinova, & Vermunt, 2016; Molenaar et al., 2015b), or sample the time discrimination or error variance from a normal (e.g., Fox et al., 2014) log-normal distribution (e.g., Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016). In the present study, the time discrimination parameter is fixed at 2, so that the logRT distributions are generated with a variance of .25. This is carried out to mimic the variance of the logRT distributions in the first empirical dataset, which has a mean of .272 and a variance of .005.

Additionally, the SD of the location shift and the intercept for regressing the item difficulty on the time intensity are set as .20 and -.30 respectively. These values are chosen based on preliminary results from the empirical data. With the SDs of location shift and item difficulty fixed at .20 and 1 respectively, manipulating the correlation between the location shift and the item difficulty at .30 and .70 in fact yields a slope ω_1 of .06 and .14. This indicates that for one unit increase in item difficulty, $u_{ij} = 1$ is associated with .06 and .14 units increase in location shift on the logRT scale compared to $u_{ij} = 0$. The intercept -.30, on the other hand, means that $u_{ij} = 1$ is .30 units faster than $u_{ij} = 0$ on the logRT scale when the item difficulty is zero. Further, variance of the random effects, denoted as σ_ϕ^2, varies depending on the correlation between item difficulty and location shift, the SD of location shift, and the proposed models. For JM-RD1 and JM-DD1, σ_ϕ^2 is manipulated at .04 and .02 when the correlation between item difficulty and location shift is .30 and .70 respectively and the SD of location shift is .20. For JM-R and JM-D, σ_ϕ^2 is .04 under all conditions as the relationship between item difficulty and location shift is not taken into account. Finally, the Rasch model and the log-normal RT model are used for modeling item responses and RTs.

The number of replications in this line of research varies from 10, 50, 100, 1,000, to 2,000 (e.g., Bolsinova & Maris, 2016; Fox & Marianti, 2016; Ingrisone, 2008; Molenaar, Oberski, et al., 2016; Patton, 2015). For Monte Carlo studies in IRT-based research, 25 has been justified as the minimum number of replications (Harwell, Stone, Hsu, & Kirisci, 1996). Harwell et al. (1996) carried out analysis of variance (ANOVA) for the RMSE of item parameters in the 2PL model and did not find significant change in RMSE after 25 replications. For more complex model structures, Li (2014) also adopted 25 replications for mixture Rasch model with covariates since there is little fluctuation in the bias and SE of both item and person parameters. Xie (2014) conducted a post hoc checking for item parameters in cross-classified IRT models and concluded that the SEs flattened out after 30 replications.

Further, 100 preliminary simulation runs are conducted for fitting JM-RD1 to data generated from the same model for 500 examinees and 20 items. The correlation between latent ability and speed is set as .5, and the correlation between item difficulty and location shift is specified as .7. All fixed parameters are simulated as specified in Table 4. As shown in Figures 5 and 6, the mean bias, SE, and RMSE for item parameters and the mean SE and RMSE for person parameters stabilize after 25 replications. The mean biases of person parameters are constrained to be zero for scale identification purpose, thus are not included in Figure 6. Similar patterns in the three error indices are also found when examining individual item and person parameters and other parameters, such as the mean vector and covariance matrix for item parameters, covariance matrix for person parameters, variance of random effect, and so on. Therefore, 30 replications are considered sufficient for evaluating item and person parameter recovery in the proposed and alternative models in the present study.

A total of $2 \times 2 \times 3 \times 2 = 24$ simulation conditions are included in simulation studies 1 and 2. Generating 30 datasets under each condition in each simulation study results in 1440 datasets in total. In simulation study 1 where JM-RD1 is the data generating model, the three joint models conditioning on observed item response (i.e., JM-RD1, JM-RD2, and JM-R) and the HM are fit to each dataset. In simulation study 2 where JM-DD1 is used to generate data, the three joint models conditioning on item-person distance (i.e., JM-DD1, JM-DD2, and JM-D) and the

HM are fit to each dataset. In simulation study 3, JM-DD1, JM-DD2, and JM-D are fit to data generated from JM-RD1, whereas JM-RD1, JM-RD2, and JM-R are fit to data generated from JM-DD1 under two simulation conditions. As such, there are a total of $4 \times 2 \times 24 + 3 \times 2 \times 2 = 204$ simulation cells with $204 \times 30 = 6120$ replications. R version 3.3.3 (R Core Team, 2017) is used to generate data, interface with JAGS using R2jags package, and evaluate model performance.

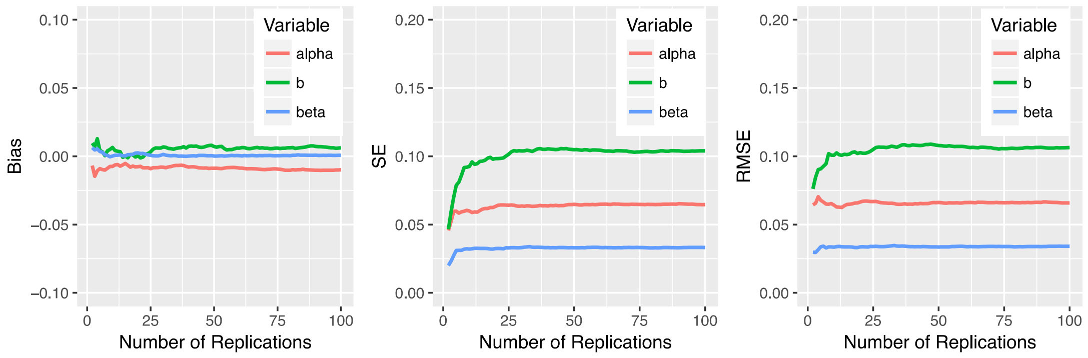


Figure 5. The mean bias, SE, and RMSE of the item parameters by the number of replications.

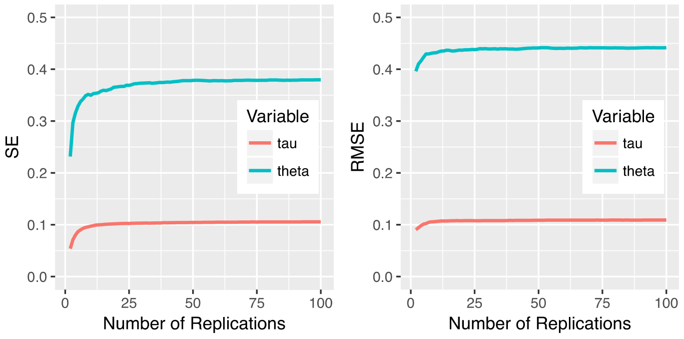


Figure 6. The mean SE and RMSE of the person parameters by the number of replications.

3.3.3 Evaluation Criteria

The proposed and the alternative models are compared under various conditions in the three simulation studies regarding the evaluation criteria summarized in this section. The evaluation criteria mainly aim at examining the recovery of model parameters and the fit of the models. In terms of model parameter recovery, the parameter estimates are compared to their corresponding true values with respect to bias, SE, and RMSE. The three error indices are chosen because they have been utilized in the large body of this line of research to represent different sources of error. They are compared descriptively first and then using ANOVA to test the significance of the manipulated factors. Specifically, the three error indices are defined as follows:

$$Bias(\hat{\boldsymbol{\eta}}) = \frac{\sum_{r=1}^{R}(\hat{\eta}_r - \eta)}{R}, \tag{3.16}$$

$$SE(\hat{\boldsymbol{\eta}}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\eta}_r - \frac{\sum_{r=1}^{R}\hat{\eta}_r}{R}\right)^2}, \tag{3.17}$$

$$RMSE(\hat{\boldsymbol{\eta}}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\eta}_r - \eta)^2}, \tag{3.18}$$

where η is a true parameter of interest, $\hat{\eta}_r$ represents the parameter estimate of η at the *r*th iteration, *R* is the total number of iterations within each simulation cell, and $\hat{\boldsymbol{\eta}}$ is a R-dimensional vector of parameter estimates.

The bias reflects the systematic errors in the estimation as it is calculated as the deviation from the true parameter averaged across iterations. The SE represents the random errors in that it quantifies the variability among the parameter estimates. The RMSE can be regarded as a measure of total error as the following equation holds for a specific parameter:

$$RMSE^2 = Bias^2 + SE^2. \tag{3.19}$$

However, this relation does not hold if bias, SE, and RMSE are averaged across multiple parameters. In the present study, the average bias of latent ability and latent speed across all examinees is zero because μ_θ and μ_τ are constrained to be zero to identify the latent scale.

Additionally, four model fit indices are used for assessing model fit and identifying the best fitting model under different simulation conditions, including Akaike's information criterion (AIC; Akaike, 1987), a modified version of AIC for adjusting small sample sizes (AICc; Sugiura, 1978), Bayesian information criterion (BIC; Schwarz, 1978), and deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). The four model fit indices are specified as follows:

$$AIC = \overline{D(\mathcal{S})} + 2p, \tag{3.20}$$

$$AICc = \overline{D(\mathcal{S})} + \frac{2Np}{N - p - 1}, \tag{3.21}$$

$$BIC = \overline{D(\mathcal{S})} + p \log N, \tag{3.22}$$

$$DIC = \overline{D(\mathcal{S})} + p_D, \tag{3.23}$$

where $\mathcal{S}$ denotes sample space of all model parameters, $\overline{D(\mathcal{S})}$ is the posterior mean of the deviance, p represents the number of parameters, N is the sample size, and p_D is calculated as the posterior mean of the deviance given parameters at each iteration minus the deviance evaluated at the posterior means of the parameters. Among these information-based model fit indices, AIC only penalizes for the number of parameters. AICc is a correction for small sample size based on AIC, which increases the penalty for small sample sizes. When the ratio of sample size to the number of parameters is smaller than 40, AICc is preferred to be used (Burnham & Anderson, 2002). BIC also considers the impact of sample size and prefer more parsimonious model since it penalizes model complexity more heavily than AIC when sample size exceeds 7. Lastly, DIC is a generalization of AIC and BIC for hierarchical modeling and Bayesian model selection when MCMC is used to obtain the posterior distribution.

In summary, the three error indices along with the four model fit indices are selected to provide different perspectives of model fit evaluation for joint modeling of RT and RA. The impact of ignoring the conditional dependence between speed and accuracy on parameter estimation and model fit is explored comprehensively. The effectiveness of these indices is also examined by whether they can identify the true data generating model as the best fitting model.

3.4 Empirical Data Analyses

The application of the proposed models in real testing scenarios is demonstrated with datasets from two large-scale tests that utilize the Rasch model for estimating examinees' ability, a large-scale credentialing exam program (Cizek & Wollack, 2017) and the 2012 Programme for International Student Assessment (PISA; Organisation for Economic Co-operation and Development [OECD], 2014).

The first dataset contains complete item responses and RTs from 1,644 examinees and 200 dichotomous items. Of the 200 items, 170 are operational items and 30 are pretest items with 10 items in three pretest sets. Examinees answer a total of 180 items, consisting of the 170 operational items and 10 items in one of the three pretest sets. Item responses are coded as 0 or 1, and the RT for each item response is also available in seconds. Such a dataset is used for the example presented in Figures 1 to 3 in Chapter 1, which motivates the present study.

Two steps were followed in the process of selecting items from the credentialing exam. Item response data from 1,644 examinees and 170 operational items were first explored to ensure item quality. Item statistics based on classical test theory, such as proportion correct and point biserial correlation, were computed to perform initial checking of item difficulty and item discrimination. Cut values of .95 and .05 were used for proportion correct, whereas .10 was used for point biserial to remove items that are too difficult, too easy, or those that do not discriminate well across examinees. 156 items met these criteria, thus were kept in the present study. Second, 40 items were randomly selected from the 156 items to demonstrate the application of the proposed models. Item responses and RTs from 1,644 examinees and the 40 operational items in this dataset are used to explore the conditional dependence between ability and speed, as well as their relationship with item difficulty.

For 2012 PISA, OECD released scored responses and log files for 30 items from three domains, with about 10 items in each domain. Log files record examinees' actions during the test-taking process and time stamps for each action in a chronological order. The present study focuses on the 10 items from computer-based mathematics domain in 2012 PISA, which contains three polytomous items with three categories and seven dichotomous items. The polytomous items were recoded into dichotomous items by collapsing partial scores and full scores, since there are only around 10% examinees who had full scores. RT information was extracted from the log files by taking the difference in time stamps for the events “START_ITEM” and “END_ITEM” for each examinee on each item. After extracting RTs from log files, they were merged with item responses for the 10 math items, resulting in a total sample size of 7,617. To remove potential impact from multiple countries, Australia was selected as it has the largest sample size among all 30 countries. Therefore, the second dataset contains item responses and RTs from 795 examinees and 10 items.

The six proposed models and van der Linden’s (2007) HM are applied to the two empirical datasets. Parameter estimates are presented to understand the test-taking behaviors in real testing programs. Further, the model fit indices summarized in section 3.3.3 are utilized to examine the performance of the proposed and alternative models.

Chapter 4: Results

Joint modeling of responses and RTs is not a new topic, yet the conditional dependence between responses and RTs and its relationship with item difficulty has not been thoroughly explored. As elaborated in Chapter 3, three simulation studies were conducted to evaluate the performance of the six proposed models compared with van der Linden’s (2007) HM as a baseline model. The six proposed models accounted for different mechanisms initiating the shift in RT distributions and different approaches to explaining the relationship between item difficulty and the shift. In section 4.1, the results of the three simulation studies were reported in terms of the error indices for parameter recovery and overall model fit indices. In section 4.2, the application of the proposed models was demonstrated using two datasets from large-scale assessment programs.

4.1 Results of the Simulation Studies

For all three simulation studies, the recovery of 17 model parameters was examined as listed in Table 5. The estimates of the covariance between item difficulty and time intensity and the covariance between ability and speed were converted to correlations using the estimated variances of the associated parameters as the correlation was manipulated in simulating different study conditions. Convergence was not an issue as all parameter estimates under all conditions and replications had an $\hat{R}$ smaller than 1.1. Diagnostic plots, including history, quantile, and density plots, were also examined to ensure model convergence.

Table 5. *Summary of parameters of interest.*

No.	Symbol	Variable Description	Level	Model
1	b_i	Item difficulty	1	IRT
2	α_i	Time discrimination	1	RT
3	β_i	Time intensity	1	RT
4	θ_j	Ability	1	IRT
5	τ_j	Speed	1	RT
6	ω_0	Intercept of RT distribution shift	1	RT
7	ω_1	Slope of RT distribution shift	1	RT
8	$\rho_{b\lambda}$	Correlation between item difficulty and the shift	1	RT
9	σ_{ϕ}^2	Variance of the random effects	2	RT
10	μ_b	Mean of item difficulty	2	IRT
11	μ_{β}	Mean of time intensity	2	RT
12	σ_b^2	Variance of item difficulty	2	IRT
13	$\rho_{b\beta}$	Correlation between item difficulty and time intensity	2	Both
14	σ_{β}^2	Variance of time intensity	2	RT
15	σ_{θ}^2	Variance of item difficulty	2	IRT
16	$\rho_{\theta\tau}$	Correlation between item difficulty and time intensity	2	Both
17	σ_{τ}^2	Variance of time intensity	2	RT

Bias, SE, and RMSE were calculated based on the estimates from 30 replications for each parameter. Following the sequence listed in Table 5, the detailed bias, SE, and RMSE for each parameter under each simulation condition are reported in Appendices A to C for the three simulation studies respectively. For the first-level item and person parameters (i.e., item difficulty, time discrimination, time intensity, ability and speed), repeated measures ANOVA was performed in SPSS Statistics (version 25.0; IBM Corp, 2017) by specifying each of the three error indices as the dependent variable and the four manipulated variables and the estimation model as factors. Specifically, the four manipulated variables were treated as between-condition factors whereas the estimation model was used as a within-condition factor. The abbreviations for the estimation model, sample size and test length in Table 6 were used in tables and figures in this chapter for clear presentation of the results.

Among the three error indices, bias was treated slightly differently than the other two. Since the sign of bias was indefinite, ANOVA was conducted to test the differences among the absolute values of mean bias for different conditions.

Table 6. *Abbreviations of manipulated factors.*

Abbreviation	Description
Model	Estimation model
J	Sample size
I	Test length

The assumptions of repeated measures ANOVA were checked before conducting the analyses. Although it is assumed that the dependent variable should be normally distributed at each level of the within-condition factor, ANOVA is known to be robust to moderate deviations from normality (e.g., Glass, Peckham, & Sanders, 1972). Inspections of P-P plots and Q-Q plots of the dependent variables showed that there was no severe violation of the normality assumption, thus normality was not considered an issue here. The sphericity assumption assumes the variances of the differences in the outcome measure between all pairs of the within-condition factor are equal. Based on Mauchly's sphericity test (Mauchly, 1940), this assumption was violated for ANOVA conducted with all error indices of all variables in the present study. As such, the Huynh-Feldt correction (Huynh & Feldt, 1976) was applied to adjust the degrees of freedom, which resulted in larger critical values and less inflation of Type I error due to violations of the sphericity assumption.

In addition, the effect size for significant main effects and interactions was computed as a measure of practical importance. In the present study, Cohen's f (Cohen, 1988) was used:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}, \tag{4.1}$$

where η^2 is defined as the proportion of total variance in the dependent variable explained by a certain manipulated factor alone. Cohen (1988) also recommended using .10, .25, and .40 as the cut values for small, medium, and large effect sizes respectively. In the following sections, only those significant effects with at least small effect sizes ($f > .10$) were presented and discussed. For such effects with more than two levels, a post-hoc pairwise comparison was carried out to explore which levels of the effects were significantly different. Dunn-Sidak test (Šidák, 1967) was used to control familywise Type I error rate (Tukey, 1953), which is more powerful than the Bonferroni test (Bonferroni, 1936). For bias, the post-hoc comparisons were performed for comparing the absolute values of mean bias across different levels of a significant effect. All decimals in this chapter were rounded to three places; those with absolute values smaller than .001 were denoted as <.001, followed by (+) or (-) to indicate their signs if significant.

In the following sections, the results from the three simulation studies are summarized and presented. In particular, simulation study 1 compares the three models conditioning on item responses with the HM, where the parameter recovery and model selection results are discussed in detail in section 4.1.1. For simulation study 2, the three models with item-person distance as the shift indicator are evaluated with the HM. Given that most findings from simulation studies 1 and 2 are similar, section 4.1.2 mainly highlights the differences between them. Lastly, section

4.1.3 summarizes the conclusions from simulation study 3, focusing on the consequences of fitting models with misspecified shift indicator.

4.1.1 Simulation Study 1

In simulation study 1, the performance of the three models conditioning on item responses (i.e., the JM-RD1, the JM-RD2, and the JM-R) was compared with the HM. A total of 24 conditions were simulated for comparing the four models under scenarios mimicking real testing situations. This section mainly demonstrates the impact of main effects and the interaction effects of the manipulated factors with respect to the error measures and model selection criteria. The bias, SE, and RMSE of all parameters under the 24 conditions were detailed in Appendix A.

Item difficulty. Based on the results from four-way repeated measures ANOVA, none of the factors or interactions was significant on the bias of item difficulty estimates with at least a small effect size. However, both the estimation model and sample size had significant impacts on SE and RMSE. The estimation model had a small effect size (f=.165) on SE and a medium effect size (f=.364) on RMSE; sample size had large effect sizes on both SE (f=.546) and RMSE (f=.453).

Figure 7 presents the main effects of the estimation model and sample size on item difficulty estimates, where bars with different colors indicate different levels of the manipulated factors. To explore the differences among the four estimation models, Dunn-Sidak test was performed (see Table 7). Although differences among the SE of the estimation models did not seem large in Figure 7, all pairwise comparisons were significant except the difference between the JM-R and the HM. In terms of RMSE, the results summarized in Table 7 were more consistent with visual inspection, where only the JM-RD2 performed worse than the other estimation models. This finding suggests that ignoring the conditional dependence did not lead to significantly worse item difficulty estimates, but modeling the conditional dependence between responses and RTs without random effects would reduce the estimation accuracy of item difficulty. Both graphical and numerical representations show that increasing sample size resulted in significantly smaller SE and RMSE.

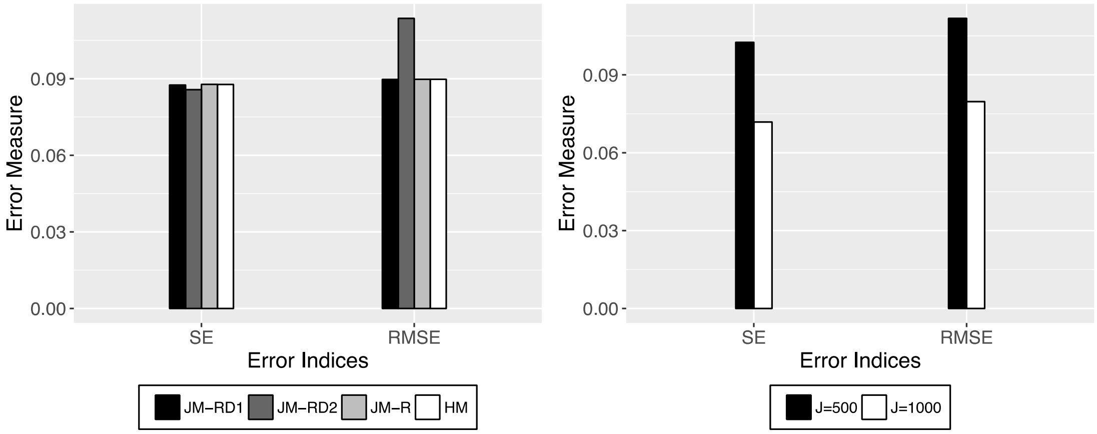


Figure 7. Significant main effects on the SE and RMSE of the item difficulty estimates.

Table 7. *Post-hoc pairwise comparison results of the estimation model on the SE and RMSE in item difficulty estimation in simulation study 1.*

		Mean Difference	
Model (m)	Model (n)	SE	RMSE
JM-RD1	JM-RD2	.002*	-.024*
	JM-R	<.001*(-)	<.001
	HM	<.001*(-)	<.001
JM-RD2	JM-R	-.002*	.024*
	HM	-.002*	.024*
JM-R	HM	<.001	<.001

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Time discrimination. Figure 8 and Table 8 present the results for the significant main effects and post-hoc pairwise comparison for the time discrimination parameters. The estimation model was a significant factor with large effect sizes on bias (*f*=.599) and RMSE (*f*=.436), and with a medium effect size on SE (*f*=.219). Sample size was significant with a large effect size on SE (*f*=.736) and a small effect size on RMSE (*f*=.140).

The mean bias depicted in Figure 8 reflects that all four estimation models consistently underestimated the time discrimination parameters. While there was no significant difference between the JM-RD1 and the JM-R, the two models with random effects produced smaller systematic and total errors than the JM-RD2. The HM yielded significantly smaller random error than the other estimation models, yet it performed the worst in terms of bias and RMSE. This finding is expected as the underspecified models usually have less uncertainty in the estimation process and thus smaller SE; but are generally more biased because some important effects or model parameters are omitted. Similar to the item difficulty parameters, a sample size of 1,000 yielded significantly lower SE and RMSE than a sample size of 500 for all four models. Detailed error indices under each condition are reported in Tables A2a to A2c.

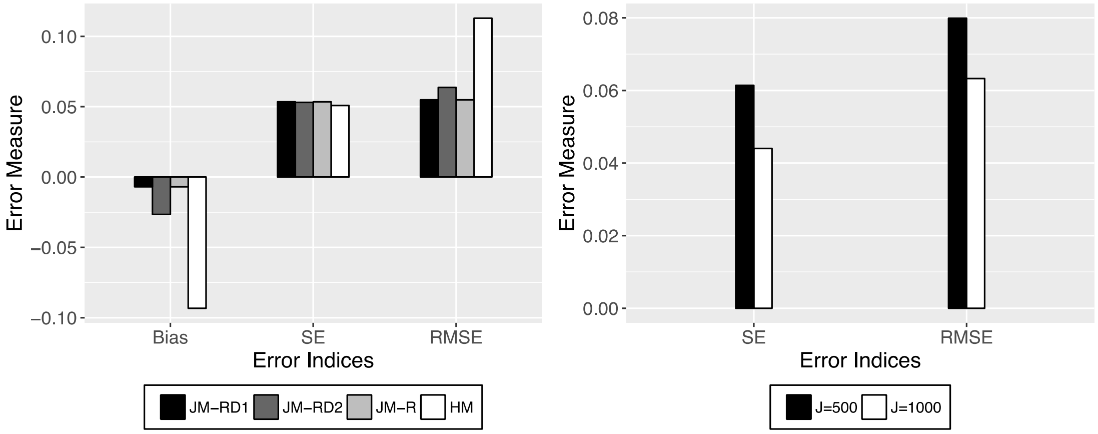


Figure 8. Significant main effects on the bias, SE, and RMSE of the time discrimination estimates.

Table 8. *Post-hoc pairwise comparison results of the estimation model on the bias, SE, and RMSE in time discrimination estimation in simulation study 1.*

		Mean Difference		
Model (m)	Model (n)	Bias	SE	RMSE
JM-RD1	JM-RD2	-.019*	<.001*(+)	-.009*
	JM-R	<.001	<.001	<.001
	HM	-.087*	.003*	-.058*
JM-RD2	JM-R	.019*	<.001*(-)	.009*
	HM	-.067*	.002*	-.049*
JM-R	HM	-.087*	.003*	-.058*

Note. *p<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Time intensity. In terms of the time intensity parameters, the estimation model was a significant factor on all three indices with large effect sizes (f=.527, .679, .579 respectively). Sample size was only significant on SE with a large effect size (f=.463). The main effects of the estimation model and sample size on bias, SE, and RMSE are depicted in Figure 9. Overall, the findings were similar to the time discrimination parameters. The two models with random effects, the JM-RD1 and the JM-R, outperformed the other estimation models on bias and RMSE, even though they produced significantly larger SE (see Table 9). In contrast with the three proposed models, the HM provided significantly smaller SE but much larger bias and RMSE in the time intensity estimates. The other significant factor, sample size, resulted in smaller SE when sample size increased from 500 to 1,000.

The time intensity estimates were negatively biased in the HM because the interpretation of the time intensity parameters in the proposed models has changed. In the HM, the time intensity parameters reflect the average time required to finish an item for the whole population. On the contrary, in the proposed models, the time intensity parameters represent the extent to which an item is time consuming for examinees who answer incorrectly to the item, or for those whose ability levels are lower than the difficulty level of the item. As a result, when data were generated from the JM-RD1, the HM would underestimate the time intensity parameters if incorrect responses were more time consuming than correct responses. In the present study, ω_0 was set at -.30, which meant that on average incorrect responses are .30 unit slower than correct responses on the logRT scale. As such, there is no surprise that the HM consistently underestimated the time intensity under all conditions (see Table A3a).

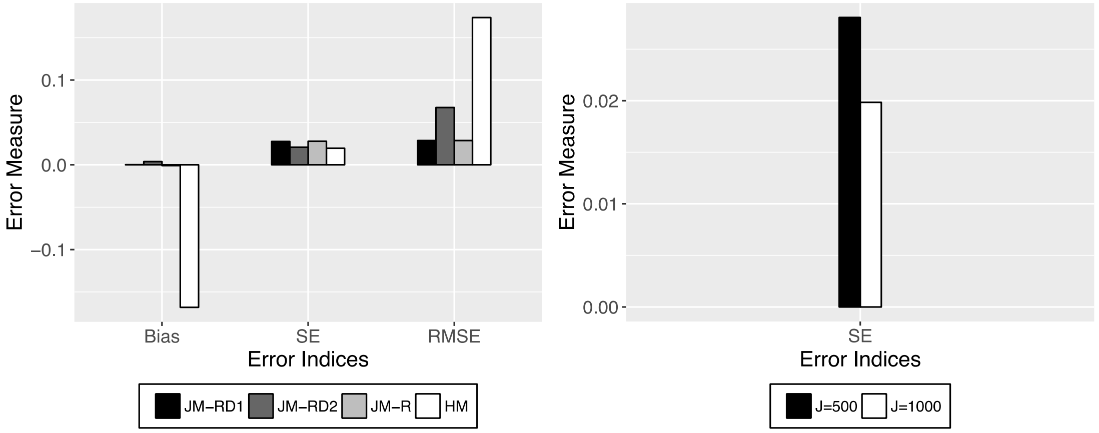


Figure 9. Significant main effects on the bias, SE, and RMSE of the time intensity estimates.

Table 9. *Post-hoc pairwise comparison results of the estimation model on the bias, SE, and RMSE in time intensity estimation in simulation study 1.*

		Mean Difference		
Model (m)	Model (n)	Bias	SE	RMSE
JM-RD1	JM-RD2	-.004	.007*	-.039*
	JM-R	-.001	<.001*(-)	<.001
	HM	-.168*	.008*	-.145*
JM-RD2	JM-R	.003	-.007*	.039*
	HM	-.164*	.001*	-.107*
JM-R	HM	-.167*	.008*	-.145*

Note. *p<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Ability parameters. While the item parameter estimates were mainly influenced by the estimation model and sample size, the results for the ability estimates were more complicated. Since the mean bias of the ability estimates was centered at zero for scale identification, ANOVA was only conducted for SE and RMSE. For the random error in the ability estimates, the estimation model was significant with a large effect size (f=.659), whereas the interaction between the estimation model and the correlation between ability and speed had a medium effect size (f=.346). The main effects of test length and the correlation between ability and speed were also significant on SE with medium effect sizes (f=.310 and .285 respectively). Only test length was significant with a small effect size (f=.202) on the total error of the ability estimates.

The significant interaction between the estimation model and the correlation between ability and speed is presented in the left panel of Figure 10, where the estimation models are marked with different line types. Overall, the SE of the ability estimates for all estimation models followed a similar pattern. As the correlation between ability and speed became stronger, the mean SE of the ability estimates from all estimation models consistently reduced. As such, the main effect of the correlation between the two latent traits was interpretable, and all three levels of correlation were significantly different (see Table 10). The HM yielded significantly smaller random error in the ability estimates than the proposed models across all levels of correlations, while the proposed models with random effects produced significantly larger SEs as expected (see Table 11 and the left panel of Figure 11). Yet, the interaction manifested itself as the discrepancy between the proposed models and the HM magnified with stronger correlation. This is an intriguing finding because studies have shown that the precision of the ability estimates increases as more information is shared between the two latent traits (Klein Entink, 2009; Patton, 2015).

Further examinations on the SD of bias showed that the SD of bias for the HM was always the largest among the four estimation models, despite that the mean bias was all constrained to zero. Moreover, the discrepancy between the HM and the other three models tended to increase when the correlation between ability and speed increased (see the right panel of Figure 10). This observation reflected that the stronger the correlation between ability and speed, the higher the variability in the bias of the ability estimates obtained from the HM. Therefore, the ability estimates from the HM for some examinees might be more biased than the three proposed models. This may also be the reason why the HM had significantly smaller SE, but the four estimation models did not differ much on RMSE of the ability estimates.

Table 10. *Post-hoc pairwise comparison results of the correlation between ability and speed on SE in ability estimation in simulation study 1.*

$\rho_{\theta\tau}$ (m)	$\rho_{\theta\tau}$ (n)	Mean Difference in SE
.2	.5	.016*
	.8	.069*
.5	.8	.053*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of $\rho_{\theta\tau}$ (n) from that of $\rho_{\theta\tau}$ (m).

Table 11. *Post-hoc pairwise comparison results of the estimation model on SE in ability estimation in simulation study 1.*

Model (m)	Model (n)	Mean Difference in SE
JM-RD1	JM-RD2	.001*
	JM-R	<.001*(-)
	HM	.005*
JM-RD2	JM-R	-.001*
	HM	.004*
JM-R	HM	.005*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

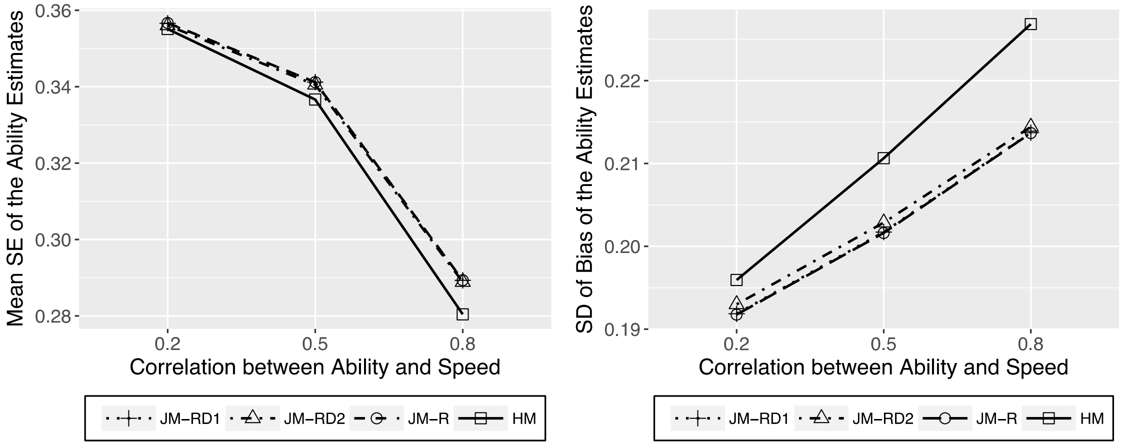


Figure 10. Two-way interaction between the estimation model and the correlation between ability and speed on the SE of the ability estimates (left panel), and on the SD of bias of the ability estimates (right panel).

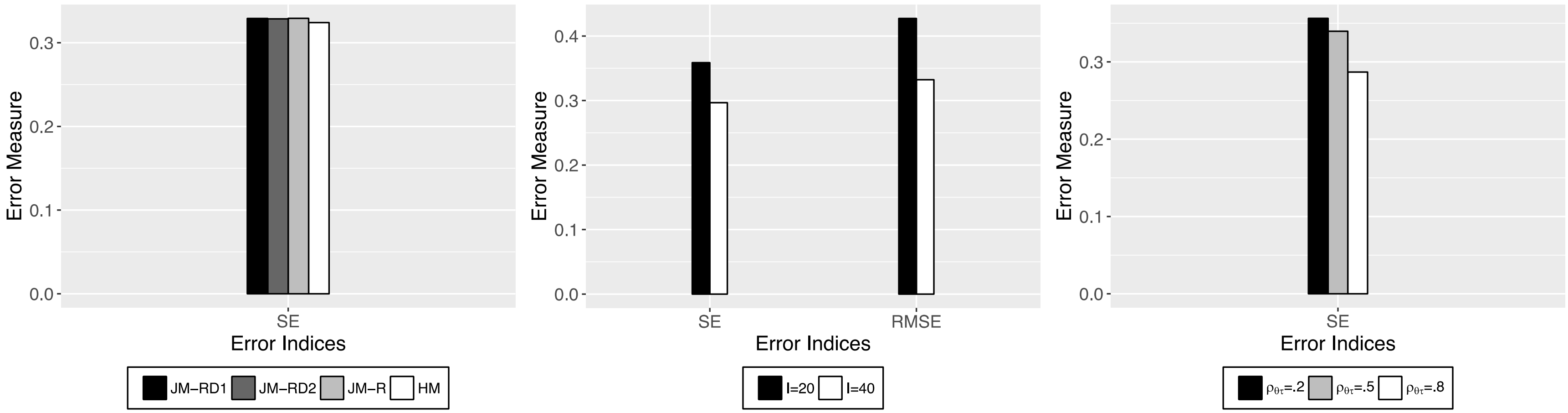


Figure 11. Significant main effects on the SE and RMSE of the ability estimates.

Speed parameters. Both the estimation model and test length had significant impacts on the SE and RMSE of the speed estimates with large effect sizes. The estimation model had large effect sizes on the random error and total error (f=.460 and .688 respectively), so did test length (f=.490 and .655 respectively). The two models with random effects, the JM-RD1 and the JM-R, performed similarly and significantly better than the model without random effects, the JM-RD2 (see Table 12). The HM produced the largest SE and RMSE than the proposed models, indicating that the accuracy of the speed parameters was negatively affected if the conditional dependence was not taken into account (see Figure 12). Similar to the ability estimates, longer test length was associated with smaller SE and RMSE.

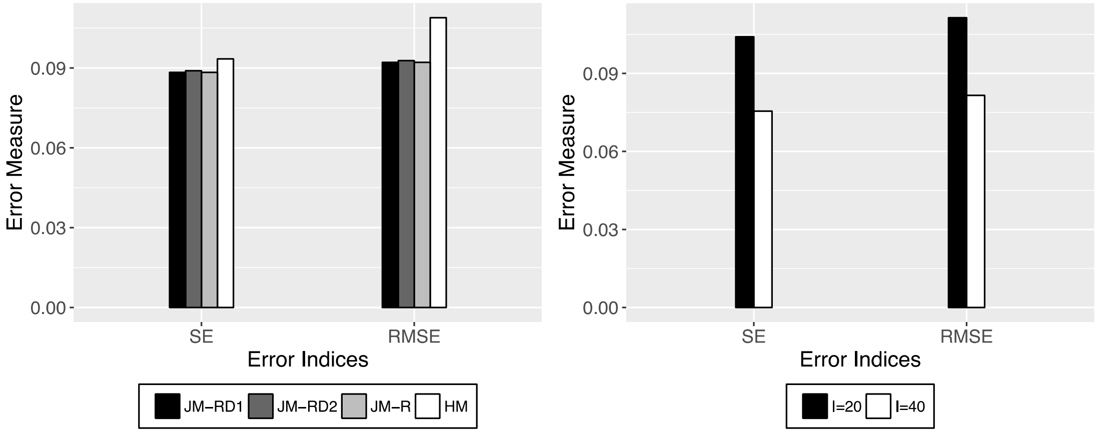


Figure 12. Significant main effects on the SE and RMSE of the speed estimates.

Table 12. *Post-hoc pairwise comparison results of the estimation model on the SE and RMSE in speed estimation in simulation study 1.*

		Mean Difference	
Model (m)	Model (n)	SE	RMSE
JM-RD1	JM-RD2	-.001*	-.001*
	JM-R	<.001	<.001
	HM	-.005*	-.017*
JM-RD2	JM-R	.001*	.001*
	HM	-.004*	-.016*
JM-R	HM	-.005*	-.017*

Note. *p<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Parameters on RT distribution shift. Regarding the parameters related to the shift in RT distributions, the recovery of the intercept ω_0, the slope ω_1, and the variance of the random effects σ_ϕ^2 was examined. The three parameters are only involved in the proposed models. Specifically, all proposed models have the intercept, the JM-RD1 and the JM-RD2 estimate the relationship between item difficulty and the shift, and the JM-RD1 and the JM-R incorporate the random effects. As such, the JM-RD1 has one unique parameter, the correlation between item difficulty and the shift $\rho_{b\lambda}$, which is a manipulated factor in the present study. This parameter was not directly estimated in the JM-RD1; but it can be resolved from the estimates of variance of item difficulty, the slope, and the variance of the random effects.

In general, the intercept and the variance of the random effects were better recovered than the intercept and the correlation between item difficulty and the shift. Increasing sample size and test length resulted in smaller bias, SE, and RMSE of the parameters on RT distribution shift with a few exceptions. For instance, it was found that the bias of the slope and of the variance of random effects from the JM-RD1 was not affected by sample size. Interestingly, the bias and RMSE of the correlation between item difficulty and the shift increased with larger sample size. On the other hand, the correlation between ability and speed and the correlation between item difficulty and the shift did not have a consistent impact on the error indices.

Figure 13 demonstrates the main effect of the estimation model on the error indices of the parameters on RT distribution shift. The random error of the intercept estimates from all three models was not remarkably different across the three estimation models, yet the two models with random effects (i.e., JM-RD1 and JM-R) yielded smaller bias and RMSE. Given that the true value of the slope was .06 and .14 when the correlation between item difficulty and the shift was .3 and .7 respectively, the bias of the slope under some conditions was not satisfactory, especially conditions with smaller correlation between item difficulty and the shift. Regardless, the JM-RD1 yielded smaller bias, SE, and RMSE than the JM-RD2. This may be because the ignored random effects in the JM-RD2 were absorbed in the slope estimates, producing larger slope estimates with higher variability.

The variance of the random effects was less biased than the slope of RT distribution shift. In general, variance estimates from the JM-R contained more systematic, random and total errors than those from the JM-RD1. Since incorporating item difficulty in modeling the shift explained a certain amount of total variance, the variance of the random effects was recovered with higher precision in the JM-RD1.

As a manipulated factor, the correlation between item difficulty and the shift from the JM-RD1 was negatively biased across all conditions. Further, the three sources of errors were greater as the true correlation between item difficulty and the shift became stronger (see Table 13). The estimated correlation between item difficulty and the shift was about 40% downward biased for both levels of true correlation, with $\rho_{b\lambda} = .3$ yielding larger bias. The large estimation error of $\rho_{b\lambda}$ could be attributed to the estimation errors of the variance of item difficulty, the slope, and the variance of the random effects. Additionally, the empirical correlation between the true item difficulty and the true RT shift would not be equal to the true correlation due to the randomness in the data generation process. Therefore, it is expected that $\rho_{b\lambda}$ would not be recovered as well as the intercept ω_0.

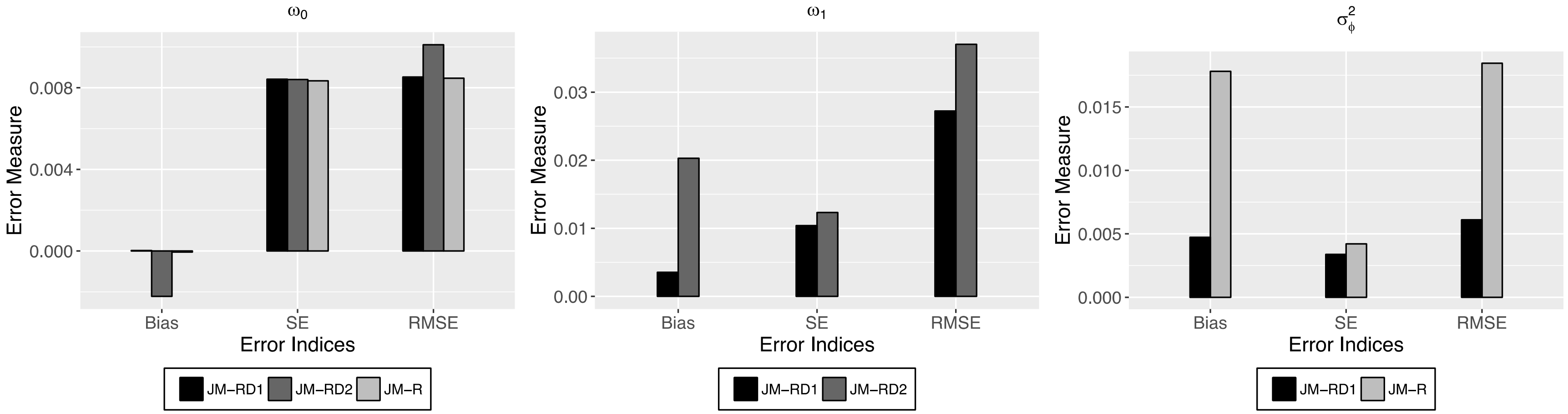


Figure 13. Main effect of the estimation model on the bias, SE, and RMSE of the parameter estimates of RT distribution shift.

Table 13. *Bias, SE, and RMSE of $\hat{\rho}_{b\lambda}$ from the JM-RD1 at different levels of $\rho_{b\lambda}$.*

		$\hat{\rho}_{b\lambda}$	
$\rho_{b\lambda}$	Bias	SE	RMSE
.3	-.125	.026	.142
.7	-.259	.034	.262

Item mean vector and item covariance matrix. The estimation accuracy of five parameters was evaluated, including the mean and variance of item difficulty, the mean and variance of time intensity, and the correlation between item difficulty and time intensity. The variances of item difficulty and time intensity were consistently overestimated regardless of the estimation model, whereas the correlation was underestimated with the JM-RD1, the JM-RD2, the JM-R but overestimated with the HM. For all five parameters, increasing sample size and test length resulted in smaller bias, SE and RMSE with a few exceptions. In general, the two manipulated correlations did not noticeably influence the error indices. The detailed bias, SE, and RMSE for the second-level item parameters under each simulation condition are presented in Tables A10 to A14.

The impact of the estimation model demonstrated an interesting pattern on the recovery of the item mean vector and item covariance matrix (see Figures 14 and 15). No big difference was found between the HM and the proposed models with respect to the two parameters solely related to the IRT model, the mean and the variance of item difficulty. However, for the mean and the variance of time intensity, and the correlation between item difficulty and time intensity, the estimates from the HM contained large systematic and total errors compared to the proposed models. In particular, the mean of time intensity from the HM was remarkably negatively biased, which may also be due to the change in the definition of time intensity parameters. The correlation between item difficulty and time intensity and the variance of time intensity were inflated in the HM, where the conditional dependence between responses and RTs may have been absorbed.

The recovery of the mean of time intensity and the correlation between item difficulty and time intensity was even worse when the correlation between item difficulty and the shift was strong. Table 14 reports the bias of the mean of time intensity, and the bias and RMSE of the correlation between item difficulty and time intensity. The HM consistently produced less accurate mean of time intensity and correlation between item difficulty and time intensity when a stronger correlation between item difficulty and the shift was ignored. This finding aligns with the expectation that the stronger the correlation between item difficulty and the shift, the more important it is to account for it.

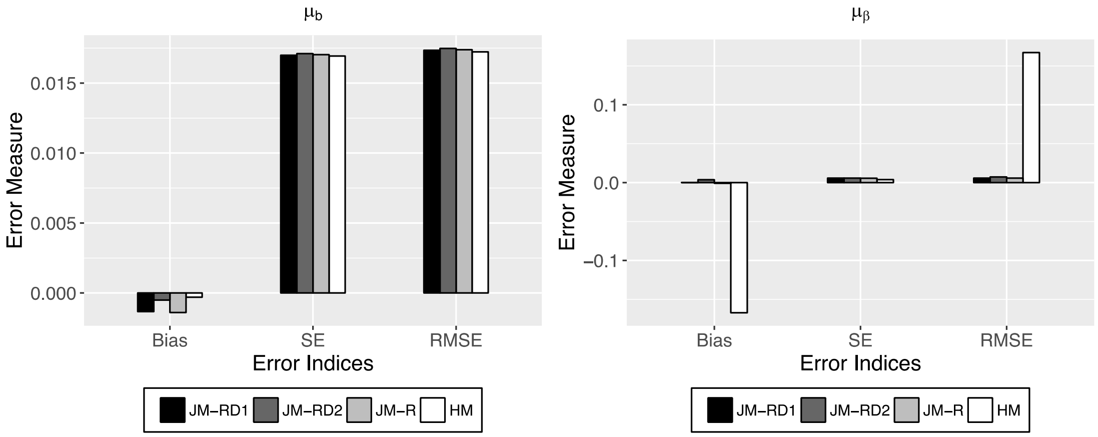


Figure 14. Main effect of the estimation model on the bias, SE, and RMSE of the item mean vector.

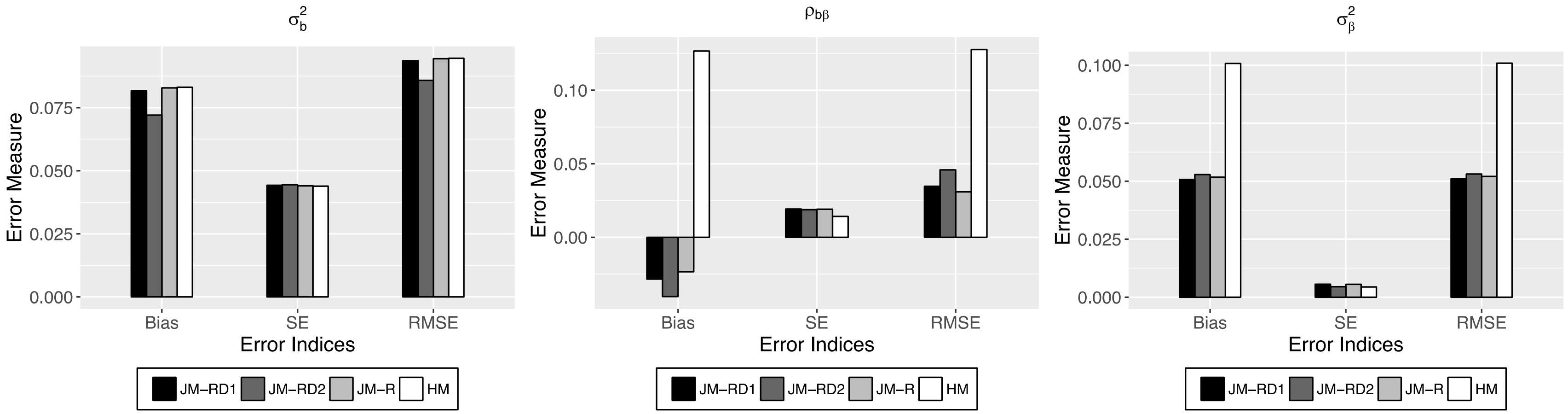


Figure 15. Main effect of the estimation model on the bias, SE, and RMSE of the item covariance matrix.

Table 14. *Bias and RMSE of $\hat{\mu}_\beta$ and $\hat{\rho}_{b\beta}$ from the HM at different levels of $\rho_{b\lambda}$.*

	$\hat{\mu}_\beta$	$\hat{\rho}_{b\beta}$	
$\rho_{b\lambda}$	Bias	Bias	RMSE
.3	-.161	.104	.105
.7	-.174	.150	.150

Person covariance matrix. As the ability and speed parameters were both constrained to have a mean of zero, only the variances of ability and speed, and the correlation between ability and speed were included as second-level person parameters. Similar to the findings from the item covariance matrix, the four estimation models performed similarly in terms of the ability variance, but the HM produced large bias and RMSE regarding the correlation between the two latent traits and the speed variance (see Figure 16). All four models overestimated the speed variance, yet the bias of the estimates from the HM was much larger than the other three. The RT model-related second-level person parameters may be inflated due to the conditional dependence between responses and RTs as well.

Sample size and test length often affected the error indices for the variances of ability and speed, and the correlation between ability and speed, except for a few cases. The correlation between speed and ability did not influence the variance estimates much, despite that the bias of speed variance from the HM increased from .015, .031 to .047 when the correlation between the two latent traits was varied at .2, .5, .8. This finding implies that when a stronger correlation between ability and speed is present, ignoring the conditional dependence would have a greater impact on the speed variance estimates.

Regarding the correlation between the two latent traits, the bias in the estimated correlation between ability and speed from the HM reduced when the true values increased (see the left panel in Figure 17). Only the bias from the HM was presented as the correlation between ability and speed did not affect the bias from other estimation models consistently. Additionally, the random error and the total error in the correlation between ability and speed also shrank as the true values became stronger for all four estimation models. As demonstrated in Figure 17, the estimation accuracy of the correlation between the two latent traits was lower when its magnitude was small. The last manipulated factor, the correlation between item difficulty and the shift, did not present a manifest main effect or interaction effect on the error indices of the person covariance matrix.

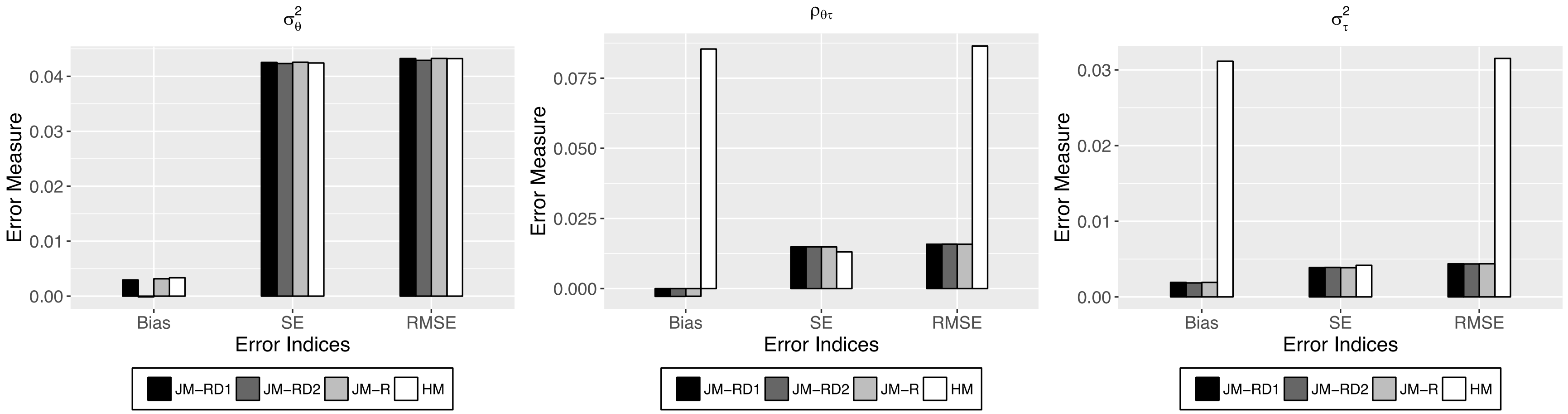


Figure 16. Main effect of the estimation model on the bias, SE, and RMSE of the person covariance matrix.

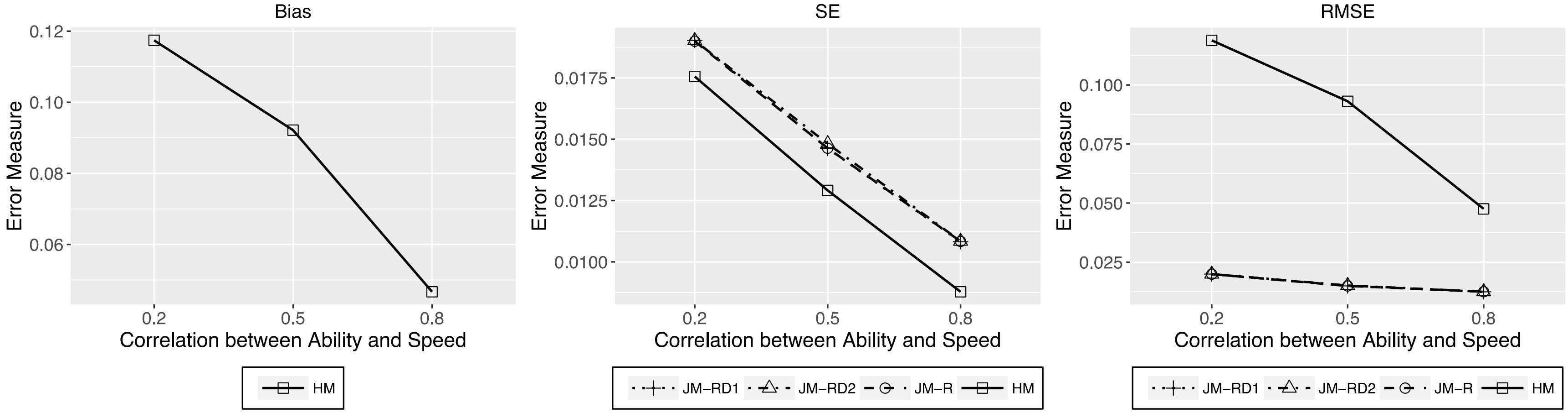


Figure 17. Two-way interaction between the estimation model and the correlation between ability and speed on the bias, SE, and RMSE of the estimated correlation between ability and speed.

Model fit indices. Table 15 reports the frequency of identifying each of the four estimation models as the best fitting model in 30 replications under each simulation condition based on deviance, AIC, AICc, BIC and DIC respectively. None of the model fit indices identified the JM-RD2 and the HM as the best fitting model under any conditions. With respect to the two selected estimation models, although the JM-RD1 offered significant improvement on some parameter estimates than the JM-R, the model fit indices generally preferred the JM-R over the JM-RD1. Deviance, as a goodness-of-fit measure without any penalty on the number of parameters, favored the JM-RD1 and the JM-R approximately equally under most conditions. There was also a general trend that when the correlation between item difficulty and the shift was stronger, the JM-RD1 was preferred by deviance more than the JM-R because it modeled the shift as a function of item difficulty.

Since AIC added a penalty term of the number of parameters to the deviance function, the JM-R was chosen as the best fitting model in more replications than deviance. With increasing penalties of the number of parameters, AICc and BIC gradually moved towards favoring the JM-R. As the model fit index with the largest penalty term in the present study, BIC chose the JM-R 100% of the replications under all conditions. DIC performed similarly to deviance, which did not distinguish the two models with random effects well.

In general, model fit indices discussed in the present study did not perform well in terms of identifying the true data generating model. Yet, they always identify one of the models that considered the conditional dependence between responses and RTs with random effects as the best fitting model. Comparing the four estimation models in each replication, the model fit indices of the JM-RD1 and the JM-R were usually close. The JM-RD2 had larger model fit indices than the two models with random effects, but it still outperformed the HM on all fit indices. This finding reflects that ignoring the conditional dependence between responses and RTs results in a remarkable drop in the overall model goodness-of-fit.

Table 15. *Frequency of identifying each model as the best fitting model in simulation study 1.*

				Deviance				AIC				AICc				BIC				DIC			
J	I	$\rho_{\theta\tau}$	$\rho_{\theta\tau}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	15	0	15	0	3	0	27	0	0	0	30	0	0	0	30	0	12	0	18	0
			.7	18	0	12	0	7	0	23	0	4	0	26	0	0	0	30	0	16	0	14	0
		.5	.3	14	0	16	0	3	0	27	0	0	0	30	0	0	0	30	0	11	0	19	0
			.7	18	0	12	0	2	0	28	0	0	0	30	0	0	0	30	0	15	0	15	0
		.8	.3	16	0	14	0	1	0	29	0	1	0	29	0	0	0	30	0	16	0	14	0
			.7	15	0	15	0	5	0	25	0	0	0	30	0	0	0	30	0	16	0	14	0
	40	.2	.3	14	0	16	0	5	0	25	0	0	0	30	0	0	0	30	0	19	0	11	0
			.7	19	0	11	0	6	0	24	0	0	0	30	0	0	0	30	0	16	0	14	0
		.5	.3	14	0	16	0	5	0	25	0	1	0	29	0	0	0	30	0	14	0	16	0
			.7	16	0	14	0	4	0	26	0	1	0	29	0	0	0	30	0	20	0	10	0
		.8	.3	14	0	16	0	2	0	28	0	0	0	30	0	0	0	30	0	12	0	18	0
			.7	16	0	14	0	3	0	27	0	1	0	29	0	0	0	30	0	14	0	16	0
1000	20	.2	.3	17	0	13	0	4	0	26	0	4	0	26	0	0	0	30	0	13	0	17	0
			.7	19	0	11	0	8	0	22	0	4	0	26	0	0	0	30	0	15	0	15	0
		.5	.3	20	0	10	0	8	0	22	0	6	0	24	0	0	0	30	0	17	0	13	0
			.7	15	0	15	0	5	0	25	0	2	0	28	0	0	0	30	0	16	0	14	0
		.8	.3	18	0	12	0	10	0	20	0	8	0	22	0	0	0	30	0	14	0	16	0
			.7	19	0	11	0	10	0	20	0	8	0	22	0	0	0	30	0	19	0	11	0
	40	.2	.3	10	0	20	0	2	0	28	0	1	0	29	0	0	0	30	0	11	0	19	0
			.7	15	0	15	0	7	0	23	0	5	0	25	0	0	0	30	0	20	0	10	0
		.5	.3	9	0	21	0	2	0	28	0	2	0	28	0	0	0	30	0	17	0	13	0
			.7	11	0	19	0	5	0	25	0	2	0	28	0	0	0	30	0	14	0	16	0
		.8	.3	14	0	16	0	6	0	24	0	2	0	28	0	0	0	30	0	12	0	18	0
			.7	15	0	15	0	7	0	23	0	6	0	24	0	0	0	30	0	14	0	16	0

To summarize, the first-level item parameters were significantly affected by sample size, whereas test length was a significant factor on the first-level person parameters. Table 16 presents a summary of the effect sizes of the significant effects. Although the correlation between item difficulty and the shift was varied in simulation study 1, it did not have a significant impact on the recovery of the five first-level parameters, hence was not listed in the Table 16. Regarding other parameters for which ANOVA was not performed, sample size and test length generally resulted in smaller estimation errors, while the two manipulated factors mainly affected the accuracy of parameter estimates from the HM. Except the slope on RT distribution shift and the correlation between item difficulty and the shift, the parameters of interest could be well recovered with the proposed models.

Comparing the four estimation models, the JM-RD1 and the JM-R performed in a similar fashion as they both considered the random effects on the shift of the RT distributions. The JM-RD2 often yielded less random error but more systematic error in the parameter estimates than the JM-RD1 and the JM-R. The HM, even though usually produced the most stable estimates, had the largest systematic and total errors especially for RT model-related parameters. Parameters such as time discrimination, time intensity, speed, the mean and variance of time intensity, the speed variance, the correlation between item difficulty and time intensity, and the correlation between ability and speed, were largely biased if the conditional dependence was omitted.

In terms of model selection criteria, deviance and DIC were preferred over AIC, AICc, and BIC with regard to identifying the data generating model. Even so, deviance and DIC were still not able to distinguish the JM-RD1 and the JM-R.

Table 16. *Summary of the effect sizes of the significant effects in simulation study 1.*

		Significant Effects				
Parameters	Error Indices	Model	J	I	$\rho_{\theta\tau}$	Model*$\rho_{\theta\tau}$
b_i	Bias					
	SE	small	large			
	RMSE	medium	large			
α_i	Bias	large				
	SE	small	large			
	RMSE	large	small			
β_i	Bias	large				
	SE	large	large			
	RMSE	large				
θ_j	SE	large		medium	medium	medium
	RMSE			small		
τ_j	SE	large		large		
	RMSE	large		large		

4.1.2 Simulation Study 2

In simulation study 2, the performance of the three models conditioning on item-person distance (i.e., the JM-DD1, the JM-DD2, and the JM-D) was evaluated with the HM. The 24 simulation conditions were the same as in simulation study 1, but data were simulated from the JM-DD1 rather than the JM-RD1. Overall, the findings from simulation study 2 were similar to simulation study 1, especially for the RT model-related parameters. Yet, some major inconsistencies were found in the results for the IRT model-related parameters. In this section, the results for all parameters are displayed in tables and figures as in simulation study 1, and the important discrepancies between simulation studies 1 and 2 were highlighted regarding the IRT model-related parameters. The bias, SE, and RMSE under each simulation condition are presented in Appendix B in detail.

Item difficulty. Same as in simulation study 1, none of the factors had a significant impact on the systematic errors of the item difficulty estimates. For the random error, the estimation model and sample size were both significant with medium effect sizes (*f*=.255 and .387 respectively). The same two factors were significant for the total error with a small effect size (*f*=.177) and a medium effect size (*f*=.368) respectively. As depicted in Figure 18, the random error and total error in the item difficulty estimates shrank significantly with increasing sample size.

In contrast to simulation study 1 where only the JM-RD2 performed worse than the other three estimation model, the results presented in Table 17 revealed that the true data generating model, the JM-DD1, was significantly better than the other three estimation models in terms of both SE and RMSE. Although performed slightly worse than the JM-DD1, the JM-D still yielded significantly smaller random and total errors than the JM-DD2 and the HM. The JM-DD2 provided significantly lower random error than the HM; but did not outperform the HM in terms of the total error. In other words, while ignoring the conditional dependence between responses and RTs did not negatively affect the item difficulty parameters in simulation study 1, employing the HM as the estimation model led to significantly larger SE and RMSE when data were generated from the JM-DD1. A possible explanation is that more information from the RTs could be utilized in estimating the item difficulty parameters when the shift indicator is based on the ability and item difficulty, thus providing higher estimation precision.

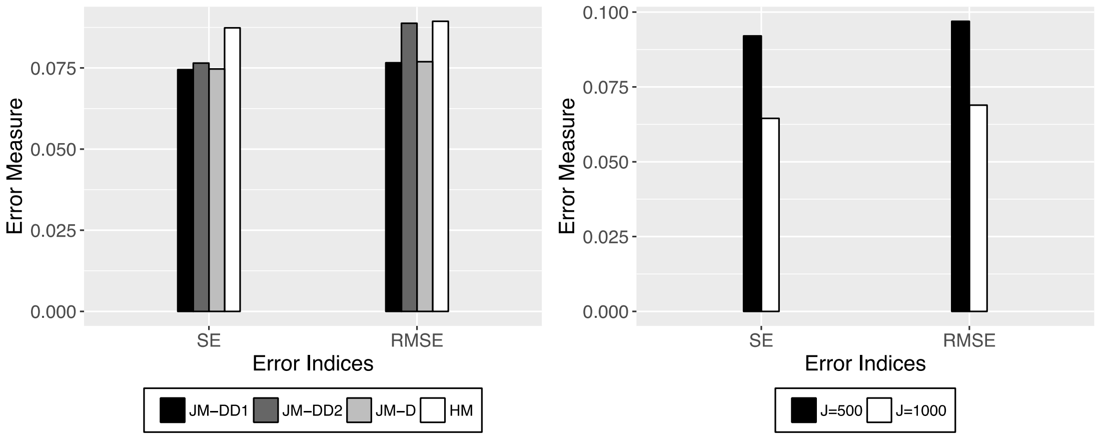


Figure 18. Significant main effects on the SE and RMSE of the item difficulty estimates.

Table 17. *Post-hoc pairwise comparison results of the estimation model on the SE and RMSE in item difficulty estimation in simulation study 2.*

		Mean Difference	
Model (m)	Model (n)	SE	RMSE
JM-DD1	JM-DD2	-.002*	-.012*
	JM-D	<.001*(-)	<.001*(-)
	HM	-.012*	-.012*
JM-DD2	JM-D	.002*	.011*
	HM	-.010*	<.001
JM-D	HM	-.012*	-.012*

Note. *p<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Time discrimination. The estimation model was a significant factor on bias, SE, and RMSE with large (f=.591), small (f=.245), and medium (f=.316) effect sizes respectively. Sample size was also a significant factor with a large effect size (f=.776) on SE and a small effect size (f=.249) on RMSE. While the random errors from the four estimation models were rather equivalent, the JM-DD1 and the JM-D produced significantly smaller bias and RMSE than the JM-DD2 and the HM (see Figure 19 and Table 18). As expected, the HM yielded the most biased time discrimination estimates with the largest total error.

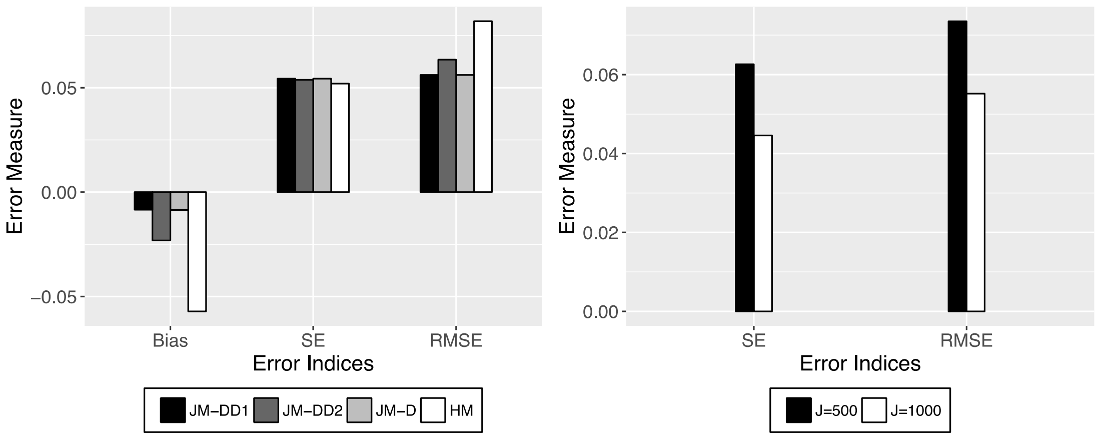


Figure 19. Significant main effects on the bias, SE, and RMSE of the time discrimination estimates.

Table 18. *Post-hoc pairwise comparison results of the estimation model on the bias, SE, and RMSE in time discrimination estimation in simulation study 2.*

		Mean Difference		
Model (m)	Model (n)	Bias	SE	RMSE
JM-DD1	JM-DD2	-.015*	.001*	-.007*
	JM-D	<.001	<.001	<.001
	HM	-.049*	.002*	-.026*
JM-DD2	JM-D	.014*	-.001*	.007*
	HM	-.034*	.002*	-.018*
JM-D	HM	-.049*	.002*	-.026*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Time intensity. Again, the estimation model had large effect sizes on bias, SE, and RMSE (*f*=.434, .415, and .474, respectively). As shown in Figure 20, sample size was associated with significantly lower SE with a small effect size (*f*=.128). As expected, even though the HM had the lowest mean SE, it had the most biased estimates for time intensity, which also led to the largest total error in time intensity estimates among the four estimation models (see Figure 20 and Table 19). The main difference between simulation studies 1 and 2 was that the bias in time intensity estimates from the JM-RD1 and the JM-D did not differ significantly, while the JM-DD1 provided significantly smaller bias than the JM-D.

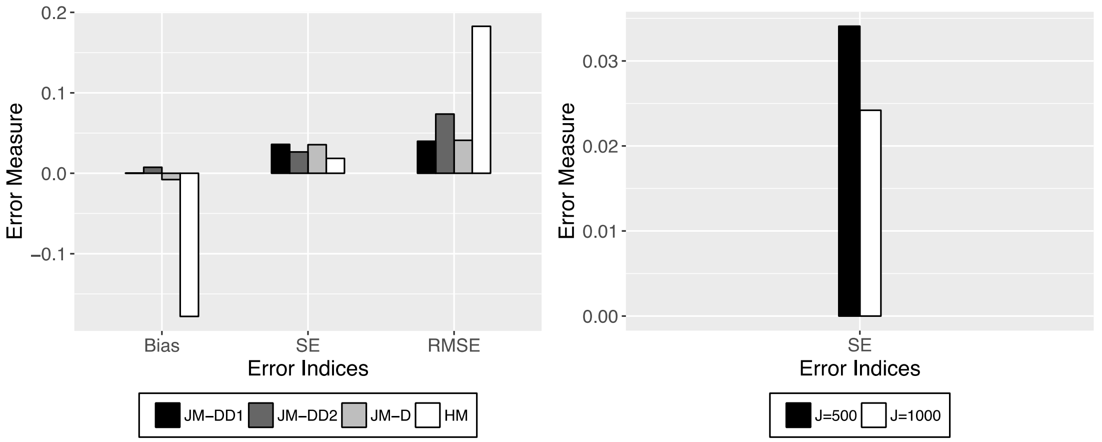


Figure 20. Significant main effects on the bias, SE, and RMSE of the time intensity estimates.

Table 19. *Post-hoc pairwise comparison results of the estimation model on the bias, SE, and RMSE in time intensity estimation in simulation study 2.*

		Mean Difference		
Model (m)	Model (n)	Bias	SE	RMSE
JM-DD1	JM-DD2	-.006	.009*	-.033*
	JM-D	-.008*	.001*	-.001
	HM	-.178*	.018*	-.143*
JM-DD2	JM-D	-.001	-.009*	.032*
	HM	-.172*	.009*	-.110*
JM-D	HM	-.170*	.017*	-.142*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Ability parameters. Same as in simulation study 1, the mean of the ability estimates was constrained to be zero under each replication for scale identification. The SE of the ability estimates was significantly affected by the estimation model, test length, and the correlation between ability and speed with medium effect sizes (*f*=.356, .352, and .325, respectively). The interaction effect between the estimation model and the correlation between the two latent traits was significant in simulation study 1, but not in simulation study 2. Although none of the manipulated factors had an impact on the RMSE of the ability estimates in simulation study 1, in simulation study 2 the estimation model and test length were significant factors on RMSE with a large effect size (*f*=.547) and a small effect size (*f*=.199) respectively.

While the HM provided the smallest SE among the four estimation models in simulation study 1, the HM yielded the largest SE and RMSE when the JM-DD1 was the data generating model. As presented in Figure 21 and Table 20, the JM-DD1 outperformed the three underspecified models on both SE and RMSE significantly. These findings may also result from the specification of the shift indicator in the JM-DD1. Similar to item difficulty, the information from RTs directly contributed to the estimation of the ability parameters in the models conditioning on item-person distance. Since the HM only utilized RT information indirectly to estimate IRT model-related parameters, it is expected that the HM would perform worse than the proposed models on these parameters as well. Nonetheless, the SE of the ability estimates gradually decreased as the correlation increased from .2, .5, to .8, which was consistent with simulation study 1 (see Figure 21 and Table 21).

Table 20. *Post-hoc pairwise comparison results of the estimation model on the SE and RMSE in ability estimation in simulation study 2.*

		Mean Difference	
Model (m)	Model (n)	SE	RMSE
JM-DD1	JM-DD2	-.007*	-.011*
	JM-D	<.001*(-)	<.001*(-)
	HM	-.031*	-.043*
JM-DD2	JM-D	.007*	.010*
	HM	-.024*	-.032*
JM-D	HM	-.030*	-.043*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Table 21. *Post-hoc pairwise comparison results of the correlation between ability and speed on the SE in ability estimation in simulation study 2.*

$\rho_{\theta\tau}$ (m)	$\rho_{\theta\tau}$ (n)	Mean Difference in SE
.2	.5	.022*
	.8	.074*
.5	.8	.053*

Note. *p<*.05. The mean difference in the error indices is calculated by subtracting each error index of $\rho_{\theta\tau}$ (n) from that of $\rho_{\theta\tau}$ (m).*

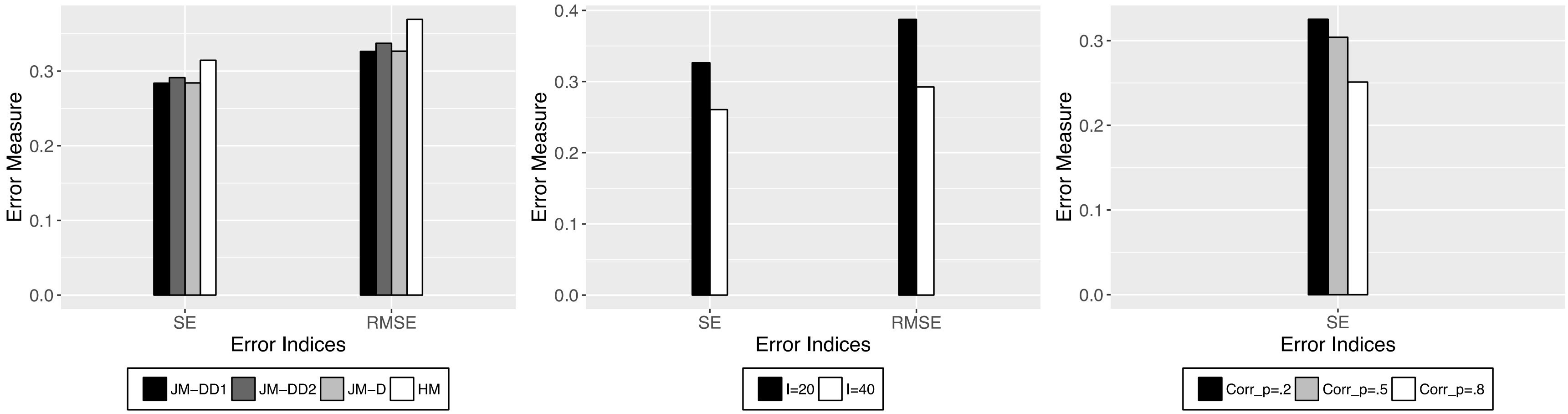


Figure 21. Significant main effects on the SE and RMSE of the ability estimates.

Speed parameters. Regarding the random error, the interaction between the estimation model and the correlation between ability and speed was a significant factor with a small effect size (*f*=.208), and test length was significant with a large effect size (*f*=.625). This is different from the ANOVA results for speed estimates in simulation study 1, where the main effects of the estimation model and test length were significant on SE. For the total error, both the estimation model and test length had large effect sizes (*f*=.490 and .476 respectively). The HM yielded the largest total error compared to the proposed models (see Figure 22). While the JM-RD1 and the JM-R performed similarly on RMSE in simulation study 1, the JM-DD1 had significantly smaller total error than the JM-D (see Table 22).

As the correlation between ability and speed became stronger, the random error of speed estimates consistently decreased for all four models (see Figure 23). In addition, the three models that considered the conditional dependence between responses and RTs (i.e., the JM-DD1, the JM-DD2, and the JM-D) performed in a similar way under each level of manipulated correlation. When the correlation was .2 and .5, the HM yielded smaller SE in speed estimates than the proposed models. When the correlation was .8, however, the three proposed models produced smaller random errors in the speed estimates than the HM. This finding suggests that the impact of ignoring the conditional dependence is not consistent across all levels of correlations. Accounting for conditional dependence in the joint modeling of responses and RTs is more likely to reduce the SE in the speed estimates when the correlation between ability and speed is larger than .5.

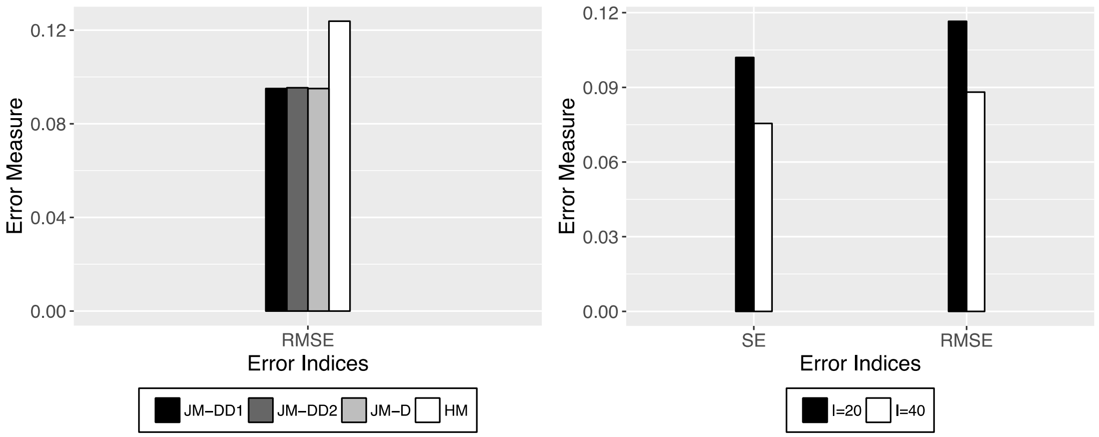


Figure 22. Significant main effects on the SE and RMSE of the speed estimates.

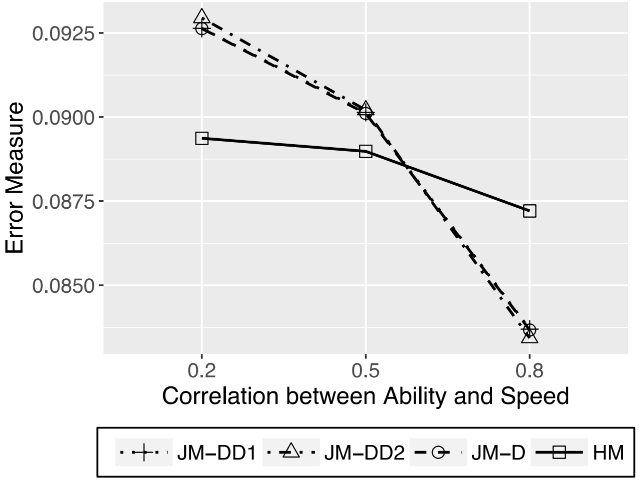


Figure 23. Significant two-way interaction between the estimation model and the correlation between ability and speed on the SE of speed estimates.

Table 22. *Post-hoc pairwise comparison results of the estimation model on the RMSE in speed estimation in simulation study 2.*

Model (m)	Model (n)	Mean Difference in RMSE
JM-DD1	JM-DD2	<.001*(-)
	JM-D	<.001*(-)
	HM	-.029*
JM-DD2	JM-D	<.001*(+)
	HM	-.028*
JM-D	HM	-.029*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Parameters on RT distribution shift. Similar to simulation study 1, the intercept and the variance of the random effects were recovered better than the slope and the correlation between item difficulty and the shift. Compared to the underspecified models, the JM-DD1 yielded estimates with the least estimation error for all parameters on RT distribution shift (see Figure 24). Overall, larger sample size and longer test length led to smaller errors in the estimates. The two manipulated correlations did not have an evident impact on the estimates, except that the bias and RMSE in the estimated correlation between item difficulty and the shift increased proportionally with the true correlation between item difficulty and the shift (see Table 23).

Table 23. *Bias and RMSE of $\hat{\rho}_{b\lambda}$ from the JM-DD1 at different levels of $\rho_{b\lambda}$.*

	$\hat{\rho}_{b\lambda}$	
$\rho_{b\lambda}$	Bias	RMSE
.3	-.118	.157
.7	-.268	.277

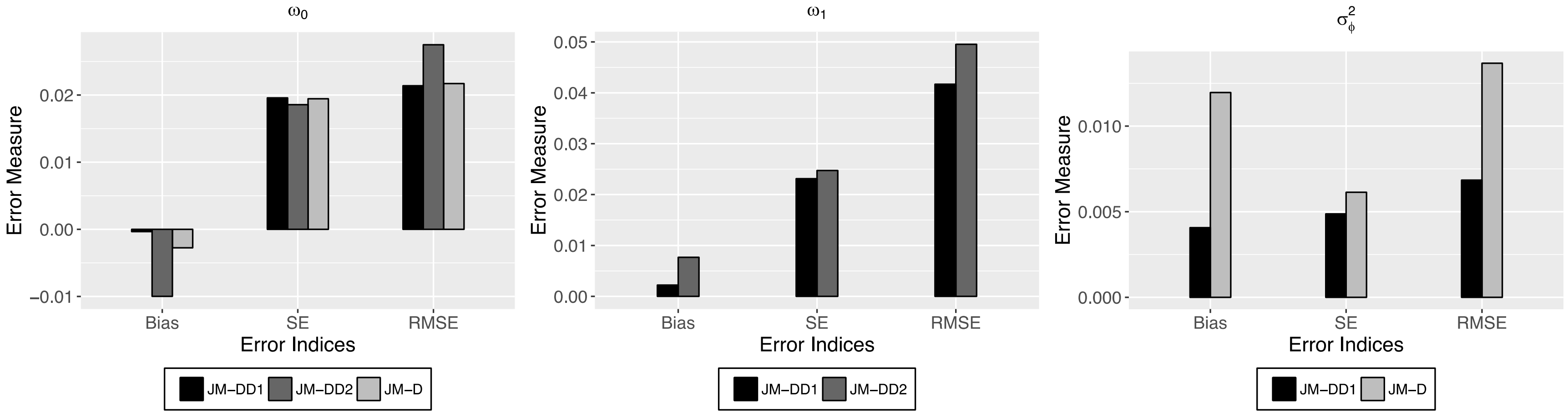


Figure 24. Main effect of the estimation model on the bias, SE, and RMSE of the parameter estimates on RT distribution shift.

Mean vector and covariance matrices. Generally speaking, the findings on the elements in item mean vector and the covariance matrices were similar to simulation study 1. The unique findings from simulation study 2 were discussed in three aspects, mainly regarding IRT model-related parameters.

First, the estimated mean of item difficulty from the HM contained more estimation error than the proposed model (see Figure 25). While the mean of item difficulty in simulation study 1 was not affected by the choice of estimation model, the JM-DD1 outperformed the HM in terms of systematic, random, and total errors. This is consistent with the findings from item difficulty parameters. The JM-DD1 utilized information from RT directly with the correctly specified model structure, thus yielding more accurate item difficulty estimates than other estimation models.

Second, even though the JM-RD2 had the smallest bias and RMSE of the estimated variance of item difficulty, the results indicated the opposite for the JM-DD2. In simulation study 2, the JM-DD2 in fact produced the largest systematic and total errors in the variance of item difficulty, despite similar random errors from all estimation models (see Figure 26). Likewise, the JM-DD2 also resulted in the largest bias and RMSE in the ability variance (see Figure 27). Therefore, omitting the random effects on RT distribution shift had opposite impact on the recovery of item difficulty variance and ability variance, depending on which shift indicator was employed in the model.

Third, the influence of the correlation between ability and speed and the correlation between item difficulty and the shift was different. The bias and RMSE in the mean of item difficulty, the correlation between item difficulty and time intensity, and the speed variance from the HM were consistently affected by the manipulated correlations (see Table 24). In simulation study 1, however, only the error indices with an asterisk were affected. The interaction effect between the estimation model and the correlation between the two latent traits was also manifested differently. Rather than decreasing uniformly in simulation study 1, the SE of the four models reduced slightly as the correlation between ability and speed increased from .2 to .5, but a bigger drop was found when the correlation further increased to .8 (see Figure 28). Additionally, the three proposed models did not perform as similar as in simulation study 1. The JM-DD2 yielded consistently larger SE than the JM-DD1 and the JM-D.

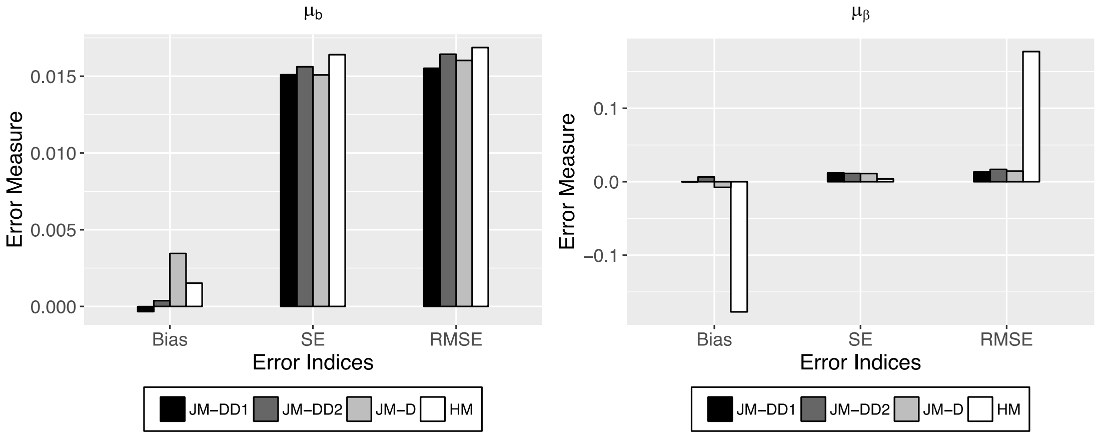


Figure 25. Main effect of the estimation model on the bias, SE, and RMSE of the item mean vector.

Table 24. *Bias and RMSE of $\hat{\mu}_\beta$, $\hat{\rho}_{b\beta}$, $\hat{\sigma}_\tau^2$ from the HM at different levels of manipulated correlations.*

	$\hat{\mu}_\beta$		$\hat{\rho}_{b\beta}$		$\hat{\sigma}_\tau^2$			$\hat{\sigma}_\tau^2$	
$\rho_{b\lambda}$	Bias*	RMSE	Bias*	RMSE*	Bias	RMSE	$\rho_{\theta\tau}$	Bias*	RMSE
.3	-.166	.166	.142	.143	.052	.052	.2	.027	.027
.7	-.188	.188	.186	.186	.054	.054	.5	.053	.053
							.8	.078	.079

Note. Error indices with an asterisk were also affected in simulation study 1.

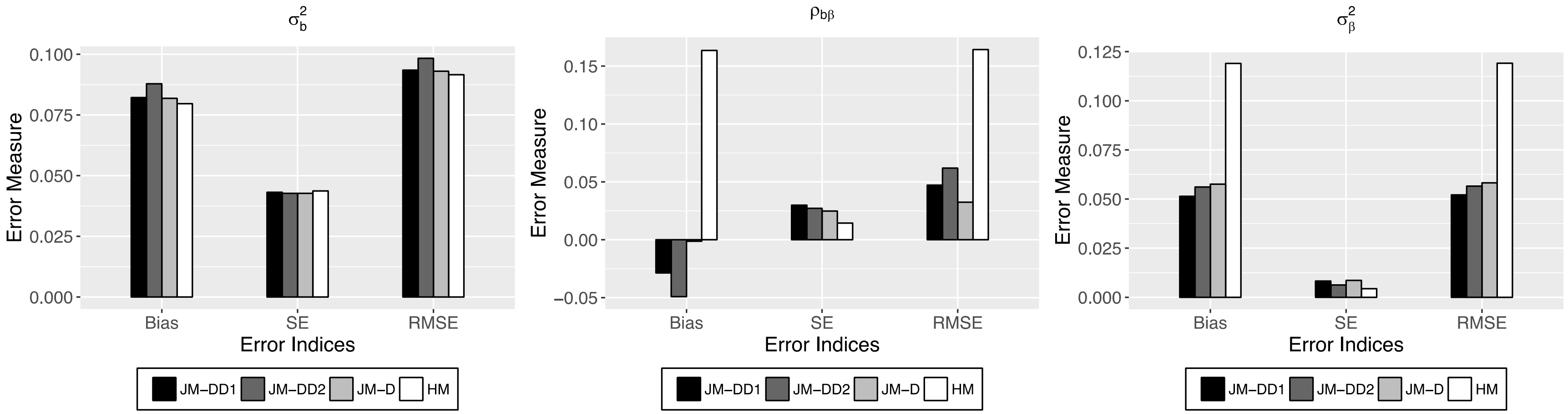


Figure 26. Main effect of the estimation model on the bias, SE, and RMSE of the item covariance matrix.

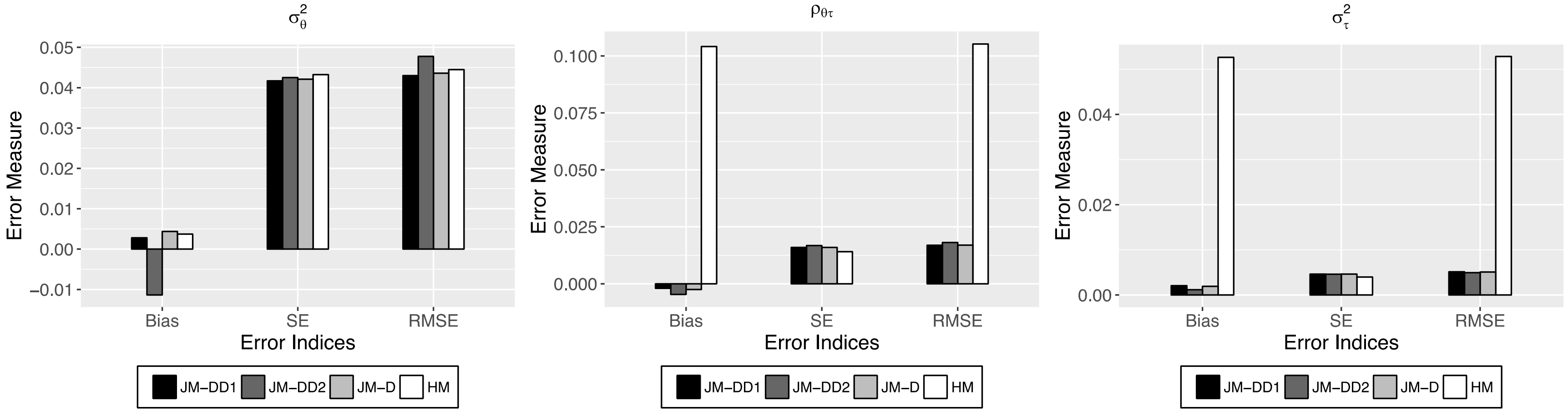


Figure 27. Main effect of the estimation model on the bias, SE, and RMSE of the person covariance matrix.

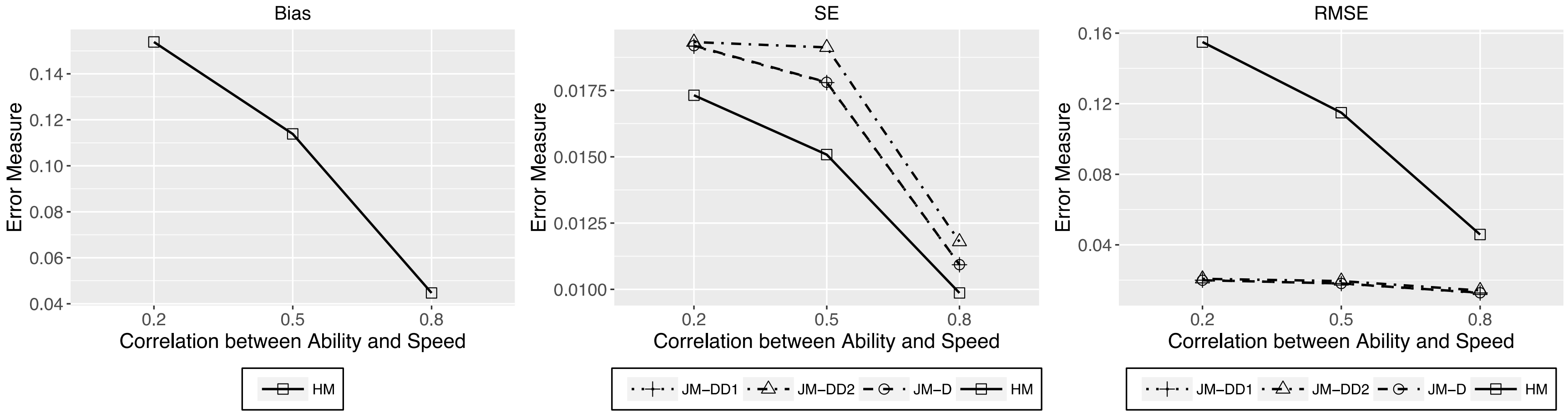


Figure 28. Two-way interaction between the estimation model and the correlation between ability and speed on the bias, SE, and RMSE of the estimated correlation between ability and speed.

Model fit indices. The frequencies of identifying each estimation model as the best fitting model based on different model fit indices are presented in Table 25. Deviance, AIC, AICc, and BIC demonstrated similar pattern as in simulation study 1, where none of them chose the JM-DD2 and the HM as the best fitting model. Specifically, deviance identified the data generating model as the best fitting model over 50% of the replications under most conditions. Compared to data generated from the JM-RD1, deviance performed better in terms of identifying the true data generating model on data generated from the JM-DD1. As such, AIC, AICc and BIC also outperformed their counterparts in simulation study 1 even though they still gradually leaned towards the underspecified JM-D as the penalty of the number of parameters became heavier. With the largest penalty term, BIC rarely selected the JM-DD1 as the best fitting model.

The performance of DIC was quite different from simulation study 1. In simulation study 2, the effectiveness of DIC was affected by the interaction between the estimation model and test length. When test length was 40, DIC mainly identified either the JM-DD1 and the JM-D as the best fitting model about 50% of the replications across conditions, which was similar to deviance. However, when test length was 20, DIC tended to favor the HM under most conditions, the simplest model among the four. One reason for this finding is that the three proposed models in the present simulation study employed a binary indicator based on the distance between ability and item difficulty, as opposed to the observed correct and incorrect responses in simulation study 1. This may add an extra layer of complexity into the proposed models, and DIC was known to prefer simpler models. While AIC, AICc and BIC required the specification of the number of parameters, the effective number of parameters was estimated by DIC (Spiegelhalter et al., 2002). Therefore, it is possible that DIC estimated more effective number of parameters in the JM-DD1, the JM-DD2, and the JM-D than the specified number of parameters when test length was small, thus penalizing these three models more. In summary, for shorter tests, deviance, AIC and AICc are recommended over BIC and DIC; when test length is long, deviance, AIC, AICc and DIC outperform BIC, which favors the underspecified model with random effects.

Table 25. *Frequency of identifying each model as the best fitting model in simulation study 2.*

				Deviance				AIC				AICc				BIC				DIC			
J	I	$\rho_{\theta\tau}$	$\rho_{\theta\tau}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	14	0	16	0	12	0	18	0	10	0	20	0	0	0	30	0	3	0	1	26
			.7	23	0	7	0	18	0	12	0	16	0	14	0	1	0	29	0	7	0	6	17
		.5	.3	13	0	17	0	6	0	24	0	4	0	26	0	0	0	30	0	6	0	3	21
			.7	18	0	12	0	9	0	21	0	9	0	21	0	0	0	30	0	6	3	4	17
		.8	.3	19	0	11	0	16	0	14	0	9	0	21	0	0	0	30	0	13	1	13	3
			.7	21	0	9	0	18	0	12	0	13	0	17	0	1	0	29	0	2	0	3	25
	40	.2	.3	20	0	10	0	14	0	16	0	9	0	21	0	0	0	30	0	15	0	15	0
			.7	22	0	8	0	16	0	14	0	11	0	19	0	0	0	30	0	15	1	14	0
		.5	.3	25	0	5	0	21	0	9	0	12	0	18	0	0	0	30	0	16	0	14	0
			.7	17	0	13	0	13	0	17	0	9	0	21	0	0	0	30	0	19	0	11	0
		.8	.3	17	0	13	0	10	0	20	0	6	0	24	0	0	0	30	0	20	0	10	0
			.7	27	0	3	0	18	0	12	0	14	0	16	0	0	0	30	0	18	0	12	0
1000	20	.2	.3	17	0	13	0	16	0	14	0	14	0	16	0	0	0	30	0	11	0	7	12
			.7	16	0	14	0	11	0	19	0	11	0	19	0	0	0	30	0	1	1	1	27
		.5	.3	14	0	16	0	11	0	19	0	11	0	19	0	0	0	30	0	7	0	7	16
			.7	16	0	14	0	13	0	17	0	13	0	17	0	0	0	30	0	10	1	6	13
		.8	.3	15	0	15	0	9	0	21	0	7	0	23	0	1	0	29	0	13	0	12	5
			.7	20	0	10	0	13	0	17	0	13	0	17	0	0	0	30	0	7	0	4	19
	40	.2	.3	15	0	15	0	10	0	20	0	8	0	22	0	0	0	30	0	15	0	15	0
			.7	26	0	4	0	24	0	6	0	20	0	10	0	2	0	28	0	11	0	19	0
		.5	.3	21	0	9	0	13	0	17	0	13	0	17	0	0	0	30	0	15	0	15	0
			.7	28	0	2	0	23	0	7	0	20	0	10	0	1	0	29	0	15	0	15	0
		.8	.3	17	0	13	0	10	0	20	0	8	0	22	0	0	0	30	0	14	0	16	0
			.7	14	0	16	0	8	0	22	0	7	0	23	0	0	0	30	0	21	0	9	0

In a nutshell, the conclusions from simulation study 2 were quite similar to simulation study 1 regarding the RT model-related parameters. The differences in the recovery of IRT model-related parameters were mainly attributed to the shift indicator based on item-person distance. Introducing the item-person distance into the RT model enabled the direct use of RT information for improving the estimation accuracy of IRT model-related parameters. As a result, the JM-DD1 performed better than the underspecified models on most parameters. While using the HM as the estimation model did not affect the IRT model-related parameters for data generated from the JM-RD1, ignoring the conditional dependence would result in larger estimation error in all parameters involved in the joint modeling of responses and RTs.

Other than the estimation model, the manipulated factors also influenced the parameter estimates in a similar way to simulation study 1, despite some changes in the magnitude of effect sizes (see Table 26). The interaction between the estimation model and the correlation between the two latent traits was a significant factor on the SE of the ability parameters in simulation study 1, but in simulation study 2 it affected the SE of the speed parameters significantly.

Lastly, the sensitivity of the model selection criteria was comparable to simulation study 1 except DIC. DIC tended to favor the HM over the proposed models when test length was small. Therefore, caution should be advised when selecting model fit indices to identify the best fitting model for different test lengths.

Table 26. *Summary of the effect sizes of the significant effects in simulation study 2*.

		Significant effects				
Parameters	Error Indices	Model	J	I	$\rho_{\theta\tau}$	Model*$\rho_{\theta\tau}$
b_i	Bias					
	SE	medium	medium			
	RMSE	small	medium			
α_i	Bias	large				
	SE	small	large			
	RMSE	medium	small			
β_i	Bias	large				
	SE	large	small			
	RMSE	large				
θ_j	SE	medium		medium	medium	
	RMSE	large		small		
τ_j	SE			large		small
	RMSE	large		large		

4.1.3 Simulation Study 3

Simulation study 1 compared the three models conditioning on observed item responses and the baseline HM, whereas simulation study 2 evaluated the three models conditioning on the item-person distance with the HM. The purpose of simulation study 3 is to compare the performance of the six proposed models and the HM when data were generated from two different mechanisms of test-taking behaviors. Since the proposed models with each shift indicator have been compared in simulation studies 1 and 2, this section focuses on exploring the consequences of fitting models with misspecified shift indicator.

Two conditions generated in simulation studies 1 and 2 were used in the present simulation study. Both conditions included responses and RTs from 500 examinees, assuming weak correlation between ability and speed and weak correlation between item difficulty and the shift. Test length was varied at 20 and 40. Thus, repeated measures ANOVA was conducted with one between-condition factor, test length, and one within-condition factor, the estimation model. The bias, SE, and RMSE under the two conditions were reported in Appendix C.

First-level item parameters. In terms of the item difficulty parameters, the estimation model was a significant factor with a medium effect size (*f*=.353) on RMSE when the data generating model was the JM-RD1. Interestingly, the seven estimation models performed similarly, except that the JM-RD2 (see Figure 29 and Table 27). The JM-RD2 appeared to have the largest RMSE among all seven estimation models, indicating that ignoring the random effects resulted in significantly larger total error in the item difficulty estimates than the misspecified models, even if the shift indicator was specified correctly.

When the JM-DD1 was the data generating model, the estimation model had a medium effect size (*f*=.387) on SE and a small effect size (*f*=.211) on RMSE. Except for the JM-DD2, models with correctly specified shift indicator significantly outperformed the models conditioning on item responses for both SE and RMSE. No significant difference was found among the three models conditioning observed item responses (i.e., the JM-RD1, the JM-RD2, and the JM-R) and the HM for both SE and RMSE. These results reflect that fitting models with misspecified shift indicator would not reduce the estimation precision of the item difficulty estimates when data were generated from the JM-RD1, but fitting the models conditioning on item responses to data generated from the JM-DD1 resulted in significantly worse estimates.

For both time discrimination and time intensity parameters, the error indices were only affected by the estimation model with a practically meaningful effect size.

In terms of the time discrimination parameters (see Figure 30 and Table 28), the estimation model had a big effect size (*f*=.518) on the systematic error, a small effect size (*f*=.104) on the random error, and a medium effect size (*f*=.381) on the total error when data were simulated from the JM-RD1. If the underlying data structure followed the JM-DD1, the estimation model was significant with a large effect size (*f*=.601) on bias, and medium effect sizes on SE and RMSE (*f*=.259 and .340, respectively).

Regarding the time intensity parameters, the estimation model was significant with large effect sizes on bias, SE, and RMSE (*f*=.453, .427, and .494, respectively) for data generated from the JM-RD1. When the JM-DD1 was the data generating model, the estimation model also had a significant large effect with a size (*f*=.435) on bias, a medium effect size (*f*=.391) on SE, and a large effect size (*f*=.440) on RMSE.

Findings on both time discrimination and time intensity parameters were similar. Applying estimation models with incorrectly specified shift indicator led to large negative bias and large RMSE, even though the seven models did not differ much on SE. For the time intensity parameters, the discrepancy between the HM and the models conditioning on item-person distance on bias and RMSE was larger when data were simulated from the JM-RD1 (see Figure 31 and Table 29). However, the HM only performed slightly worse than models with a shift mechanism different from the data generating model for data generated from the JM-DD1. This finding suggests that the underlying structure specified in the JM-DD1 is more sensitive to the specification of the shift indicator in terms of the recovery of time intensity parameters.

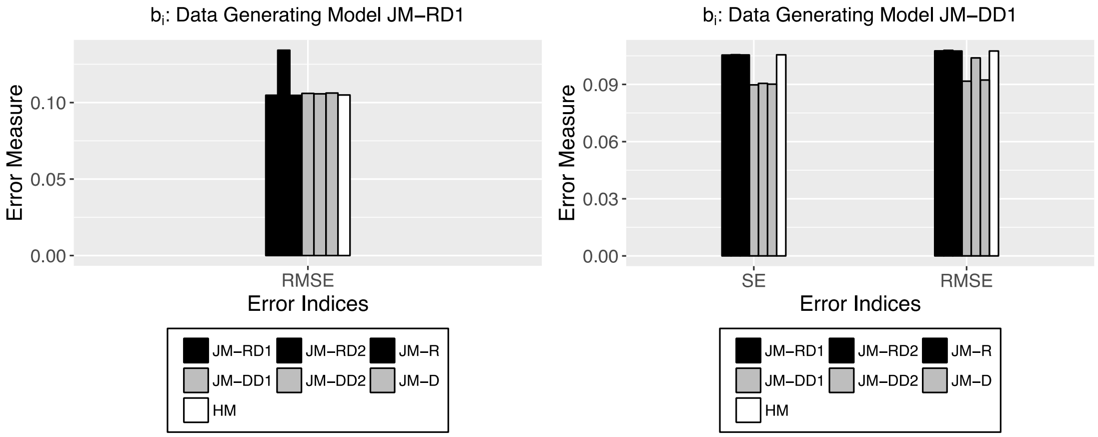


Figure 29. Significant main effects on the SE and RMSE of the item difficulty estimates.

Table 27. *Post-hoc pairwise comparison results of the estimation model on the SE and RMSE in item difficulty estimation in simulation study 3.*

		Mean Difference		
		JM-RD1	JM-DD1	
Model (m)	Model (n)	RMSE	SE	RMSE
JM-RD1	JM-RD2	-.030*	<.001	<.001
	JM-R	<.001	<.001	<.001
	JM-DD1	-.002*	.014*	.015*
	JM-DD2	-.001	.013*	.004
	JM-D	-.002	.014*	.014*
	HM	<.001	<.001	<.001
JM-RD2	JM-R	.030*	<.001	<.001
	JM-DD1	.028*	.015*	.015*
	JM-DD2	.029*	.013*	.005
	JM-D	.028*	.014*	.014*
	HM	.030*	<.001	<.001
JM-R	JM-DD1	-.002	.014*	.014*
	JM-DD2	-.001	.013*	.004
	JM-D	-.002*	.014*	.014*
	HM	<.001	<.001	<.001
JM-DD1	JM-DD2	.001	-.001	-.010*
	JM-D	<.001	-.001	-.001
	HM	.002	-.014*	-.014*
JM-DD2	JM-D	-.001	.001	.010
	HM	.001	-.013*	-.004
JM-D	HM	.002	-.014*	-.014*

Note. *p<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

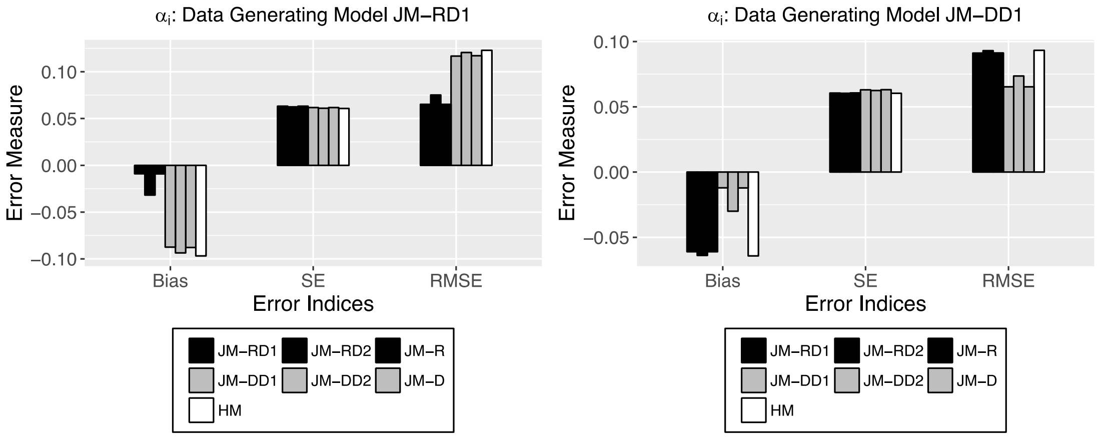


Figure 30. Significant main effects on the bias, SE, and RMSE of the time discrimination estimates.

Table 28. *Post-hoc pairwise comparison results of the estimation model on the bias, SE, and RMSE in time discrimination estimation in simulation study 3.*

		Mean Difference					
		JM-RD1			JM-DD1		
Model (m)	Model (n)	Bias	SE	RMSE	Bias	SE	RMSE
JM-RD1	JM-RD2	-.022*	.001	-.010*	-.003*	<.001	-.002*
	JM-R	<.001	<.001	<.001	<.001	<.001	<.001
	JM-DD1	-.077*	.001	-.050*	.047*	-.003*	.024*
	JM-DD2	-.083*	.002*	-.054*	.029*	-.002*	.016*
	JM-D	-.077*	.001	-.051*	.046*	-.003*	.024*
	HM	-.086*	.002*	-.057*	-.003*	<.001	-.002*
JM-RD2	JM-R	.022*	-.001	.010*	.002*	<.001	.002*
	JM-DD1	-.055*	.001	-.041*	.049*	-.003*	.026*
	JM-DD2	-.061*	.001	-.045*	.032*	-.002*	.018*
	JM-D	-.055*	.001	-.041*	.049*	-.003*	.026*
	HM	-.065*	.002	-.047*	<.001	<.001	<.001
JM-R	JM-DD1	-.077*	.001	-.050*	.047*	-.003*	.024*
	JM-DD2	-.083*	.002*	-.054*	.029*	-.002*	.016*
	JM-D	-.077*	.001	-.051*	.047*	-.003*	.024*
	HM	-.086*	.002*	-.057*	-.003*	<.001	-.002*
JM-DD1	JM-DD2	-.006*	.001*	-.004*	-.017*	.001	-.008*
	JM-D	<.001	<.001	<.001	<.001	<.001	<.001
	HM	-.010*	.001*	-.006*	-.049*	.003*	-.026*
JM-DD2	JM-D	.006*	-.001*	.003*	.017*	-.001	.008*
	HM	-.003*	<.001	-.003*	-.032*	.002*	-.018*
JM-D	HM	-.009*	.001*	-.006*	-.049*	.003*	-.026*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

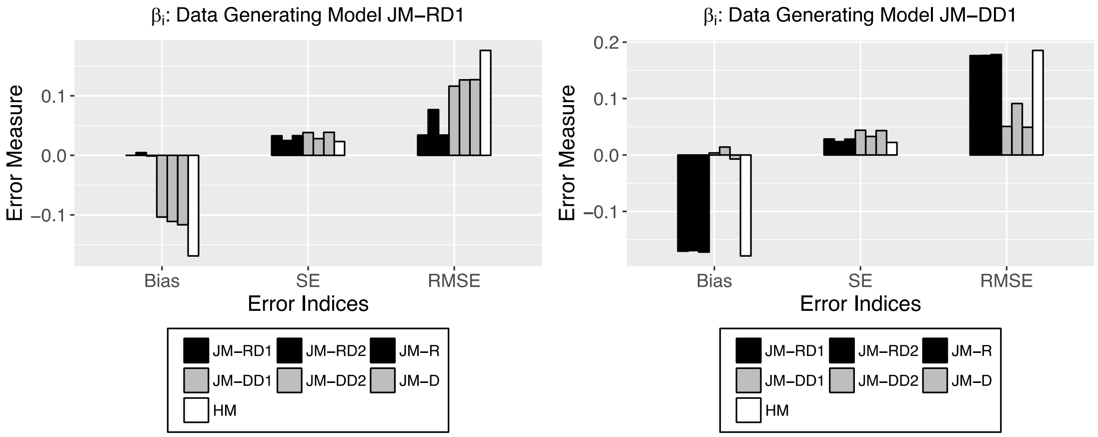


Figure 31. Significant main effects on the bias, SE, and RMSE of the time intensity estimates.

Table 29. *Post-hoc pairwise comparison results of the estimation model on the bias, SE, and RMSE in time intensity estimation in simulation study 3.*

		Mean Difference					
		JM-RD1			JM-DD1		
Model (m)	Model (n)	Bias	SE	RMSE	Bias	SE	RMSE
JM-RD1	JM-RD2	-.003	.008*	-.042*	.001	.004*	<.001
	JM-R	-.001	<.001*(-)	<.001	-.002	<.001	-.002
	JM-DD1	-.104*	-.006*	-.082*	.172*	-.016*	.126*
	JM-DD2	-.114*	.005*	-.096*	.168*	-.005*	.087*
	JM-D	-.117*	-.007*	-.093*	.161*	-.016*	.126*
	HM	-.170*	.010*	-.143*	-.009	.006*	-.009
JM-RD2	JM-R	.002	-.008*	.042*	-.003	-.004*	-.002
	JM-DD1	-.101*	-.014*	-.040*	.171*	-.021*	.127*
	JM-DD2	-.111*	-.003*	-.053*	.167*	-.010*	.087*
	JM-D	-.114*	-.015*	-.051*	.161*	-.020*	.127*
	HM	-.167*	.002*	-.100*	-.009*	.002*	-.009*
JM-R	JM-DD1	-.103*	-.006*	-.082*	.173*	-.016*	.128*
	JM-DD2	-.113*	.005*	-.095*	.169*	-.005*	.088*
	JM-D	-.116*	-.007*	-.093*	.163*	-.016*	.128*
	HM	-.169*	.010*	-.143*	-.007	.006*	-.007
JM-DD1	JM-DD2	-.010	.011*	-.014	-.004	.011*	-.040*
	JM-D	-.013	-.001	-.011*	-.010	<.001	<.001
	HM	-.066*	.016*	-.061*	-.180*	.022*	-.135*
JM-DD2	JM-D	-.003	-.012*	.003	-.006	-.011*	.040*
	HM	-.056*	.005*	-.047*	-.176*	.011*	-.096*
JM-D	HM	-.053*	.017*	-.050*	-.170*	.022*	-.135*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

First-level person parameters. Unlike item parameters, the first-level person parameters were affected by both the estimation model and test length. With regard to the ability parameters, for data generated from the JM-RD1, the estimation model and test length were significant with a small effect size (*f*=.184) and a medium effect size (*f*=.383) on the random error, and with a small effect size (*f*=.148) and a medium effect size (*f*=.266) on the total error. For data generated from the JM-DD1, both the estimation model and test length had significant impacts on the SE of ability estimates with large effect sizes (*f*=.574 and .503 respectively). The interaction term between the estimation model and test length also significantly affected the random error in the ability estimates with a small effect size (*f*=.132). Regarding RMSE, the estimation model and test length were significant with a large effect size (*f*=.716) and a medium effect size (*f*=.330) respectively.

If the latent structure followed the JM-RD1, models conditioning on item-person distance yielded significantly larger SE and RMSE than the models conditioning on item responses, but the discrepancies were rather small (see Figure 32 and Table 30). In contrast, if the shift indictor was determined by item-person distance, models with the correct shift indicator yielded much smaller SE and RMSE than other misspecified models. Models conditioning on the observed responses produced SE and RMSE comparable to the HM, indicating that it is more important to specify the correct shift indicator for data generated from the JM-DD1. Similar to the item difficulty parameters, this finding may also be because RTs directly take part in the estimation of ability parameters. Thus, the ability estimates from other models that utilize RT information through the correlation between ability and speed are expected to be less accurate.

The significant interaction effect on the SE of the ability estimates is depicted in Figure 33. The JM-RD1, the JM-RD2, the JM-R, and the HM performed similarly on the random error of the ability estimates when data were simulated from the JM-DD1. The three models conditioning on item-person distance, however, produced smaller random errors. This pattern was the same for both levels of test length, despite that the differences in the SEs between the three models conditioning on item-person distance and the other four models was more evident when test length was 40.

Regarding the speed parameters, when the shift indicator was determined by item responses, the estimation model and test length were significant factors with a medium effect size (*f*=.314) and a large effect size (*f*=.772) on the random error, and large effect sizes (*f*=.439 and .671 respectively) on the total error. Similarly, when the JM-DD1 was the data generating model, the estimation model had a medium effect size (*f*=.233) on SE and a large effect size (*f*=.450) on RMSE, whereas test length had a large effect size (*f*=.712) on SE and a medium effect size (*f*=.229) on RMSE.

For both data generating models, fitting models with correctly specified shift indicator resulted in significantly smaller RMSE, even though the differences in SE was small (see Figure 34 and Table 31). Similar to the time intensity parameters, applying models conditioning on the item responses to data generated from the JM-DD1 resulted in the speed estimates containing approximately the same amount of error as the HM. This finding aligns with the results from the first-level item parameters, that the JM-DD1 is more sensitive to misspecified shift indicator.

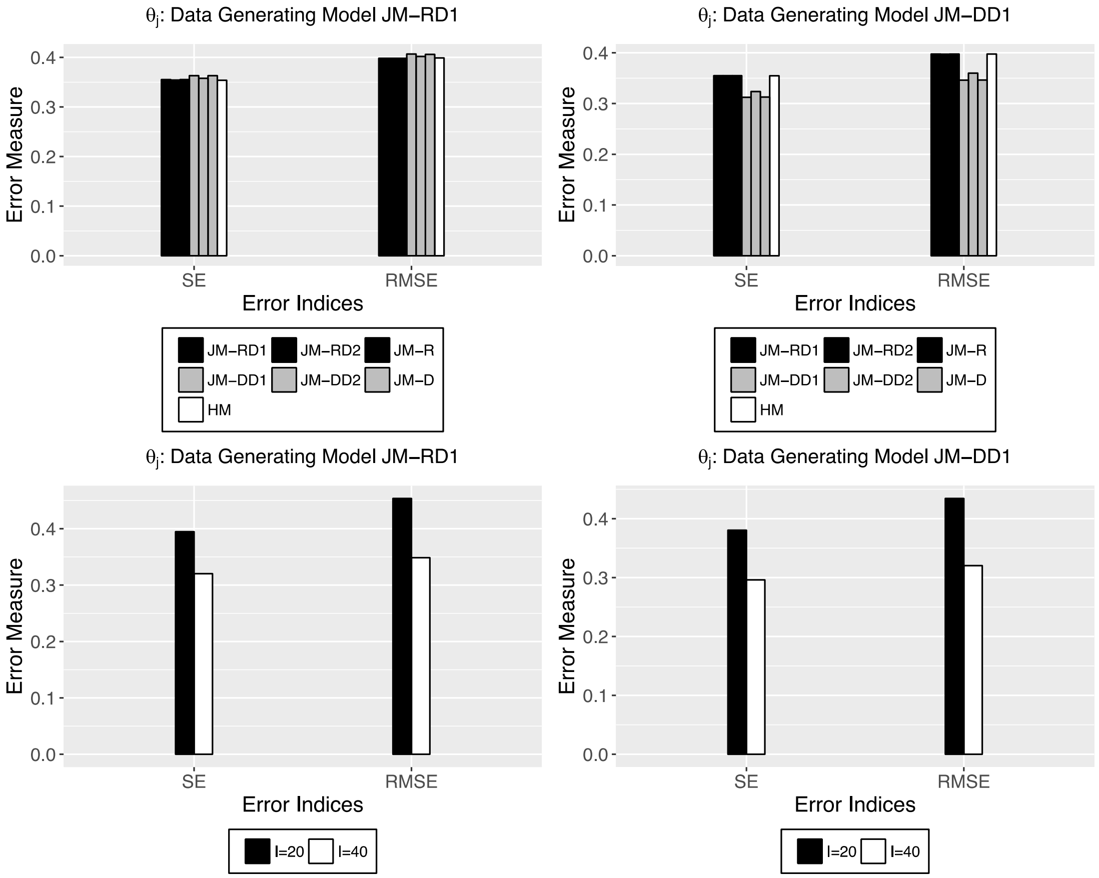


Figure 32. Significant main effects on the SE and RMSE of the ability estimates.

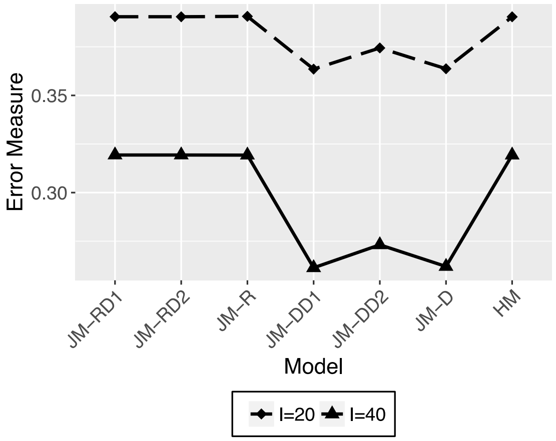


Figure 33. Significant two-way interaction between the estimation model and test length on the SE of the ability estimates for data generated from the JM-DD1.

Table 30. *Post-hoc pairwise comparison results of the estimation model on the SE and RMSE in ability estimation in simulation study 3.*

		Mean Difference			
		JM-RD1		JM-DD1	
Model (m)	Model (n)	SE	RMSE	SE	RMSE
JM-RD1	JM-RD2	.001*	<.001	<.001	<.001
	JM-R	<.001	<.001	<.001	<.001
	JM-DD1	-.008*	-.009*	.042*	.051*
	JM-DD2	-.003*	-.004*	.031*	.038*
	JM-D	-.008*	-.008*	.042*	.051*
	HM	.001*	-.001	<.001	<.001
JM-RD2	JM-R	-.001*	<.001	<.001	<.001
	JM-DD1	-.009*	-.009*	.042*	.051*
	JM-DD2	-.003*	-.004*	.031*	.038*
	JM-D	-.009*	-.008*	.042*	.051*
	HM	.001*	-.001	<.001	<.001
JM-R	JM-DD1	-.008*	-.008*	.043*	.051*
	JM-DD2	-.003*	-.004*	.031*	.038*
	JM-D	-.008*	-.008*	.042*	.051*
	HM	.001*	-.001	<.001	<.001
JM-DD1	JM-DD2	.005*	.005*	-.011*	-.014*
	JM-D	<.001	.001*	<.001	<.001
	HM	.009*	.008*	-.042*	-.051*
JM-DD2	JM-D	-.005*	-.004*	.011*	.014*
	HM	.004*	.003*	-.031*	-.038*
JM-D	HM	.009*	.007*	-.042*	-.051*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

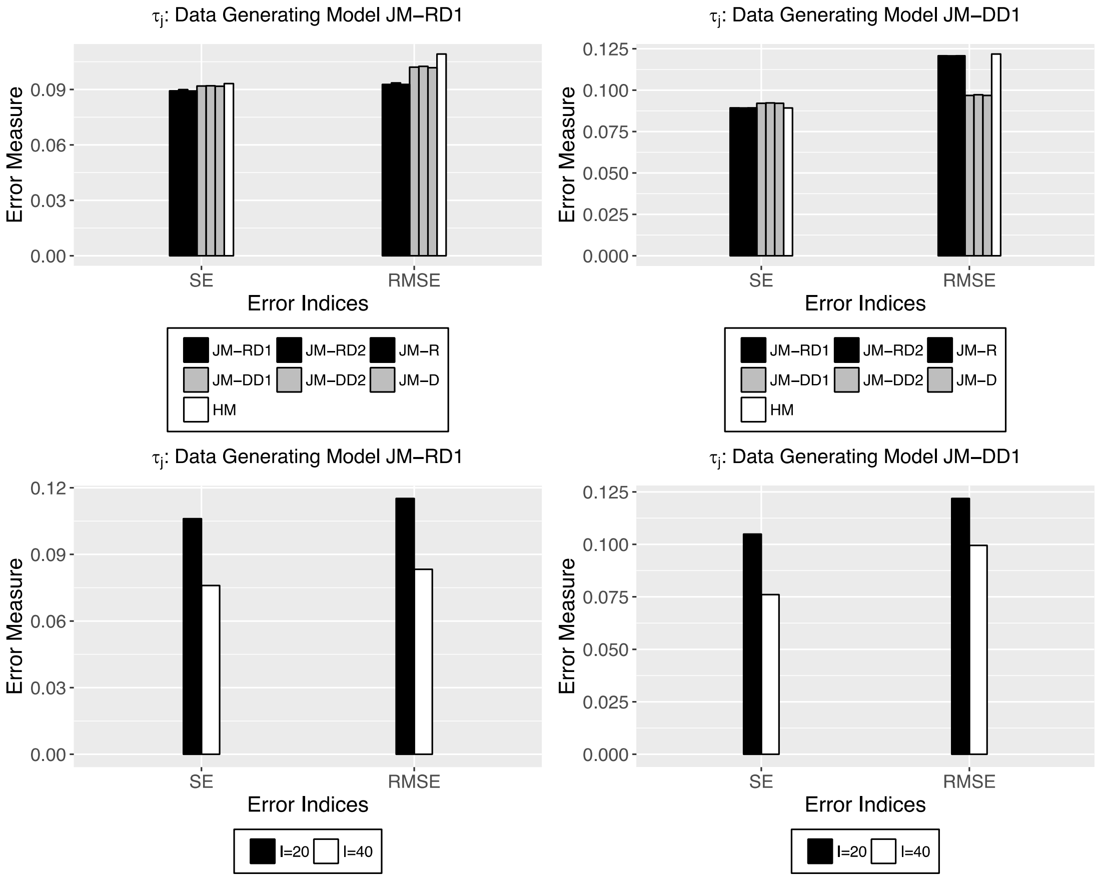


Figure 34. Significant main effects on the SE and RMSE of the speed estimates.

Table 31. *Post-hoc pairwise comparison results of the estimation model on the SE and RMSE in speed estimation in simulation study 3.*

		Mean Difference			
		JM-RD1		JM-DD1	
Model (m)	Model (n)	SE	RMSE	SE	RMSE
JM-RD1	JM-RD2	-.001*	-.001*	<.001	<.001*(+)
	JM-R	<.001	<.001	<.001	<.001
	JM-DD1	-.003*	-.009*	-.003*	.024*
	JM-DD2	-.003*	-.010*	-.003*	.024*
	JM-D	-.003*	-.009*	-.003*	.024*
	HM	-.004*	-.016*	<.001	-.001*
JM-RD2	JM-R	.001*	.001*	<.001	<.001*(-)
	JM-DD1	-.002*	-.009*	-.003*	.024*
	JM-DD2	-.002*	-.009*	-.003*	.023*
	JM-D	-.002*	-.008*	-.003*	.024*
	HM	-.003*	-.016*	<.001	-.001*
JM-R	JM-DD1	-.003*	-.009*	-.003*	.024*
	JM-DD2	-.003*	-.010*	-.003*	.024*
	JM-D	-.003*	-.009*	-.003*	.024*
	HM	-.004*	-.016*	<.001	-.001*
JM-DD1	JM-DD2	<.001*(-)	<.001*(-)	<.001*(-)	<.001*(-)
	JM-D	<.001*(+)	<.001*(+)	<.001	<.001
	HM	-.001*	-.007*	.003*	-.025*
JM-DD2	JM-D	<.001*(+)	.001*	<.001*(+)	<.001*(+)
	HM	-.001*	-.007*	.003*	-.025*
JM-D	HM	-.001*	-.007*	.003*	-.025*

Note. **p*<.05. The mean difference in the error indices is calculated by subtracting each error index of Model (n) from that of Model (m).

Parameters on RT distribution shift, mean vector and covariance matrices. Figures 35 to 38 show the impact of the estimation model on the error indices of the parameters on RT distribution shift, item mean vector, item covariance matrix, and person covariance matrix, respectively. In general, fitting data generated from a model with one shift indicator to models with another shift indicator resulted in much larger bias and RMSE, despite that the SE from the models conditioning on item-person distance was always larger.

One exception was the slope parameter on RT distribution shift. For both data generating models, the estimated slope from models with misspecified shift indicator always had smaller bias, SE, and RMSE (see Figure 35). However, if the slope estimates were converted to the correlation between item difficulty and the shift, the model with the correct shift indicator always yielded smaller bias, SE, and RMSE. Given that the slope parameter was not well recovered, this finding indicates that the estimation error in the slope estimates may be influenced by other estimates, such as the variance of random effects and the variance of item difficulty parameters.

As with the first-level parameters, models with incorrectly specified shift indicator did not substantially reduce the estimation accuracy of IRT model-related parameters when the JM-RD1 reflected the true latent structure. For the mean and variance of item difficulty, the models conditioning on item-person distance even provided slightly smaller bias and RMSE than the JM-RD1 and the JM-R. However, fitting models conditioning on item responses to data simulated from the JM-DD1 usually led to errors as large as the HM. As such, the latent structure of the data should be carefully examined prior to choosing the estimation models.

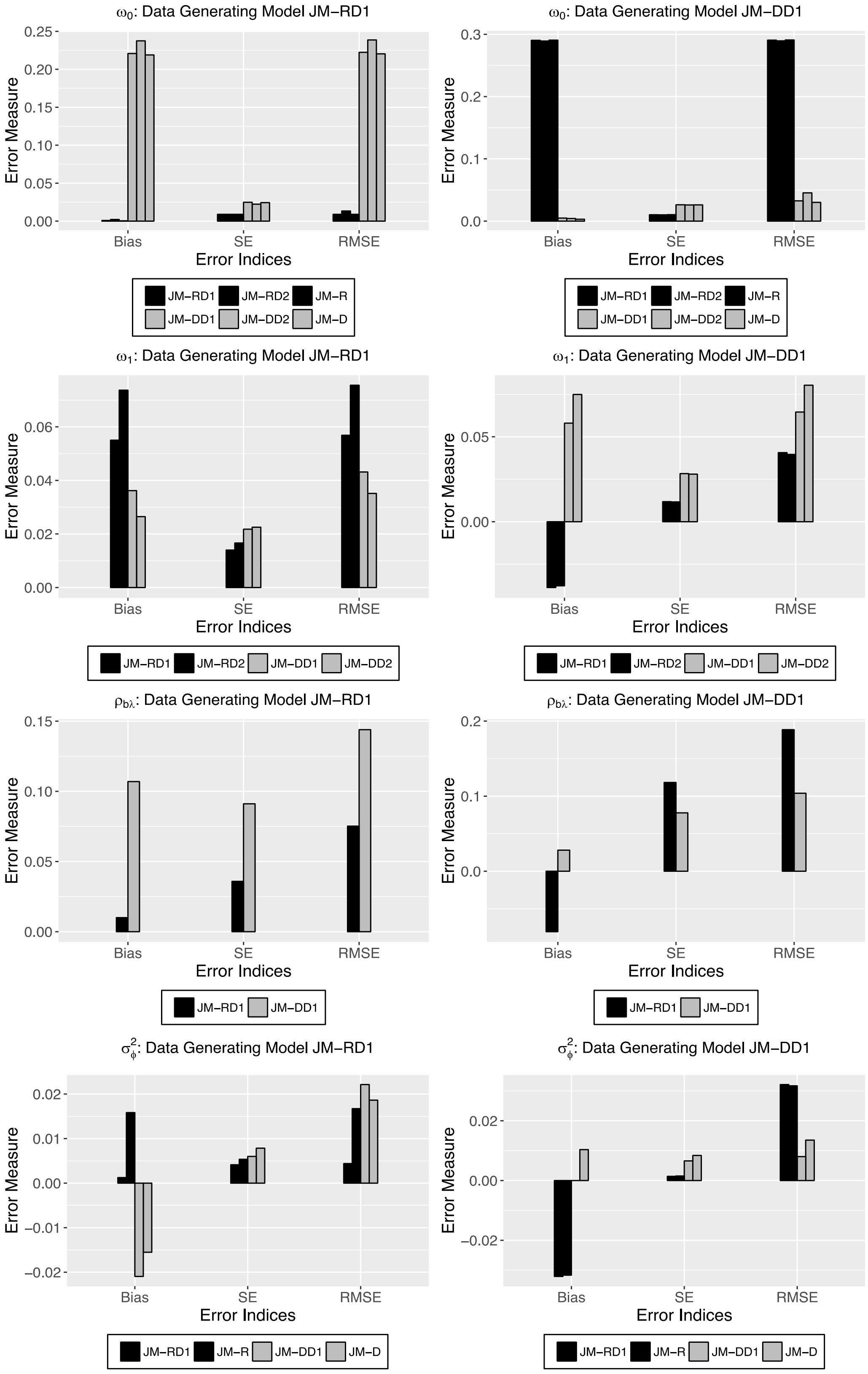


Figure 35. Main effect of the estimation model on the bias, SE, and RMSE of the parameter estimates on RT distribution shift.

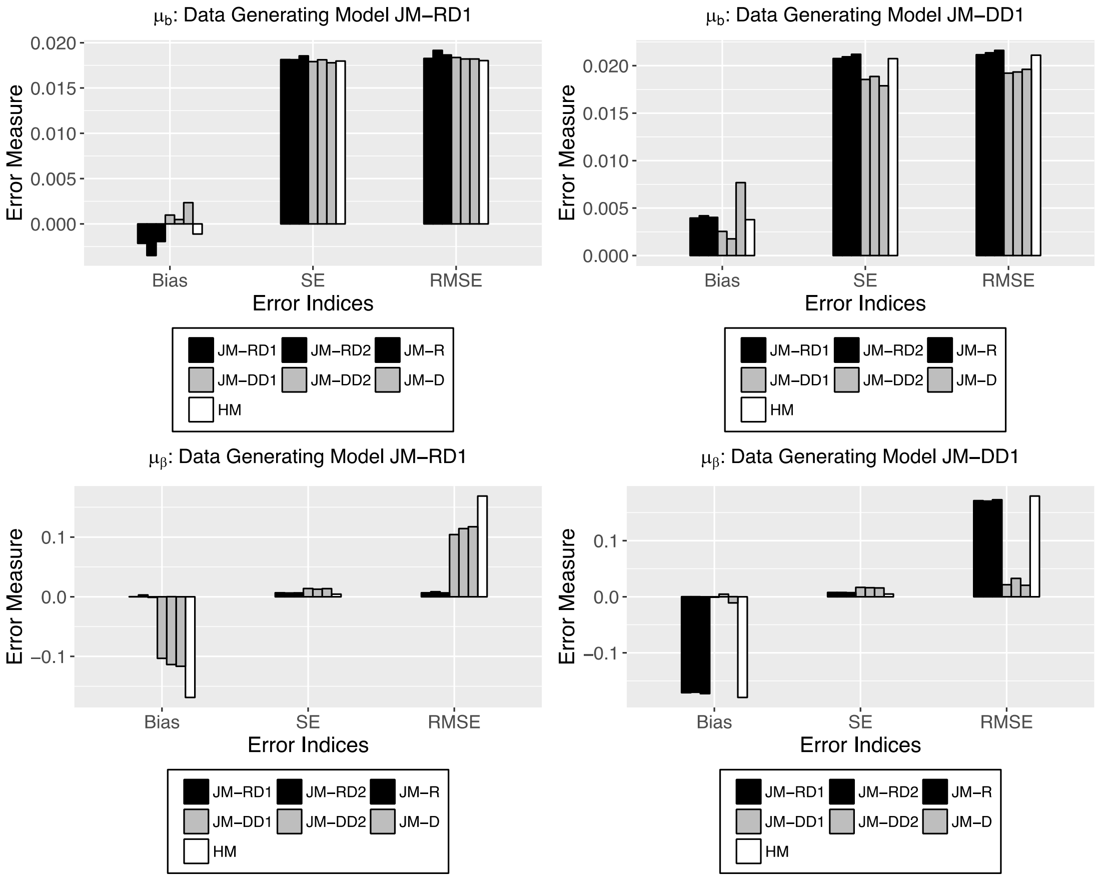
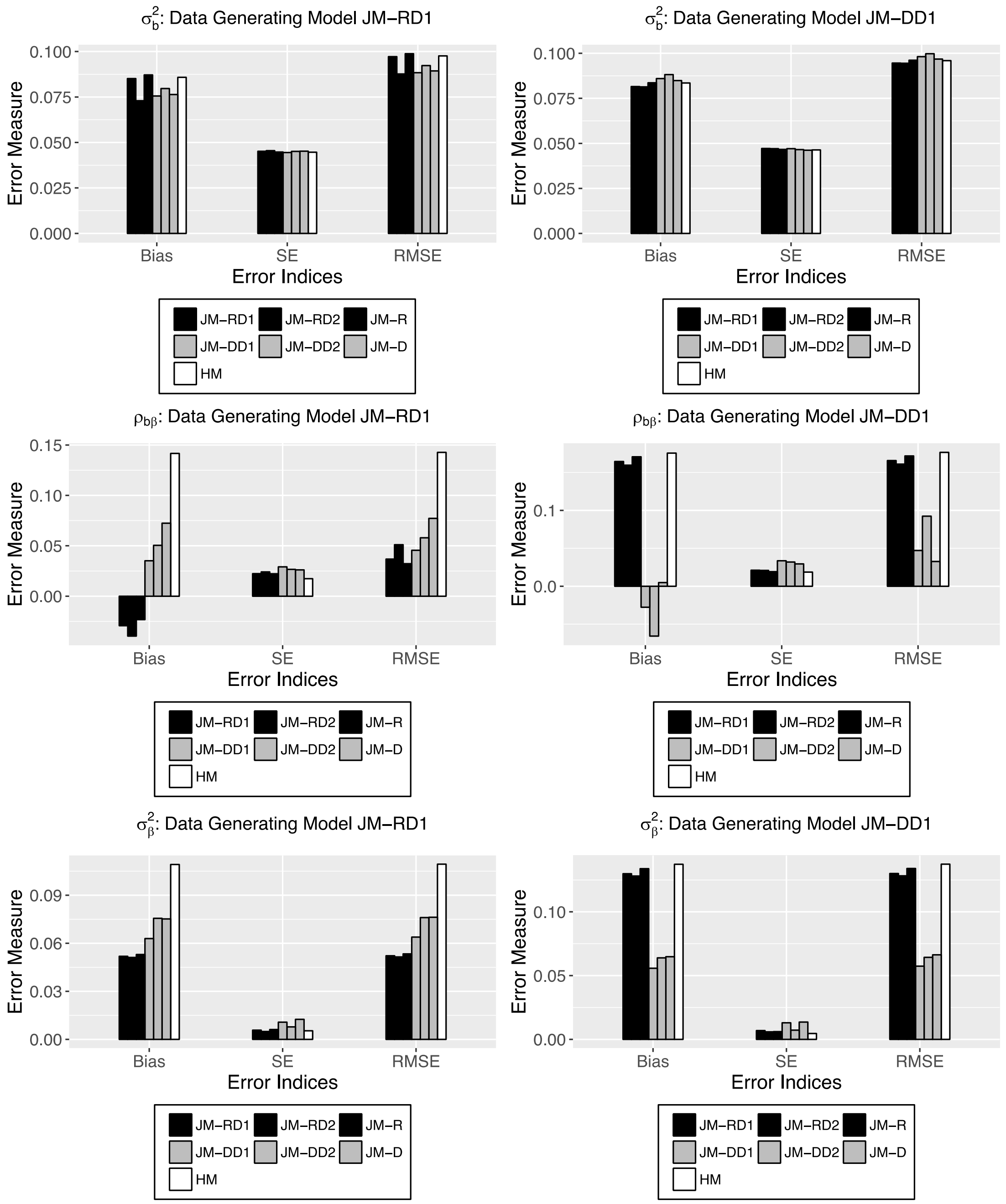


Figure 36. Main effect of the estimation model on the bias, SE, and RMSE of the item mean vector.



σ_b^2: Data Generating Model JM−RD1

σ_b^2: Data Generating Model JM−DD1

$\rho_{b\beta}$: Data Generating Model JM−RD1

$\rho_{b\beta}$: Data Generating Model JM−DD1

σ_β^2: Data Generating Model JM−RD1

σ_β^2: Data Generating Model JM−DD1

Figure 37. Main effect of the estimation model on the bias, SE, and RMSE of the item covariance matrix.

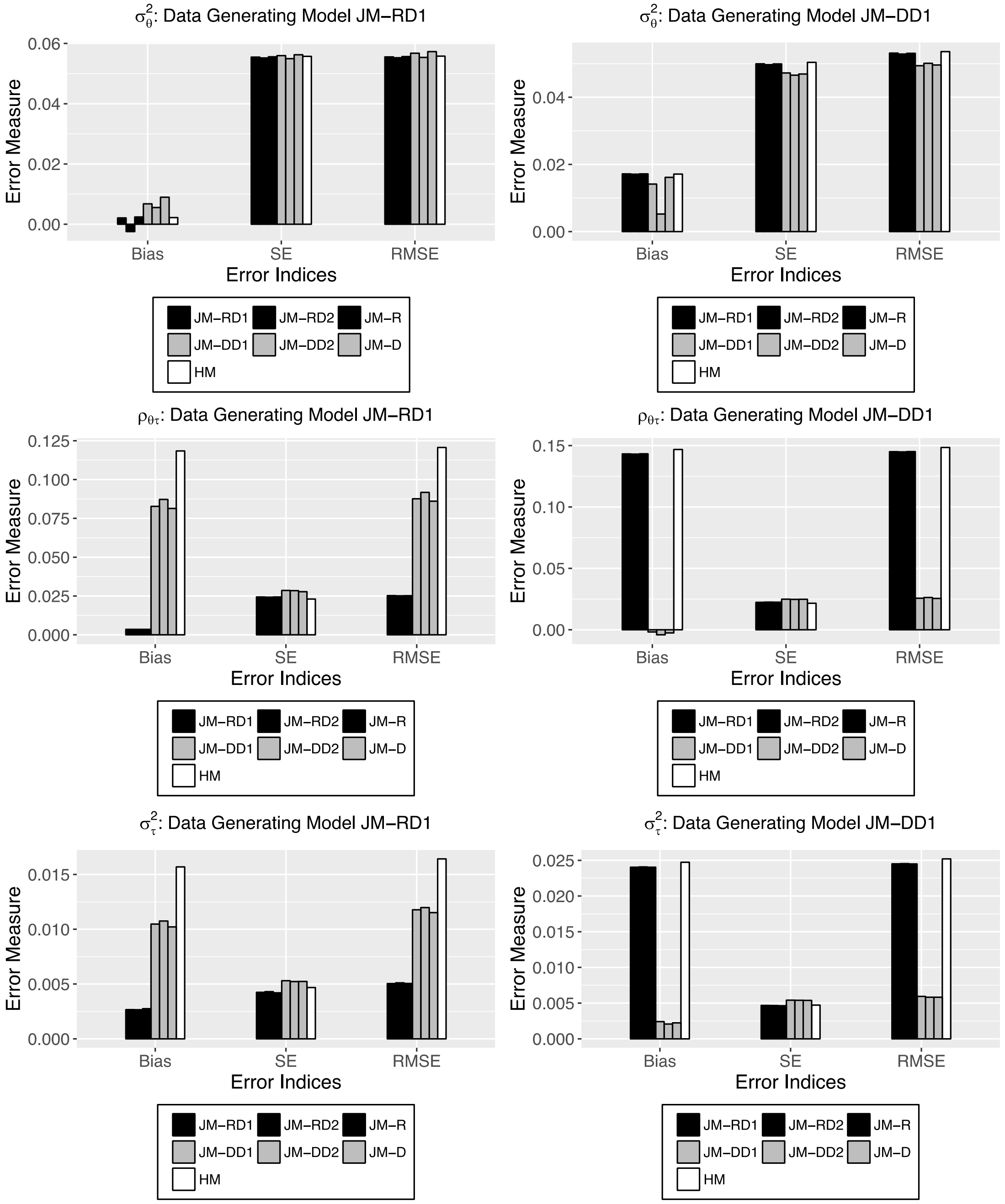


Figure 38. Main effect of the estimation model on the bias, SE, and RMSE of the person covariance matrix.

Model fit indices. Table 32 presents the frequency of each model fit index selecting each estimation model as the best fitting model. If the underlying structure followed the JM-RD1, the performance of the model selection criteria was similar to simulation study 1. Deviance and DIC chose the JM-RD1 and the JM-R about 50% of the replications, which outperformed the other model fit indices. When the JM-DD1 was the data generating model, deviance, AIC, and AICc performed better than their counterparts on data generated from the JM-RD1 under both conditions. BIC still selected the JM-D in all replications.

DIC was the second best model fit index when test length was 40 as it selected the JM-DD1 in about 50% of the replications. However, for a test with 20 items, DIC tended to favor models conditioning on observed responses more than the other four estimation models. Among the three models conditioning on item-person distance, only the JM-D was chosen as the best fitting model in one of the 30 replications, whereas the HM was selected in only three of the replications. On the contrary, the JM-R was the most frequently selected model, despite that the JM-DD1 was in fact the data generating model. Given that DIC tended to favor the HM in simulation study 2 when test length was small, this finding is not unexpected. Similar to the HM, the effective number of parameters for the models with item responses as the shift indicator estimated by DIC was also smaller than the JM-DD1. As such, DIC was not able to identify the true model when test length was small. Results from the other model fit indices may be more trustworthy for shorter tests.

Table 32. *Frequency of identifying each model as the best fitting model in simulation study 3.*

					JM-RD1							JM-DD1						
J	I	$\rho_{\theta\tau}$	$\rho_{\theta\tau}$	Model Fit Indices	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	Deviance	15	0	15	0	0	0	0	0	0	0	14	0	16	0
				AIC	3	0	27	0	0	0	0	0	0	0	12	0	18	0
				AICc	0	0	30	0	0	0	0	0	0	0	10	0	20	0
				BIC	0	0	30	0	0	0	0	0	0	0	0	0	30	0
				DIC	12	0	18	0	0	0	0	8	5	13	0	0	1	3
500	40	.2	.3	Deviance	14	0	16	0	0	0	0	0	0	0	20	0	10	0
				AIC	5	0	25	0	0	0	0	0	0	0	14	0	16	0
				AICc	0	0	30	0	0	0	0	0	0	0	9	0	21	0
				BIC	0	0	30	0	0	0	0	0	0	0	0	0	30	0
				DIC	19	0	11	0	0	0	0	0	0	0	15	0	15	0

To summarize, most parameters were affected by the choice of the estimation model, while only first-level person parameters were significantly influenced by test length (see Table 33). Applying models with the same shift indicator as the data generating model in general resulted in more accurate model parameter estimates, regardless of the data generating model. The consequences of employing estimation models with misspecified shift indicator were manifested on RT model-related parameters. If the true latent structure followed the JM-RD1, fitting models conditioning on item-person distance was not as harmful as the HM. However, if the JM-DD1 reflected the true model structure, models conditioning on item responses produced similar estimation error as the HM. As such, the underlying model structure should be identified before interpreting the parameter estimates. Even though the model fit indices considered in this study did not perform well in recognizing the data generating models, they were able to identify models with the correctly specified shift indicator, especially for longer tests.

Table 33. *Summary of the effect sizes of the significant effects in simulation study 3.*

		Significant effects				
		JM-RD1		JM-DD1		
Parameters	Error Indices	Model	I	Model	I	Model*I
b_i	Bias					
	SE			medium		
	RMSE	medium		small		
α_i	Bias	large		large		
	SE	small		medium		
	RMSE	medium		medium		
β_i	Bias	large		large		
	SE	large		medium		
	RMSE	large		large		
θ_j	SE	small	medium	large	large	small
	RMSE	small	medium	large	medium	
τ_j	SE	medium	large	small	large	
	RMSE	large	large	large	small	

4.2 Empirical Data Analyses

In addition to the simulation studies, the performance of different approaches to modeling or ignoring the speed-accuracy-difficulty interaction was evaluated with empirical data analyses as well. Datasets from two large-scale assessment programs that both used the Rasch model as the operational scoring model were utilized to demonstrate the application of the proposed models. The first dataset came from a large-scale credentialing exam program (Cizek & Wollack, 2017), and the second one was extracted from the math domain in 2012 PISA (OECD, 2014). After carrying out the data cleaning procedures described in section 3.4, the first dataset included item responses and RTs from 1,644 examinees and 40 items, whereas the second dataset contained item responses and RTs from 795 examinees and 10 items. The six proposed models and the baseline HM (van der Linden, 2007) were applied to both datasets. Model fit indices and parameter estimates based on the best fitting models for the two datasets were discussed in the sections 4.2.1 and 4.2.2 respectively.

4.2.1 Dataset 1

The model fit indices and parameter estimates for dataset 1 were reported in Tables 34 and 35. Convergence criteria were satisfied as all parameter estimates had a $\hat{R}$ smaller than 1.1, and no convergence issue was found through the examination of diagnostic plots.

Model fit indices summarized in Table 34 provided information from at least three aspects. First, the HM consistently yielded the largest values on all model fit indices, indicating that models accounting for the conditional dependence between responses and RTs provided much better overall model fit, regardless of the shift mechanisms. Second, for both shift mechanisms, the model that took into account conditional dependence but ignored the random effects usually produced larger model fit indices than the models with random effects. In particular, the JM-RD2 performed worse than the JM-RD1 and the JM-R, whereas the JM-DD2 performed worse than the JM-DD1 and the JM-D. Even so, all models with conditional dependence had better overall model fit than the HM.

Third, the three models conditioning on the observed responses yielded smaller model fit indices than the three models conditioning on item-person distance, implying that the underlying shift may more likely be dependent on the observed responses. Given that the model fit indices evaluated in the present study did not distinguish well between the two models with random effects in the simulation studies, both the JM-RD1 and the JM-R might be the better fitting underlying model. Therefore, the posterior mean and SD of the key parameters from both the JM-RD1 and the JM-R were presented in Table 35 and discussed.

For dataset 1, the estimated mean of item difficulty was -.912, indicating that items were in general easy for the examinee population. The correlation between item difficulty and time intensity was estimated to be .100 and .114 from the JM-RD1 and the JM-R respectively, reflecting a weak association between item difficulty and time intensity. Additionally, the examinees in dataset 1 were rather homogeneous in terms of ability and test-taking speed as the variances of ability and speed estimates were small. The correlation between ability and speed was also weak. The mean of time discrimination parameters was approximately 2, which is similar to the simulation setting of the present study.

In terms of the intercept of RT distribution shift, the estimate from the JM-RD1 was -.128, indicating that for an item with difficulty of 0, a correct response was .128 unit faster than an incorrect response. The interpretation of the estimated ω_0 from the JM-R was different from the JM-RD1. An estimated intercept of -.255 in the JM-R reflected that on average correct responses were .255 unit faster than incorrect responses for all items in dataset 1. Given that the item difficulty was positively related to the shift, this indicates that most items in dataset 1 had an item difficulty smaller than 0, which also aligned with the estimated mean of item difficulty. The estimate of ω_1 was .139, meaning that one unit increase in item difficulty led to .139 unit increase in the RT shift. This is consistent with the motivating example of this study (see Figures 1 and 2) as item difficulty was positively associated with RT distribution shift.

Taking the estimates of ω_0, σ_ϕ^2, and σ_b^2, the correlation between item difficulty and shift and the variance of the shift could also be derived from the parameter estimates. The correlation between item difficulty and shift was .727, and the variance of the shift was .028. Notice that the variance of the shift resolved from the estimates obtained from the JM-RD1 was equal to the estimated variance of the random effects from the JM-R. This is because in the JM-R where no predictor of RT distribution shift was included, the variance of random effects was theoretically equal to the variance of shift. Although the variance of shift may seem small, it was approximately the same as the estimated variance of speed parameters, and about one fourth of the estimated variance of time intensity. Adding item difficulty as a predictor of the shift explained half of the total variance of the shift.

Table 34. *Model fit indices for dataset 1.*

Model	Deviance	AIC	AICc	BIC	DIC
JM-RD1	**163441.2**	163703.2	163743.1	165251.1	**166975.6**
JM-RD2	164038.0	164298.0	164337.2	165834.1	167617.4
JM-R	163443.0	**163703.0**	**163742.2**	**165239.0**	167003.8
JM-DD1	166320.5	166582.5	166622.3	168130.3	170872.3
JM-DD2	166717.4	166977.4	167016.6	168513.4	170868.5
JM-D	166325.0	166585.0	166624.2	168121.0	170695.9
HM	167074.1	167330.1	167368.0	168842.5	170518.5

Table 35. *Parameter estimates for the dataset 1.*

	JM-RD1		JM-R	
Parameters	Mean	SD	Mean	SD
Model parameters				
μ_b	-.912	.138	-.909	.138
μ_β	4.166	.051	4.165	.052
σ_b^2	.754	.179	.757	.180
$\sigma_{b\beta}$	.028	.046	.032	.047
σ_β^2	.104	.025	.104	.025
σ_θ^2	.272	.015	.272	.015
$\sigma_{\theta\tau}$	.015	.003	.015	.003
σ_τ^2	.029	.001	.029	.001
ω_0	-.128	.021	-.255	.005
ω_1	.139	.023	NA	NA
σ_ϕ^2	.013	.004	.028	.007
Derived parameters				
mean(α_i)	2.067		2.067	
$\rho_{b\beta}$	.100		.114	
$\rho_{\theta\tau}$	.170		.169	
$\rho_{b\lambda}$	.727		NA	
σ_λ^2	.028		.028	

4.2.2 Dataset 2

Dataset 2 told a different story than dataset 1. Tables 36 and 37 summarize the model fit indices and posterior mean and SD for dataset 2. Convergence of all parameter estimates was checked through numerical and graphical diagnostics before interpreting the results.

Similar to dataset 1, the HM still yielded the largest model fit indices based on deviance, AIC, AICc and BIC. Yet, the HM produced smaller DIC than the JM-DD2 and the JM-D. This is not unexpected because both simulation studies 2 and 3 show that DIC tended to over-penalize models conditioning on item-person distance when data were generated from the JM-DD1 and test length was small. Yet, models with random effects still performed better than models with the same shift mechanism but without random effects, regardless of which model selection criteria was used.

Unlike dataset 1, models assuming one shift mechanism were not consistently better than models assuming the other shift mechanism. Nonetheless, the JM-DD1 and the JM-D were identified as the best fitting model by deviance, AIC, AICc, and BIC, indicating that the sample in dataset 2 was more likely to follow the second shift mechanism based on item-person distance. DIC was not considered as a model selection criteria for dataset 2 as it was not able to identify the data generating model when true underlying model followed the JM-DD1 and test length was small. Therefore, parameter estimates from the JM-DD1 and the JM-D were illustrated.

Regarding the item mean vector and covariance matrices, the JM-DD1 and the JM-D provided similar parameter estimates with differences in the third decimal place. Specifically, the estimated mean of item difficulty was .517 and .518 from the

JM-DD1 and the JM-D respectively. As such, items in dataset 2 were on average more difficult than items in dataset 1, even though they had similar time intensity and time discrimination parameter estimates as items in dataset 1. The estimated correlation between item difficulty and time intensity parameters was both .440, indicating a moderate positive association between item difficulty and time intensity.

The variance of time intensity was estimated to be larger than dataset 1. This shows that items in dataset 2 were more diverse regarding the time intensiveness. Along the same lines, examinees in dataset 2 had greater variabilities in terms of their ability and speed. Further, the latent ability and speed was strongly negatively correlated. In other words, examinees with higher ability tended to respond to the items in dataset 2 at a slower pace, whereas low-ability examinees were likely to spend less time on the items. As PISA was generally considered a low-stakes assessment, examinees seemed less motivated to work on the items.

Considering the intercept estimate from the JM-D, it is expected that the mean of item difficulty in dataset 2 should be above 0 given the positive correlation between item difficulty and the shift. While the correlation between item difficulty and the shift in dataset 1 was strong, the correlation for dataset 2 was weak. This may be due to unstable estimation with 10 items. Moreover, this parameter was on average about 40% underestimated in both simulation studies 1 and 2. As a result, the true correlation between item difficulty and the shift is expected to be stronger.

Nevertheless, the estimated variance of random effects was close to the estimated variance of speed parameters for both models. Consistent with the results from simulation studies and dataset 1, this finding suggests that the random effects are non-negligible given the magnitude of the variance. Except for the person covariance matrix, the SDs of all parameter estimates were quite large compared to dataset 1, indicating that model estimation was not as stable due to limited number of items.

Table 36. *Model fit indices for dataset 2.*

Model	Deviance	AIC	AICc	BIC	DIC
JM-RD1	20618.8	20700.8	20705.4	21166.4	**22420.6**
JM-RD2	20725.3	20805.3	20809.7	21259.6	22498.7
JM-R	20622.1	20702.1	20706.5	21156.4	22435.5
JM-DD1	**20500.4**	**20582.4**	**20587.0**	21048.0	22599.6
JM-DD2	20663.6	20743.6	20748.0	21197.9	23099.5
JM-D	20507.3	20587.3	20591.7	**21041.6**	22750.6
HM	20828.5	20904.5	20908.4	21336.1	22612.0

Table 37. *Parameter estimates for the dataset 2.*

	JM-DD1		JM-D	
Parameters	Mean	SD	Mean	SD
Model parameters				
μ_b	.517	.338	.518	.340
μ_β	4.115	.178	4.112	.180
σ_b^2	1.222	.692	1.216	.685
$\sigma_{b\beta}$	.278	.267	.278	.272
σ_β^2	.327	.190	.329	.195
σ_θ^2	.961	.085	.950	.086
$\sigma_{\theta\tau}$	-.306	.026	-.304	.026
σ_τ^2	.208	.013	.207	.013
ω_0	-.052	.062	-.009	.047
ω_1	.102	.090	NA	NA
σ_ϕ^2	.257	.179	.204	.133
Derived parameters				
mean(α_i)	1.959		1.958	
$\rho_{b\beta}$	.440		.440	
$\rho_{\theta\tau}$	-.684		-.686	
$\rho_{b\lambda}$	.217		NA	
σ_λ^2	.270		.204	

Chapter 5: Discussion

The conditional independence assumption between responses and RTs has been widely adopted in the joint modeling framework, yet the consequences of violating this assumption has not been thoroughly explored until recently (e.g., Bolsinova & Tijmstra, 2017; Meng et al., 2015; Ranger & Ortner, 2012). The present study focused on a phenomenon where the direction of the conditional dependence between responses and RTs appeared to have a systematic association with item difficulty. Different approaches to modeling the interaction among speed, accuracy, and difficulty were proposed. Their performance was evaluated with the HM that did not account for conditional dependence in simulation studies and empirical data analyses. In the first two sections of this chapter, findings from the three simulation studies and the empirical data analyses were summarized, and the implications of the findings in research and practical settings were addressed. In the last section of this chapter, limitations and future research directions were discussed in detail.

5.1 Discussion of the Simulation Results

In this section, findings from the three simulation studies were discussed in terms of the recovery of model parameters, impact of manipulated factors, performance of the proposed and alternative modeling approaches as well as model selection criteria. Generally speaking, all parameters were well recovered except the slope of the shift in RT distributions and the correlation between item difficulty and the shift. Repeated measures ANOVA was conducted for item difficulty, time discrimination, time intensity, ability, and speed parameter estimates, where only statistically significant and practically meaningful effects were reported. Other parameters were summarized in terms of bias, SE, and RMSE for each parameter under each condition.

5.1.1 Impact of the Manipulated Factors

Regarding item difficulty, time discrimination, time intensity, ability, and speed parameter estimates, conclusions were rather consistent for both simulation studies 1 and 2. Sample size was often significant with at least small effect sizes for item difficulty, time discrimination, and time intensity parameter estimates. On the other hand, test length was significant for person-related parameter estimates, ability and speed. Although test length was the only manipulated between-condition factor in simulation study 3, the same pattern was found where test length was a significant and practically important factor for both ability and speed estimates. Other than the five parameters, increasing sample size and test length were also in general associated with more accurate parameter estimates.

The conclusion that increasing sample size and test length both led to smaller errors in the parameter estimates was consistent with the literature (e.g., Kang, 2016; Marianti, 2015; Suh, 2010; Wang, Fan, et al., 2013). However, the finding that sample size only affected item-related parameters and test length only influenced person-related parameters was not commonly reported in this line of research. Several reasons may lead to this inconsistency. One possibility was that only a few studies manipulated sample size and/or test length in the context of modeling conditional dependence (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2017; Fox & Marianti, 2016; Meng et al., 2015; Ranger & Kuhn, 2012).

Among them, even fewer conducted ANOVA for the error indices and screened the significant factors with an effect size measure. It was also possible that the findings were different due to different data generating models.

Another manipulated factor, the correlation between ability and speed, only had a significant impact on the random error of the ability estimates. The SE of the ability parameters consistently increased as the correlation between ability and speed became stronger, which aligned with the findings from Klein Entink (2009) and Patton (2015). As more information was shared between the two latent traits, the ability estimates were more stable, regardless of the estimation model. On the other hand, the speed estimates were not significantly affected by this factor due to asymmetrical share of information. Other than the ability parameters, the correlation between ability and speed overall did not noticeably affect the parameter estimates, but it did lead to consistently decreasing bias from the HM and decreasing random and total errors from all estimation models of the estimated correlation between ability and speed. This finding implies that the estimation precision of the correlation between ability and speed was improved when the true person correlation was stronger. On the contrary, the estimated speed variance from the HM was more biased when the correlation between ability and speed increased.

The correlation between item difficulty and the shift, however, did not appear as a significant factor for any of the ANOVAs. This may result from the fact that the correlation between item difficulty and the shift operated on a rather small scale, thus did not lead to remarkable changes in the parameter estimates overall. However, the bias of estimates of the mean of time intensity, the correlation between item difficulty and time intensity, and the speed variance from the HM tended to increase as the correlation between item difficulty and the shift became stronger.

5.1.2 Different Approaches to Modeling the Speed-Accuracy-Difficulty Interaction

In the present study, six models were proposed with two different shift mechanisms and three different approaches to modeling the speed-accuracy-difficulty interaction. Three simulation studies were designed to evaluate the performance of the six proposed models as opposed to the baseline HM (van der Linden, 2007). In general, the proposed models yielded smaller bias and RMSE in the parameter estimates than the HM. The two models with random effects (i.e., the JM-RD1 and the JM-D, the JM-DD1 and the JM-D) tended to perform in a similar fashion for both shift mechanisms, which outperformed the models without random effects on bias and RMSE. Regarding random errors, the rank ordering was the opposite in most cases. For both simulation studies 1 and 2, the HM usually produced the smallest random errors, the model without random effects had the second smallest SEs, and the two models with random effects yielded the largest random errors. This is expected since when the complexity of the estimation model increases, the systematic error usually goes down, but the random error would be magnified. However, as the discrepancies among random errors from different estimation models was small compared to systematic errors, it was often found that the pattern of bias dominated the pattern of RMSE.

The consequences of ignoring the conditional dependence could be concluded through comparing parameter recovery under the HM and the proposed models. If the JM-RD1 reflected the true latent structure, ignoring the conditional dependence would not lead to significantly worse parameter estimates related to the IRT model, including item difficulty, ability, the mean and variance of item difficulty, and the ability variance. However, when data were simulated from the JM-DD1, estimates of the five IRT model-related parameters mentioned above from the HM were subject to more estimation errors. As the item-person distance was introduced into the RT model, the information from RT could contribute to improving the estimation accuracy of IRT model-related parameters both directly and indirectly. Further, the JM-DD1 performed significantly better than the JM-D on item difficulty and ability parameters, as opposed to rather equivalent performance of the JM-RD1 and the JM-R. As such, model parameter recovery of the JM-DD1 relied on correct model specification more than the JM-RD1.

Meanwhile, for both data generating model, parameter estimation related to the RT model would be negatively affected if conditional independence was assumed. Parameters such as time discrimination, time intensity, speed, the mean and variance of time intensity, the correlation between item difficulty and time intensity, the correlation between ability and speed, and the speed variance, would contain large systematic and total errors. In particular, time discrimination, time intensity, the mean of time intensity were underestimated, whereas the correlation between item difficulty and time intensity, the variance of time intensity, the correlation between two latent traits, and the variance of speed were overestimated to a larger extent compared to other estimation models. Even though the bias of the speed estimates was fixed at zero, the SDs of bias in the speed estimates were inflated, which led to the largest total errors in the speed estimates from the HM as well.

A closer examination of the SDs of bias in the ability estimates reflected that the proposed models produced smaller SDs of bias than the HM for both data generating models, despite that the mean bias of ability estimates was also constrained to zero. Further, the stronger the correlation between the two latent traits, the larger the discrepancy between the SD of bias in the ability estimates from the proposed models and the HM. Meng et al. (2015) had a similar conclusion that modeling the conditional dependence yielded less biased estimates, but the improvements were not noticeable when the correlation between latent traits was low. Bolsinova and Tijmstra (2017) also showed that accounting for the conditional dependence between responses and RTs resulted in a decrease of MSE in ability estimates only when the correlation between latent traits was non-zero.

In simulation study 3, the consequences of applying estimation models with misspecified shift indicator were also explored thoroughly. Overall, fitting data generated from one shift mechanism with models assuming the same shift mechanism yielded smaller bias and RMSE than models assuming the misspecified shift indicator. Additionally, the impact of employing models with misspecified shift indicator was more remarkable on the RT model-related parameters, rather than the IRT model-related parameters. One important difference between the two data generating models was that when the shift indicator was determined by item-person distance, estimates from the models conditioning on item responses were as biased as the HM. Therefore, accounting for the conditional dependence between responses and RTs does not guarantee the improvement of the parameter estimation accuracy, especially when the shift indicator is misspecified.

5.1.3 Model Selection Criteria

Sensitivity of several information-based relative model fit indices was also examined in the present study. Researchers have adopted different information criteria for model selection. For example, Bolsinova, De Boeck, and Tijmstra (2017) and Bolsinova and Tijmstra (2017) both used DIC, Ranger and Ortner's (2012) study was based on AIC, and Bolsinova, Tijmstra, and Molenaar (2017) compared AIC and BIC. Nevertheless, a comprehensive comparison has not been conducted among AIC, AICc, BIC, and DIC in simulated settings when the conditional independence assumption is violated. As such, results from this study regarding model selection criteria provided unique information about the effectiveness of these information-based model fit indices in the joint modeling framework assuming conditional dependence between responses and RTs. Note that because each simulation condition was only replicated 30 times, the percentage of correctly-identified model may be subject to larger random variations. Yet, the comparison among model fit indices across conditions provided some general conclusions.

In the present study, findings based on deviance, AIC, AICc, and BIC were similar for data generated from both models. For both data generating models, deviance, AIC, AICc, and BIC all chose one of the two models with the correctly specified shift indicator and random effects as the best fitting model. Surprisingly, as a goodness-of-fit statistic without penalty of number of parameters, deviance outperformed AIC, AICc, and BIC in terms of identifying the best fitting model. Even so, deviance only favored the true data generating model in around 50% of the replications when data were generated from the JM-RD1, and slightly higher when data were generated from the JM-DD1. As the penalty of number of parameters got stronger, AIC, AICc, and BIC were inclined to favoring the simpler model with the correctly specified shift indicator and random effects. The model fit index with the largest penalty of number of parameters, BIC, almost always yielded the smallest values for the JM-R or the JM-D. This contradicted with the findings from Bolsinova, Tijmstra, and Molenaar (2017), where BIC was recommended over AIC for model selection regarding models accounting for conditional dependence.

The performance of DIC, however, varied depending on the data generating model. For data generated from the JM-RD1, DIC performed as good as deviance, which identified the true data generating model in about 50% of the replications. For data generated from the JM-DD1, DIC also operated equivalently to deviance when test length was 40. Yet, when test length was 20, DIC tended to favor the HM and the models conditioning on observed responses more than the models with the correctly specified shift indicator. As illustrated before, DIC may over-penalize the models conditioning on item-person distance because the shift indicator was determined by two latent variables. Fox and Marianti (2016) also pointed out that a straightforward implementation of DIC would not produce reliable results since the estimation of the number of effective parameters would be very complex when the model contained many random effects, outcomes of different types (categorical and continuous), and multiple link functions (linear and nonlinear). To simplify the computation of the penalty term, a modified version of DIC based on the integrated likelihood (e.g., Berger, Brunero, & Wolpert, 1999) could be considered as an alternative (Fox, 2010; Klein Entink, Fox, & van der Linden, 2009).

In summary, the model selection criteria considered in the present study did not perform well in term of distinguishing models with the same shift indicator and random effects. This is not unforeseen because such models often performed similarly in the simulation studies regarding parameter recovery. However, they were generally able to identify models with the correctly specified shift indicator. When the true data generating model was unknown and test length was small, models selected based on deviance, AIC, AICc, and BIC were more likely to reflect the true latent structure than DIC.

5.2 Applications of the Speed-Accuracy-Difficulty Interaction

In the present study, the results from simulation studies showed that the parameter estimates would be biased when the conditional dependence between responses and RTs was ignored, especially for the RT model-related parameters. Additionally, the application of the proposed models was demonstrated through two datasets in the empirical data analyses section.

The six proposed models included three models for each of the two shift indicators, representing different shift mechanisms. The first shift indicator depended on the observed item responses, where the locations of RT distributions for correct and incorrect responses were different. Models employed this shift indicator were inspired by the phenomenon depicted in Figures 1 and 2. In fact, this observation reflected different pacing strategies in examinees' responding behaviors.

For an easier item, those who were able to provide a correct answer to this item tended to respond fast and correctly, while others responded slowly and incorrectly. This may relate to two common concepts in cognitive psychology, simple

RT task and choice RT task (e.g., Fowler, Brown, Sabadini, & Weihing, 2003; Logan, Cowan, & Davis, 1984). A simple RT task only has one stimulus and a choice RT task has multiple stimuli which the examinees need to respond to. According to Hick's law, the more stimuli there are, the slower the responses (Hyman, 1953). Based on this theory, an easier item may perform like a simple RT task or appear to have fewer stimuli for examinees who answered correctly. However, it may resemble a choice RT task with more stimuli for those who answered incorrectly, which led to the differences in RT distributions. Examinees who solved the item correctly with fewer stimuli were likely to have higher ability levels than those who failed to provide a correct answer even with longer RTs.

For a relatively difficult item, examinees who answered correctly were likely to be slow and correct, whereas those who were not able to answer correctly respond fast and incorrectly. Same logic also applies to difficult items that an item may function differently for different groups of examinees regarding the number of stimuli. Yet another important consideration in timed tests is time allocation. For difficult items, examinees may intentionally skip them or randomly select an answer to save time on items that they were able to respond correctly to, resulting in incorrect answers with short RTs. This aligns with the pacing strategy discussed in Wang and Zhang (2006) that examinees tend to spend more time on items with similar difficulty levels as their ability levels. As such, examinees' ability levels often appear to be associated with the number of stimuli of an item, as well as the choice of time allocation. A shift indicator based on the distance between ability and item difficulty was therefore proposed.

In the present study, two datasets from large-scale assessment programs provided different perspectives in exploring the conditional independence assumption in real testing scenarios. Considering the effectiveness of the model fit indices, assuming conditional dependence existed between responses and RTs was more realistic than conditional independence. Yet, the two datasets demonstrated different latent structures underlying the data. Ability and speed was weakly positively correlated for dataset 1, but strongly negatively correlated for dataset 2. This may be due to the fact that dataset 1 came from a high-stakes exam, whereas dataset 2 was low-stakes in nature. As such, examinees in dataset 2 might be less motivated to spend time and do well on the assessment.

Additionally, models with item responses as the shift indicator performed better on dataset 1, and models with item-person distance as the shift indicator was selected on dataset 2. While tem difficulty and the shift in RT distributions was positively strongly correlated in dataset 1, they were only weakly related in dataset 2. Since the correlation between item difficulty and the shift was found to be negatively biased in the simulation studies, the true correlation between item difficulty and the shift was expected to be stronger than the estimates. Another possible reason was that the estimation accuracy was lower with only 10 items in dataset 2. Nonetheless, the choice of pacing strategies existed in both high-stakes and low-stakes assessment, even though the correlation between ability and speed might vary. Lastly, the variance of random effects was estimated to be about the same magnitude as the speed variance or even larger, indicating that this effect may not be negligible.

5.3 Limitations and Future Directions

Despite the findings in the present study, there are a number of limitations that need to be addressed in future explorations, especially regarding model extensions, simulation design, and the choice of priors. In terms of model extensions, the present study only explored a few modeling options limited by the scope of the study. Yet, there are much more possible modeling approaches to be further investigated. First, this study adopted a shift indicator based on the relative distance between ability and item difficulty, where the threshold was fixed at zero. It might be more reasonable to estimate the threshold to be item-specific or the same across items. Previous studies have shown that low-ability students tend to benefit more from extended RT (Clauser, Margolis, & von Davier, 2017; Harik, 2017), embedding a variable threshold could provide detailed examination regarding which examinees may need more time on each item.

Second, the current study only considered item difficulty as a predictor of the shift magnitude. Although item difficulty alone explained up to 50% of the total variance in the shift, the models with random effects that are conditioned on the same shift indicator could not be well distinguished in neither the simulation studies nor the empirical data analyses. In the future studies, more item features could be included to explore what features are associated with differences in RT distributions for different examinee groups, which may provide more predictive power to further differentiate the models with random effects. For instance, Mulholland, Pellegrino, and Glaser (1980) conducted ANOVA to investigate the relationship between item features and RT distributions. In the joint modeling framework, the linear logistic test model

(LLTM; Fischer, 1973) could be utilized to incorporate other item features, such as word count, dichotomous or polytomous items, with or without figures, content domain, etc.

Third, changes in the variance of RT distributions could be modeled to provide additional information to the changes in location depending on different examinee groups, similar to what van Rijn and Ali (2017) mentioned in their study. Fourth, the models used for item responses and RTs at the first level could easily be substituted by other IRT models, cognitive diagnostic models (e.g., Rupp, Templin, & Henson, 2010), and RT models. Lastly, given that the examinee sample was selected from a larger population, sampling effects also could be incorporated to ensure the generalizability of the results.

In terms of the design of simulation studies, the present study put constraints on some factors to ensure that the study can be completed in a reasonable time frame. Nonetheless, other levels of manipulated factors could be taken into account, and the factors fixed in this study could be varied as well. In particular, the variance of speed parameters and the variance of random effects was fixed at .25 and .04 in the present study. However, it was found in real data examples that the variance of random effects were almost as large as or even larger than the variance of speed parameters. This may explain why the manipulated correlation between item difficulty and the shift did not significantly affect parameter recovery. Even so, models with random effects outperformed those without random effects in most simulated scenarios. Future studies may consider the relative magnitude of these variances in conducting simulation studies. Moreover, this study only compared the effectiveness of several relative model fit indices. Assessing the performance of available absolute model fit indices in the context of joint modeling of responses and RTs might be another interesting direction, such as the posterior predictive checks (e.g., Gelman, Meng, & Stern, 1996; Rubin, 1984).

In the present study, the time discrimination parameters were assumed to be independent of the time intensity and item difficulty parameters. In other words, the correlation between time discrimination and time intensity and the correlation between time discrimination and item difficulty were not modeled; only the bivariate relationship between time intensity and item difficulty was taken into account in the prior for item parameters. Such a prior was chosen based on the empirical evidence from van der Linden (2007) and Molenaar, Tuerlinckx, and van der Maas (2015b). It was also much more computationally efficient than modeling the trivariate relationship among the item parameters. However, these benefits were obtained at the cost of neglecting other associations among the item parameters. If these associations are of interest in the future studies, a fully specified prior proposed in van der Linden (2007) could be used instead.

Despite the limitations, this study contributes to the literature about joint modeling of responses and RTs with a focus on the conditional dependence between responses and RTs. As computer-based assessment becomes increasingly popular, more information should be extracted from examinees' RTs and carefully interpreted. This study provides important evidence about how different groups of examinees allocate test time, depending on the observed responses or the distance between their ability levels and the item difficulty level. Further, this study explores how examinees’ pacing strategies are related to item difficulty, the most important psychometric feature of an item. Models with two different mechanisms that might lead to the speed-accuracy-difficulty interaction have been evaluated with simulation studies. Empirical data analyses also show evidence that advocates the use of the proposed models. Nevertheless, the consequences of ignoring the conditional dependence between responses and RTs have been summarized to provide inference on modeling choices to practitioners. In sum, this study complements existing literature and builds a good foundation for further explorations.

Appendix A

Table A1a. *Mean and SD of bias in item difficulty estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	-.003	-.009	-.002	-.003	.023	.108	.021	.021
			.7	.005	.005	.005	.005	.022	.090	.020	.021
		.5	.3	-.004	-.003	-.004	-.004	.018	.082	.016	.016
			.7	-.008	-.004	-.009	-.008	.020	.098	.018	.018
		.8	.3	-.001	.004	-.001	-.001	.020	.064	.020	.019
			.7	-.003	-.002	-.003	-.003	.020	.071	.022	.023
	40	.2	.3	-.001	.002	-.002	-.002	.025	.094	.024	.025
			.7	-.001	-.003	-.001	-.001	.035	.108	.031	.031
		.5	.3	.001	.003	.001	.001	.020	.103	.019	.020
			.7	-.005	-.003	-.005	-.005	.029	.079	.027	.027
		.8	.3	.000	.001	.000	.000	.026	.071	.025	.025
			.7	-.001	-.002	-.001	-.001	.023	.092	.023	.023
1000	20	.2	.3	-.010	-.009	-.010	-.010	.014	.062	.014	.014
			.7	.002	.004	.002	.002	.016	.086	.015	.016
		.5	.3	.002	.002	.002	.002	.018	.026	.018	.018
			.7	.000	.000	.000	.000	.021	.110	.020	.020
		.8	.3	-.002	-.001	-.002	-.002	.017	.089	.016	.016
			.7	.000	.002	.000	.000	.015	.085	.014	.014
	40	.2	.3	.002	.002	.002	.002	.015	.032	.015	.015
			.7	-.001	-.001	-.001	-.001	.013	.065	.012	.013
		.5	.3	.001	.002	.001	.001	.017	.064	.017	.017
			.7	-.005	-.002	-.005	-.005	.016	.087	.014	.014
		.8	.3	-.002	-.003	-.002	-.002	.014	.082	.014	.014
			.7	.003	.005	.003	.004	.017	.087	.015	.015

Table A1b. *Mean and SD of SE in item difficulty estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.101	.098	.101	.102	.016	.017	.016	.016
			.7	.102	.099	.103	.102	.020	.020	.020	.020
		.5	.3	.102	.101	.102	.102	.020	.020	.020	.019
			.7	.098	.094	.099	.099	.015	.016	.015	.015
		.8	.3	.103	.103	.103	.103	.016	.015	.016	.016
			.7	.106	.102	.106	.106	.022	.023	.022	.022
	40	.2	.3	.103	.103	.103	.103	.015	.015	.015	.015
			.7	.101	.097	.101	.101	.018	.017	.018	.018
		.5	.3	.102	.100	.103	.103	.017	.018	.017	.017
			.7	.100	.098	.101	.101	.013	.013	.013	.013
		.8	.3	.107	.106	.108	.108	.017	.017	.017	.017
			.7	.108	.102	.108	.108	.016	.017	.015	.015
1000	20	.2	.3	.073	.074	.073	.073	.016	.016	.015	.015
			.7	.074	.070	.074	.074	.010	.010	.010	.010
		.5	.3	.071	.072	.071	.071	.016	.016	.016	.017
			.7	.069	.065	.069	.069	.010	.010	.010	.010
		.8	.3	.074	.074	.073	.073	.021	.020	.021	.021
			.7	.075	.074	.075	.075	.015	.016	.015	.015
	40	.2	.3	.072	.072	.072	.072	.012	.013	.012	.012
			.7	.073	.071	.073	.073	.014	.013	.014	.013
		.5	.3	.071	.071	.072	.071	.010	.010	.010	.010
			.7	.073	.071	.073	.073	.013	.012	.013	.013
		.8	.3	.071	.070	.071	.071	.012	.012	.012	.012
			.7	.071	.068	.072	.072	.014	.012	.014	.014

Table A1c. *Mean and SD of RMSE in item difficulty estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.103	.135	.103	.104	.016	.054	.016	.016
			.7	.104	.129	.105	.105	.020	.038	.020	.020
		.5	.3	.103	.126	.103	.103	.020	.034	.020	.020
			.7	.101	.128	.101	.101	.015	.044	.016	.016
		.8	.3	.105	.119	.104	.104	.015	.023	.016	.015
			.7	.108	.121	.109	.109	.021	.036	.022	.022
	40	.2	.3	.106	.134	.105	.106	.016	.039	.016	.016
			.7	.106	.138	.105	.105	.021	.045	.020	.020
		.5	.3	.104	.136	.104	.104	.018	.048	.017	.017
			.7	.104	.123	.104	.104	.014	.027	.014	.015
		.8	.3	.110	.126	.110	.110	.019	.025	.019	.019
			.7	.110	.132	.111	.111	.016	.041	.016	.016
1000	20	.2	.3	.075	.094	.075	.075	.016	.026	.016	.016
			.7	.075	.104	.075	.075	.009	.037	.009	.009
		.5	.3	.073	.076	.073	.073	.019	.018	.018	.019
			.7	.071	.116	.072	.071	.013	.048	.012	.012
		.8	.3	.075	.110	.075	.075	.022	.035	.022	.022
			.7	.076	.107	.077	.077	.015	.035	.015	.015
	40	.2	.3	.073	.079	.073	.073	.014	.015	.013	.013
			.7	.074	.092	.074	.074	.014	.031	.014	.014
		.5	.3	.073	.091	.073	.073	.012	.029	.012	.012
			.7	.074	.107	.074	.074	.014	.033	.014	.014
		.8	.3	.072	.102	.072	.072	.012	.034	.012	.012
			.7	.073	.103	.073	.073	.014	.040	.015	.015

Table A2a. *Mean and SD of bias in time discrimination estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	-.009	-.028	-.009	-.092	.013	.028	.013	.084
			.7	-.008	-.022	-.008	-.092	.011	.021	.011	.080
		.5	.3	-.011	-.036	-.011	-.099	.009	.029	.009	.067
			.7	-.009	-.021	-.009	-.100	.012	.021	.012	.066
		.8	.3	-.010	-.036	-.010	-.103	.013	.038	.013	.083
			.7	-.010	-.023	-.010	-.092	.014	.023	.014	.055
	40	.2	.3	-.009	-.034	-.009	-.099	.013	.035	.013	.083
			.7	-.011	-.025	-.011	-.099	.010	.017	.010	.069
		.5	.3	-.008	-.032	-.008	-.099	.009	.037	.009	.090
			.7	-.008	-.021	-.008	-.090	.012	.025	.012	.071
		.8	.3	-.007	-.033	-.007	-.093	.012	.032	.012	.068
			.7	-.008	-.022	-.008	-.095	.012	.023	.012	.080
1000	20	.2	.3	-.003	-.031	-.003	-.095	.008	.035	.008	.073
			.7	-.007	-.020	-.007	-.096	.007	.018	.007	.060
		.5	.3	-.006	-.033	-.006	-.088	.009	.037	.009	.087
			.7	-.005	-.015	-.005	-.100	.011	.018	.011	.073
		.8	.3	-.006	-.033	-.006	-.091	.007	.033	.007	.079
			.7	-.006	-.021	-.006	-.093	.010	.019	.010	.077
	40	.2	.3	-.004	-.033	-.004	-.089	.010	.035	.010	.073
			.7	-.005	-.019	-.005	-.083	.009	.025	.009	.062
		.5	.3	-.005	-.031	-.005	-.089	.008	.038	.008	.083
			.7	-.005	-.019	-.005	-.090	.007	.017	.007	.063
		.8	.3	-.005	-.032	-.005	-.090	.010	.035	.010	.099
			.7	-.005	-.019	-.005	-.093	.010	.022	.010	.067

Table A2b. *Mean and SD of SE in time discrimination estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.064	.063	.064	.062	.010	.010	.010	.010
			.7	.065	.065	.065	.062	.007	.007	.007	.007
		.5	.3	.061	.061	.061	.058	.007	.007	.007	.006
			.7	.062	.062	.062	.059	.008	.008	.008	.009
		.8	.3	.063	.062	.063	.059	.007	.007	.007	.008
			.7	.064	.065	.064	.061	.008	.008	.008	.009
	40	.2	.3	.063	.062	.063	.060	.008	.008	.008	.009
			.7	.061	.061	.061	.058	.008	.008	.008	.008
		.5	.3	.063	.062	.063	.060	.007	.007	.007	.007
			.7	.063	.062	.063	.059	.010	.010	.010	.009
		.8	.3	.060	.060	.060	.058	.008	.008	.008	.007
			.7	.061	.061	.061	.058	.008	.008	.008	.007
1000	20	.2	.3	.048	.048	.048	.047	.006	.007	.006	.007
			.7	.044	.043	.044	.042	.007	.007	.007	.008
		.5	.3	.046	.045	.046	.043	.007	.007	.007	.007
			.7	.048	.048	.048	.046	.007	.007	.007	.006
		.8	.3	.046	.044	.046	.044	.007	.008	.007	.007
			.7	.046	.046	.046	.044	.007	.007	.007	.008
	40	.2	.3	.044	.043	.044	.042	.005	.005	.005	.005
			.7	.044	.044	.044	.042	.006	.005	.006	.006
		.5	.3	.043	.043	.043	.041	.006	.006	.006	.006
			.7	.044	.044	.044	.042	.006	.006	.006	.006
		.8	.3	.044	.043	.044	.041	.006	.005	.006	.006
			.7	.045	.045	.045	.043	.006	.006	.006	.006

Table A2c. *Mean and SD of RMSE in time discrimination estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.066	.074	.066	.119	.009	.015	.009	.071
			.7	.066	.071	.066	.120	.007	.012	.007	.065
		.5	.3	.063	.075	.063	.120	.007	.018	.007	.056
			.7	.063	.067	.063	.122	.009	.013	.009	.054
		.8	.3	.065	.078	.065	.125	.007	.023	.007	.073
			.7	.067	.071	.067	.115	.008	.014	.008	.045
	40	.2	.3	.065	.076	.065	.125	.008	.022	.008	.069
			.7	.063	.068	.063	.120	.007	.010	.007	.058
		.5	.3	.064	.076	.064	.123	.007	.025	.007	.078
			.7	.064	.069	.064	.116	.010	.016	.010	.058
		.8	.3	.061	.073	.061	.116	.008	.019	.008	.056
			.7	.063	.068	.063	.119	.008	.015	.008	.068
1000	20	.2	.3	.049	.063	.049	.111	.006	.023	.006	.064
			.7	.045	.050	.045	.109	.007	.010	.007	.053
		.5	.3	.047	.061	.047	.105	.007	.028	.007	.079
			.7	.050	.052	.050	.115	.006	.009	.006	.064
		.8	.3	.046	.060	.046	.106	.006	.022	.006	.071
			.7	.048	.053	.048	.109	.007	.012	.007	.068
	40	.2	.3	.045	.060	.045	.105	.005	.023	.005	.064
			.7	.045	.052	.045	.098	.006	.015	.006	.054
		.5	.3	.044	.059	.044	.104	.006	.028	.006	.076
			.7	.045	.050	.045	.104	.006	.009	.006	.055
		.8	.3	.045	.059	.045	.107	.005	.026	.005	.090
			.7	.046	.052	.046	.108	.006	.014	.006	.058

Table A3a. *Mean and SD of bias in time intensity estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.000	-.002	-.001	-.173	.009	.096	.009	.181
			.7	-.003	.000	-.005	-.176	.010	.067	.011	.146
		.5	.3	.001	.010	.000	-.160	.013	.108	.010	.129
			.7	-.001	.004	-.003	-.181	.010	.063	.010	.147
		.8	.3	.002	.012	.001	-.161	.010	.090	.008	.097
			.7	.000	.003	-.002	-.171	.010	.059	.011	.134
	40	.2	.3	.000	.008	-.001	-.167	.010	.094	.008	.128
			.7	.001	.003	.000	-.176	.010	.064	.010	.146
		.5	.3	.000	.005	-.002	-.171	.009	.090	.008	.147
			.7	.002	.007	.000	-.169	.012	.076	.010	.123
		.8	.3	.000	.004	-.001	-.161	.010	.088	.008	.125
			.7	.001	.003	.000	-.176	.011	.072	.010	.151
1000	20	.2	.3	-.001	.004	-.001	-.163	.005	.091	.005	.112
			.7	.000	.003	-.001	-.177	.005	.058	.005	.132
		.5	.3	.000	-.004	.000	-.146	.006	.089	.006	.101
			.7	.000	.003	-.001	-.185	.007	.053	.004	.169
		.8	.3	.002	.007	.000	-.159	.009	.103	.006	.148
			.7	.001	.004	.000	-.173	.006	.076	.005	.132
	40	.2	.3	.000	.002	.000	-.153	.005	.091	.005	.116
			.7	.000	.002	-.001	-.167	.006	.064	.007	.133
		.5	.3	-.001	.001	-.001	-.162	.006	.103	.006	.139
			.7	.000	.005	-.001	-.173	.007	.072	.007	.138
		.8	.3	.000	.002	-.001	-.164	.006	.104	.006	.146
			.7	.000	.004	-.001	-.175	.006	.072	.006	.132

Table A3b. *Mean and SD of SE in time intensity estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.033	.026	.033	.022	.008	.006	.008	.004
			.7	.033	.025	.033	.023	.008	.005	.008	.004
		.5	.3	.032	.025	.031	.023	.008	.006	.007	.002
			.7	.033	.025	.033	.022	.008	.005	.008	.002
		.8	.3	.034	.024	.034	.023	.009	.006	.009	.002
			.7	.032	.025	.032	.024	.009	.005	.009	.003
	40	.2	.3	.033	.025	.033	.024	.007	.004	.007	.003
			.7	.032	.025	.033	.023	.007	.004	.008	.003
		.5	.3	.031	.024	.031	.024	.008	.004	.008	.003
			.7	.030	.023	.031	.023	.006	.003	.006	.003
		.8	.3	.033	.024	.033	.024	.009	.004	.009	.002
			.7	.032	.025	.033	.023	.008	.004	.008	.003
1000	20	.2	.3	.024	.017	.024	.017	.006	.002	.006	.002
			.7	.023	.017	.023	.015	.006	.003	.006	.002
		.5	.3	.024	.018	.023	.017	.007	.005	.006	.002
			.7	.024	.018	.024	.016	.007	.003	.008	.002
		.8	.3	.024	.020	.024	.016	.008	.009	.007	.002
			.7	.023	.017	.023	.016	.006	.002	.006	.002
	40	.2	.3	.024	.018	.023	.017	.007	.003	.007	.002
			.7	.023	.017	.023	.016	.006	.003	.006	.002
		.5	.3	.023	.016	.023	.016	.007	.003	.007	.002
			.7	.022	.017	.023	.016	.005	.003	.006	.002
		.8	.3	.023	.016	.023	.016	.007	.002	.007	.002
			.7	.023	.016	.023	.015	.005	.003	.005	.002

Table A3c. *Mean and SD of RMSE in time intensity estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.034	.074	.034	.177	.009	.065	.009	.178
			.7	.034	.060	.035	.181	.009	.038	.010	.141
		.5	.3	.033	.085	.033	.167	.011	.068	.009	.121
			.7	.034	.054	.034	.189	.008	.040	.009	.138
		.8	.3	.035	.076	.035	.169	.011	.054	.010	.086
			.7	.033	.056	.033	.179	.010	.028	.011	.125
	40	.2	.3	.034	.078	.034	.175	.009	.058	.008	.118
			.7	.033	.058	.034	.182	.008	.037	.009	.139
		.5	.3	.032	.073	.032	.180	.008	.058	.008	.138
			.7	.032	.064	.032	.174	.008	.048	.007	.119
		.8	.3	.034	.079	.034	.168	.010	.045	.010	.117
			.7	.034	.062	.034	.183	.009	.043	.009	.145
1000	20	.2	.3	.024	.076	.024	.169	.006	.050	.006	.104
			.7	.023	.050	.024	.184	.006	.032	.006	.122
		.5	.3	.024	.078	.024	.149	.007	.045	.006	.098
			.7	.025	.047	.025	.192	.008	.028	.008	.162
		.8	.3	.025	.084	.025	.165	.009	.061	.007	.142
			.7	.024	.060	.024	.177	.006	.048	.006	.128
	40	.2	.3	.024	.077	.024	.159	.007	.050	.007	.109
			.7	.023	.054	.023	.170	.007	.037	.007	.129
		.5	.3	.023	.076	.024	.165	.007	.070	.007	.136
			.7	.023	.060	.023	.179	.006	.042	.007	.131
		.8	.3	.024	.083	.024	.166	.007	.064	.007	.145
			.7	.023	.054	.023	.178	.005	.050	.006	.128

Table A4a. *Mean and SD of bias in ability estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.000	.000	.000	.000	.239	.241	.239	.245
			.7	.000	.000	.000	.000	.239	.241	.239	.244
		.5	.3	.000	.000	.000	.000	.250	.251	.250	.262
			.7	.000	.000	.000	.000	.246	.248	.246	.259
		.8	.3	.000	.000	.000	.000	.254	.254	.254	.268
			.7	.000	.000	.000	.000	.259	.260	.259	.272
	40	.2	.3	.000	.000	.000	.000	.146	.147	.146	.148
			.7	.000	.000	.000	.000	.133	.134	.133	.136
		.5	.3	.000	.000	.000	.000	.161	.163	.161	.168
			.7	.000	.000	.000	.000	.164	.164	.163	.170
		.8	.3	.000	.000	.000	.000	.172	.172	.172	.183
			.7	.000	.000	.000	.000	.171	.172	.171	.181
1000	20	.2	.3	.000	.000	.000	.000	.243	.243	.243	.249
			.7	.000	.000	.000	.000	.239	.240	.239	.245
		.5	.3	.000	.000	.000	.000	.238	.239	.238	.249
			.7	.000	.000	.000	.000	.243	.245	.243	.254
		.8	.3	.000	.000	.000	.000	.253	.254	.253	.269
			.7	.000	.000	.000	.000	.254	.255	.254	.270
	40	.2	.3	.000	.000	.000	.000	.150	.150	.150	.152
			.7	.000	.000	.000	.000	.147	.147	.146	.148
		.5	.3	.000	.000	.000	.000	.154	.155	.154	.161
			.7	.000	.000	.000	.000	.156	.157	.156	.162
		.8	.3	.000	.000	.000	.000	.174	.175	.174	.185
			.7	.000	.000	.000	.000	.173	.174	.173	.186

Table A4b. *Mean and SD of SE in ability estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.392	.391	.392	.390	.054	.054	.054	.054
			.7	.398	.397	.398	.396	.056	.056	.056	.056
		.5	.3	.378	.377	.378	.372	.050	.050	.050	.049
			.7	.375	.374	.375	.369	.051	.051	.051	.051
		.8	.3	.315	.315	.315	.304	.042	.043	.042	.042
			.7	.309	.308	.309	.298	.041	.041	.041	.040
	40	.2	.3	.318	.318	.318	.317	.046	.046	.046	.045
			.7	.316	.315	.316	.315	.042	.042	.042	.042
		.5	.3	.308	.307	.308	.306	.044	.044	.044	.044
			.7	.306	.305	.306	.303	.042	.042	.042	.042
		.8	.3	.265	.265	.265	.259	.037	.037	.037	.036
			.7	.268	.268	.269	.262	.037	.037	.037	.036
1000	20	.2	.3	.394	.394	.394	.392	.055	.055	.055	.055
			.7	.398	.397	.398	.396	.055	.055	.055	.055
		.5	.3	.373	.373	.373	.367	.052	.052	.052	.051
			.7	.376	.375	.376	.370	.050	.050	.050	.049
		.8	.3	.309	.308	.309	.298	.044	.044	.044	.043
			.7	.311	.310	.311	.299	.041	.041	.041	.040
	40	.2	.3	.319	.319	.319	.318	.044	.044	.044	.044
			.7	.318	.318	.319	.318	.043	.043	.043	.043
		.5	.3	.307	.307	.307	.305	.043	.043	.043	.042
			.7	.306	.306	.306	.303	.042	.042	.042	.042
		.8	.3	.270	.269	.270	.263	.036	.035	.036	.034
			.7	.267	.267	.267	.260	.035	.035	.036	.034

Table A4c. *Mean and SD of RMSE in ability estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.450	.450	.450	.451	.104	.106	.104	.107
			.7	.457	.457	.457	.457	.099	.100	.099	.101
		.5	.3	.443	.443	.443	.443	.108	.108	.108	.112
			.7	.439	.439	.439	.440	.104	.105	.104	.110
		.8	.3	.392	.392	.392	.391	.108	.108	.108	.114
			.7	.389	.390	.390	.388	.111	.111	.111	.119
	40	.2	.3	.346	.346	.346	.346	.070	.071	.070	.070
			.7	.339	.339	.340	.340	.061	.062	.061	.061
		.5	.3	.341	.341	.341	.341	.083	.083	.083	.084
			.7	.337	.337	.337	.337	.091	.092	.091	.092
		.8	.3	.310	.310	.310	.310	.072	.072	.072	.076
			.7	.312	.312	.312	.312	.072	.072	.071	.075
1000	20	.2	.3	.453	.453	.453	.454	.109	.109	.109	.111
			.7	.455	.455	.455	.456	.105	.106	.105	.108
		.5	.3	.434	.434	.434	.434	.101	.102	.101	.105
			.7	.440	.440	.440	.440	.098	.099	.097	.101
		.8	.3	.388	.388	.388	.387	.104	.105	.104	.113
			.7	.390	.390	.390	.389	.102	.103	.102	.112
	40	.2	.3	.347	.347	.347	.347	.073	.073	.073	.072
			.7	.345	.345	.345	.345	.074	.075	.074	.074
		.5	.3	.340	.340	.340	.340	.069	.069	.069	.070
			.7	.338	.338	.338	.338	.073	.074	.074	.075
		.8	.3	.316	.315	.316	.315	.069	.069	.069	.073
			.7	.312	.312	.312	.312	.072	.073	.073	.078

Table A5a. *Mean and SD of bias in speed estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.000	.000	.000	.000	.032	.033	.032	.061
			.7	.000	.000	.000	.000	.031	.032	.031	.060
		.5	.3	.000	.000	.000	.000	.032	.033	.032	.059
			.7	.000	.000	.000	.000	.032	.032	.032	.062
		.8	.3	.000	.000	.000	.000	.035	.036	.035	.059
			.7	.000	.000	.000	.000	.033	.034	.033	.057
	40	.2	.3	.000	.000	.000	.000	.019	.019	.019	.057
			.7	.000	.000	.000	.000	.019	.019	.019	.058
		.5	.3	.000	.000	.000	.000	.019	.019	.019	.055
			.7	.000	.000	.000	.000	.019	.020	.019	.057
		.8	.3	.000	.000	.000	.000	.021	.022	.021	.054
			.7	.000	.000	.000	.000	.021	.021	.021	.054
1000	20	.2	.3	.000	.000	.000	.000	.031	.032	.031	.064
			.7	.000	.000	.000	.000	.031	.031	.031	.063
		.5	.3	.000	.000	.000	.000	.032	.033	.032	.061
			.7	.000	.000	.000	.000	.032	.032	.032	.062
		.8	.3	.000	.000	.000	.000	.035	.036	.035	.056
			.7	.000	.000	.000	.000	.034	.034	.034	.058
	40	.2	.3	.000	.000	.000	.000	.019	.019	.019	.055
			.7	.000	.000	.000	.000	.019	.019	.019	.056
		.5	.3	.000	.000	.000	.000	.020	.020	.020	.055
			.7	.000	.000	.000	.000	.020	.020	.020	.058
		.8	.3	.000	.000	.000	.000	.022	.022	.022	.053
			.7	.000	.000	.000	.000	.021	.021	.021	.057

Table A5b. *Mean and SD of SE in speed estimation in simulation study 1.*

J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	Mean				SD			
				JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.104	.105	.104	.108	.013	.013	.013	.014
			.7	.104	.104	.104	.108	.014	.014	.014	.014
		.5	.3	.103	.104	.103	.109	.013	.014	.013	.014
			.7	.103	.104	.103	.109	.013	.013	.013	.014
		.8	.3	.101	.102	.101	.110	.013	.013	.013	.014
			.7	.100	.100	.100	.108	.013	.013	.013	.014
	40	.2	.3	.074	.075	.074	.078	.010	.010	.010	.011
			.7	.075	.075	.075	.078	.010	.010	.010	.011
		.5	.3	.074	.075	.074	.078	.010	.010	.010	.010
			.7	.075	.075	.075	.079	.010	.010	.010	.010
		.8	.3	.074	.075	.074	.079	.010	.010	.010	.011
			.7	.073	.074	.073	.078	.010	.010	.010	.010
1000	20	.2	.3	.103	.104	.103	.108	.014	.014	.014	.014
			.7	.104	.105	.104	.109	.014	.014	.014	.014
		.5	.3	.103	.104	.103	.108	.013	.013	.013	.014
			.7	.104	.104	.104	.110	.014	.014	.014	.015
		.8	.3	.100	.101	.100	.108	.013	.013	.013	.015
			.7	.100	.100	.100	.108	.013	.013	.013	.014
	40	.2	.3	.076	.077	.076	.079	.010	.010	.010	.011
			.7	.075	.075	.075	.078	.010	.010	.010	.010
		.5	.3	.075	.076	.075	.079	.010	.010	.010	.011
			.7	.075	.075	.075	.078	.010	.010	.010	.010
		.8	.3	.073	.074	.073	.078	.010	.010	.010	.010
			.7	.073	.073	.073	.078	.010	.010	.010	.010

Table A5c. *Mean and SD of RMSE in speed estimation in simulation study 1.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.109	.110	.109	.123	.014	.014	.014	.022
			.7	.108	.109	.108	.122	.014	.015	.014	.022
		.5	.3	.108	.109	.108	.123	.015	.015	.015	.022
			.7	.108	.109	.108	.125	.014	.014	.014	.022
		.8	.3	.107	.108	.107	.124	.015	.015	.014	.021
			.7	.105	.106	.105	.121	.014	.014	.014	.020
	40	.2	.3	.077	.078	.077	.095	.010	.010	.010	.020
			.7	.077	.078	.077	.096	.010	.010	.010	.020
		.5	.3	.077	.077	.077	.094	.010	.010	.010	.019
			.7	.078	.078	.078	.095	.010	.010	.010	.020
		.8	.3	.077	.078	.077	.095	.011	.011	.011	.020
			.7	.076	.077	.076	.093	.010	.010	.010	.019
1000	20	.2	.3	.108	.109	.108	.124	.015	.015	.015	.023
			.7	.108	.109	.109	.125	.015	.015	.015	.021
		.5	.3	.107	.108	.107	.123	.014	.015	.015	.020
			.7	.108	.109	.108	.125	.015	.014	.015	.022
		.8	.3	.106	.107	.106	.121	.015	.015	.015	.020
			.7	.105	.106	.105	.122	.015	.015	.015	.020
	40	.2	.3	.078	.079	.078	.095	.011	.011	.011	.019
			.7	.077	.077	.077	.094	.010	.010	.010	.019
		.5	.3	.077	.078	.077	.095	.010	.011	.010	.020
			.7	.077	.078	.077	.096	.010	.010	.010	.020
		.8	.3	.076	.077	.076	.093	.010	.010	.010	.018
			.7	.076	.076	.076	.095	.010	.010	.010	.020

Table A6. *Bias, SE, and RMSE of ω_0 in simulation study 1.*

				Bias			SE			RMSE		
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-RD1	JM-RD2	JM-R	JM-RD1	JM-RD2	JM-R
500	20	.2	.3	.001	.012	.000	.009	.009	.010	.009	.015	.010
			.7	.002	.003	.003	.012	.012	.012	.012	.013	.012
		.5	.3	.001	-.006	.001	.013	.012	.013	.013	.014	.013
			.7	.002	-.004	.001	.012	.012	.011	.012	.013	.011
		.8	.3	-.003	-.021	-.003	.011	.010	.011	.012	.023	.012
			.7	.001	-.001	.001	.011	.011	.011	.011	.011	.011
	40	.2	.3	.000	-.008	.000	.009	.009	.008	.009	.012	.008
			.7	-.002	.001	-.004	.010	.011	.010	.011	.011	.011
		.5	.3	-.001	-.005	-.001	.009	.009	.009	.009	.010	.009
			.7	-.002	-.007	-.002	.008	.009	.008	.008	.011	.008
		.8	.3	.001	-.003	.001	.008	.007	.008	.008	.008	.008
			.7	-.001	.000	-.002	.008	.009	.008	.008	.009	.008
1000	20	.2	.3	.001	-.004	.000	.006	.006	.006	.006	.008	.006
			.7	.001	-.004	.001	.008	.008	.008	.008	.009	.008
		.5	.3	.001	.005	.001	.009	.009	.009	.009	.010	.009
			.7	.001	.001	.001	.008	.007	.008	.008	.007	.008
		.8	.3	-.002	-.005	-.001	.008	.009	.007	.008	.010	.008
			.7	.000	-.004	.000	.009	.009	.009	.009	.010	.009
	40	.2	.3	.000	.000	.000	.007	.006	.007	.007	.006	.007
			.7	-.001	-.001	.000	.006	.006	.006	.006	.006	.006
		.5	.3	.000	.001	.001	.005	.004	.004	.005	.005	.004
			.7	.002	-.005	.001	.005	.005	.005	.005	.007	.006
		.8	.3	.001	.005	.001	.005	.005	.005	.005	.007	.005
			.7	-.002	-.005	-.001	.006	.006	.006	.006	.007	.006

Table A7. *Bias, SE, and RMSE of ω_1 in simulation study 1.*

				Bias		SE		RMSE	
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-RD1	JM-RD2	JM-RD1	JM-RD2
500	20	.2	.3	.078	.088	.016	.019	.079	.090
			.7	.006	.018	.017	.019	.018	.026
		.5	.3	-.001	.038	.013	.015	.013	.040
			.7	.026	.043	.016	.017	.030	.046
		.8	.3	-.014	.018	.015	.020	.021	.027
			.7	-.014	-.006	.015	.015	.021	.016
	40	.2	.3	.033	.059	.012	.014	.035	.061
			.7	.012	.028	.012	.013	.016	.031
		.5	.3	.049	.082	.008	.009	.050	.082
			.7	-.036	-.018	.008	.008	.037	.020
		.8	.3	.004	.019	.007	.010	.008	.022
			.7	.003	.022	.010	.012	.011	.025
1000	20	.2	.3	.001	.013	.009	.011	.009	.017
			.7	.007	.010	.012	.013	.014	.017
		.5	.3	-.074	-.081	.010	.013	.074	.082
			.7	.067	.090	.009	.012	.068	.091
		.8	.3	.016	.047	.010	.015	.019	.050
			.7	-.007	.002	.009	.010	.011	.010
	40	.2	.3	-.032	-.023	.007	.009	.032	.025
			.7	-.040	-.027	.006	.008	.040	.028
		.5	.3	.005	.017	.007	.008	.009	.019
			.7	-.005	.003	.008	.009	.009	.009
		.8	.3	.014	.043	.006	.007	.015	.043
			.7	-.012	.004	.009	.010	.015	.011

Table A8. *Bias, SE, and RMSE of $\rho_{b\lambda}$ in simulation study 1.*

				Bias	SE	RMSE
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD1	JM-RD1
500	20	.2	.3	.076	.041	.086
			.7	-.241	.058	.248
		.5	.3	-.150	.034	.154
			.7	-.200	.045	.205
		.8	.3	-.180	.039	.184
			.7	-.299	.056	.304
	40	.2	.3	-.056	.031	.064
			.7	-.227	.032	.229
		.5	.3	-.013	.021	.025
			.7	-.344	.030	.345
		.8	.3	-.131	.017	.132
			.7	-.257	.037	.260
1000	20	.2	.3	-.147	.025	.149
			.7	-.247	.033	.249
		.5	.3	-.338	.027	.339
			.7	-.088	.025	.091
		.8	.3	-.108	.025	.110
			.7	-.292	.022	.293
	40	.2	.3	-.225	.017	.226
			.7	-.361	.023	.362
		.5	.3	-.131	.019	.132
			.7	-.265	.025	.267
		.8	.3	-.097	.018	.099
			.7	-.292	.025	.293

Table A9. *Bias, SE, and RMSE of σ_{ϕ}^{2} in simulation study 1.*

J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	Bias		SE		RMSE	
				JM-RD1	JM-R	JM-RD1	JM-R	JM-RD1	JM-R
500	20	.2	.3	.001	.018	.005	.007	.005	.019
			.7	.005	.009	.005	.007	.007	.011
		.5	.3	.012	.009	.006	.006	.013	.011
			.7	.006	.017	.004	.008	.007	.019
		.8	.3	.009	.005	.006	.006	.011	.008
			.7	.006	.002	.005	.006	.008	.006
	40	.2	.3	.002	.006	.003	.004	.004	.007
			.7	.002	.004	.002	.005	.003	.006
		.5	.3	.001	.008	.003	.003	.003	.009
			.7	.001	-.008	.002	.003	.002	.009
		.8	.3	.003	.003	.004	.004	.005	.005
			.7	.004	.005	.003	.004	.004	.006
1000	20	.2	.3	.013	.011	.004	.004	.014	.012
			.7	.007	.012	.002	.005	.007	.013
		.5	.3	.006	.000	.004	.003	.007	.003
			.7	.002	.029	.003	.005	.004	.030
		.8	.3	.012	.012	.004	.004	.012	.012
			.7	.008	.007	.003	.004	.009	.008
	40	.2	.3	.004	.000	.001	.001	.004	.001
			.7	.002	-.008	.002	.002	.003	.008
		.5	.3	.004	.004	.003	.003	.005	.005
			.7	.002	.001	.002	.003	.002	.003
		.8	.3	.000	.001	.003	.003	.003	.003
			.7	.003	.000	.002	.003	.003	.003

Table A10. *Bias, SE, and RMSE of μ_b in simulation study 1.*

J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	Bias				SE				RMSE			
				JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	-.003	-.009	-.002	-.002	.018	.019	.019	.018	.018	.021	.019	.018
			.7	.006	.004	.005	.006	.025	.025	.026	.025	.026	.026	.026	.025
		.5	.3	-.004	-.003	-.004	-.003	.026	.027	.026	.026	.027	.027	.026	.026
			.7	-.009	-.005	-.009	-.006	.019	.020	.020	.019	.021	.020	.022	.021
		.8	.3	-.001	.004	-.001	.001	.025	.026	.026	.025	.025	.027	.026	.025
			.7	-.003	-.002	-.003	-.001	.024	.023	.024	.024	.024	.024	.024	.024
	40	.2	.3	-.002	.002	-.002	-.001	.018	.017	.018	.018	.018	.018	.019	.018
			.7	-.001	-.003	-.001	.000	.016	.016	.015	.015	.016	.016	.015	.015
		.5	.3	.002	.004	.001	.002	.017	.018	.018	.018	.017	.018	.018	.018
			.7	-.005	-.003	-.005	-.005	.014	.013	.013	.014	.015	.013	.014	.015
		.8	.3	.000	.001	.000	.001	.017	.018	.018	.017	.017	.018	.018	.017
			.7	-.001	-.002	-.001	.000	.020	.019	.020	.019	.020	.019	.020	.019
1000	20	.2	.3	-.009	-.009	-.010	-.009	.021	.021	.021	.021	.023	.023	.023	.023
			.7	.001	.004	.002	.004	.017	.016	.016	.016	.017	.016	.016	.016
		.5	.3	.002	.002	.001	.003	.015	.016	.016	.016	.015	.016	.016	.016
			.7	.000	.000	.000	.001	.012	.014	.012	.013	.012	.014	.012	.013
		.8	.3	-.002	-.001	-.002	.000	.014	.014	.013	.013	.014	.014	.014	.013
			.7	.000	.002	.000	.002	.017	.018	.017	.017	.017	.018	.017	.017
	40	.2	.3	.002	.002	.001	.002	.013	.013	.013	.013	.013	.013	.013	.013
			.7	-.001	-.001	-.002	-.001	.011	.011	.010	.011	.011	.011	.010	.011
		.5	.3	.001	.002	.002	.001	.014	.014	.014	.014	.014	.014	.014	.014
			.7	-.004	-.001	-.005	-.004	.011	.011	.012	.010	.012	.011	.013	.011
		.8	.3	-.002	-.003	-.002	-.001	.013	.012	.013	.013	.014	.013	.013	.013
			.7	.003	.004	.003	.004	.010	.010	.010	.010	.010	.011	.010	.011

Table A11. *Bias, SE, and RMSE of μ_β in simulation study 1.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.000	-.002	-.001	-.171	.008	.008	.008	.006	.008	.008	.008	.171
			.7	-.003	.000	-.005	-.174	.007	.007	.007	.005	.008	.007	.008	.174
		.5	.3	.001	.010	.000	-.158	.009	.009	.008	.006	.009	.013	.008	.158
			.7	-.001	.004	-.003	-.179	.009	.008	.008	.005	.009	.009	.008	.179
		.8	.3	.002	.012	.001	-.160	.008	.008	.008	.005	.008	.014	.008	.160
			.7	.000	.003	-.002	-.170	.009	.009	.009	.006	.009	.010	.009	.170
	40	.2	.3	.000	.008	-.001	-.166	.005	.005	.005	.003	.005	.009	.005	.166
			.7	.001	.003	.000	-.175	.006	.006	.006	.004	.006	.007	.006	.175
		.5	.3	.000	.005	-.002	-.170	.004	.005	.004	.004	.005	.007	.005	.171
			.7	.002	.007	.000	-.169	.006	.006	.005	.004	.006	.009	.005	.169
		.8	.3	.000	.004	-.001	-.160	.006	.006	.006	.004	.006	.007	.006	.160
			.7	.001	.003	.000	-.175	.005	.005	.005	.003	.005	.006	.005	.175
1000	20	.2	.3	-.001	.004	-.001	-.162	.005	.005	.005	.004	.005	.007	.005	.162
			.7	.000	.003	-.001	-.176	.005	.006	.005	.004	.005	.006	.005	.176
		.5	.3	.000	-.004	.001	-.144	.006	.006	.006	.004	.006	.007	.007	.144
			.7	.000	.003	-.001	-.182	.004	.004	.005	.004	.005	.005	.005	.182
		.8	.3	.002	.007	.001	-.157	.006	.006	.005	.005	.006	.009	.005	.157
			.7	.001	.004	.000	-.172	.005	.006	.005	.003	.005	.007	.005	.172
	40	.2	.3	.000	.002	.000	-.153	.004	.005	.005	.003	.004	.005	.005	.153
			.7	.001	.002	-.001	-.166	.004	.003	.003	.003	.004	.004	.004	.166
		.5	.3	-.001	.001	-.001	-.161	.003	.003	.003	.003	.003	.003	.003	.161
			.7	.000	.005	-.001	-.172	.004	.004	.004	.003	.004	.006	.004	.172
		.8	.3	.000	.002	.000	-.163	.004	.004	.004	.003	.004	.005	.004	.163
			.7	.000	.004	-.001	-.174	.004	.004	.004	.003	.004	.005	.004	.174

Table A12. *Bias, SE, and RMSE of σ_b^2 in simulation study 1.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.116	.109	.118	.117	.048	.049	.047	.047	.125	.119	.127	.126
			.7	.116	.098	.118	.116	.057	.058	.056	.056	.129	.114	.131	.129
		.5	.3	.106	.085	.110	.109	.070	.069	.069	.068	.127	.110	.130	.129
			.7	.123	.112	.124	.125	.058	.060	.058	.059	.136	.127	.138	.138
		.8	.3	.119	.115	.120	.123	.061	.060	.060	.061	.134	.130	.134	.137
			.7	.120	.115	.122	.123	.056	.057	.054	.056	.133	.129	.134	.136
	40	.2	.3	.054	.037	.057	.055	.043	.042	.043	.042	.069	.056	.071	.069
			.7	.030	.015	.033	.033	.037	.038	.036	.036	.048	.041	.049	.049
		.5	.3	.066	.046	.066	.066	.042	.042	.042	.041	.078	.062	.078	.078
			.7	.038	.027	.038	.040	.044	.044	.043	.043	.058	.052	.058	.059
		.8	.3	.056	.049	.057	.057	.044	.043	.043	.043	.071	.066	.071	.072
			.7	.052	.040	.053	.054	.040	.040	.041	.040	.066	.056	.067	.067
1000	20	.2	.3	.122	.116	.122	.123	.043	.042	.041	.042	.130	.123	.129	.130
			.7	.114	.118	.114	.114	.039	.041	.040	.040	.121	.124	.121	.121
		.5	.3	.095	.092	.098	.095	.053	.054	.054	.053	.109	.106	.112	.108
			.7	.095	.072	.096	.096	.041	.045	.044	.043	.104	.085	.105	.105
		.8	.3	.105	.086	.106	.109	.053	.051	.052	.051	.118	.100	.118	.120
			.7	.112	.110	.112	.113	.052	.054	.051	.052	.123	.122	.123	.124
	40	.2	.3	.049	.046	.049	.049	.027	.027	.028	.027	.056	.054	.056	.056
			.7	.056	.053	.057	.057	.037	.037	.037	.037	.067	.064	.068	.068
		.5	.3	.048	.046	.048	.048	.027	.028	.028	.028	.055	.054	.056	.056
			.7	.052	.048	.052	.053	.028	.028	.029	.029	.059	.056	.060	.060
		.8	.3	.062	.047	.062	.063	.028	.028	.028	.029	.068	.055	.068	.069
			.7	.055	.046	.055	.056	.031	.030	.030	.031	.063	.055	.063	.064

Table A13. *Bias, SE, and RMSE of* $\rho_{b\beta}$ *in simulation study 1.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	-.036	-.007	-.032	.149	.027	.025	.026	.021	.044	.026	.041	.150
			.7	-.032	-.055	-.022	.158	.024	.026	.023	.017	.040	.061	.032	.159
		.5	.3	-.042	-.074	-.037	.080	.029	.029	.029	.025	.051	.080	.047	.083
			.7	-.032	-.040	-.026	.145	.022	.018	.022	.015	.039	.043	.034	.145
		.8	.3	-.047	-.085	-.045	.040	.025	.021	.025	.019	.054	.088	.051	.044
			.7	-.031	-.049	-.019	.158	.028	.029	.027	.022	.042	.057	.033	.160
	40	.2	.3	-.023	-.072	-.015	.135	.018	.023	.018	.014	.029	.076	.023	.135
			.7	-.017	-.018	-.010	.164	.023	.023	.023	.017	.029	.029	.025	.165
		.5	.3	-.018	-.044	-.011	.161	.020	.022	.020	.016	.027	.049	.023	.162
			.7	-.025	-.052	-.014	.142	.020	.020	.020	.015	.032	.056	.024	.143
		.8	.3	-.020	-.032	-.016	.117	.015	.016	.016	.012	.025	.036	.022	.118
			.7	-.020	-.038	-.010	.183	.020	.020	.020	.014	.028	.043	.022	.184
1000	20	.2	.3	-.040	-.060	-.038	.068	.015	.016	.016	.012	.043	.062	.042	.069
			.7	-.035	-.025	-.034	.120	.020	.017	.019	.015	.040	.030	.039	.121
		.5	.3	-.027	.003	-.029	.060	.020	.019	.020	.015	.034	.019	.035	.062
			.7	-.039	-.044	-.036	.163	.014	.011	.016	.012	.041	.045	.040	.163
		.8	.3	-.044	-.055	-.038	.093	.020	.020	.020	.012	.049	.059	.043	.094
			.7	-.040	-.053	-.037	.108	.017	.016	.017	.011	.044	.055	.041	.108
	40	.2	.3	-.021	-.025	-.020	.090	.014	.012	.014	.010	.025	.027	.025	.090
			.7	-.020	-.026	-.014	.155	.014	.013	.014	.010	.024	.029	.020	.155
		.5	.3	-.018	-.025	-.016	.125	.013	.014	.012	.009	.022	.028	.020	.126
			.7	-.018	-.034	-.013	.157	.014	.013	.014	.010	.023	.037	.019	.158
		.8	.3	-.017	-.022	-.015	.126	.014	.013	.013	.011	.022	.026	.020	.127
			.7	-.020	-.035	-.016	.143	.012	.013	.013	.009	.023	.038	.021	.143

Table A14. *Bias, SE, and RMSE of σ_{β}^{2} in simulation study 1.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.072	.085	.073	.157	.006	.006	.006	.006	.072	.086	.073	.157
			.7	.069	.055	.071	.110	.007	.007	.007	.007	.069	.055	.071	.110
		.5	.3	.067	.066	.068	.106	.005	.006	.005	.006	.067	.066	.068	.106
			.7	.070	.082	.071	.149	.007	.005	.007	.006	.071	.082	.072	.149
		.8	.3	.070	.100	.070	.128	.009	.007	.008	.007	.070	.100	.071	.128
			.7	.066	.051	.068	.099	.008	.005	.008	.006	.066	.051	.068	.099
	40	.2	.3	.032	.017	.033	.061	.005	.004	.006	.005	.032	.017	.034	.061
			.7	.032	.034	.033	.078	.005	.004	.005	.004	.032	.034	.033	.078
		.5	.3	.034	.021	.035	.074	.005	.004	.005	.004	.034	.021	.035	.074
			.7	.031	.023	.032	.066	.005	.003	.005	.004	.031	.023	.033	.066
		.8	.3	.032	.042	.033	.081	.005	.005	.005	.004	.033	.042	.033	.082
			.7	.032	.021	.034	.071	.005	.004	.005	.004	.032	.021	.034	.071
1000	20	.2	.3	.067	.063	.067	.096	.007	.004	.007	.004	.067	.063	.067	.096
			.7	.071	.089	.071	.143	.006	.004	.006	.004	.071	.089	.071	.143
		.5	.3	.071	.052	.072	.072	.005	.004	.004	.003	.072	.052	.072	.072
			.7	.068	.067	.070	.165	.008	.006	.008	.004	.069	.067	.070	.165
		.8	.3	.070	.109	.071	.163	.006	.006	.005	.005	.070	.109	.071	.163
			.7	.069	.080	.070	.134	.008	.006	.008	.005	.069	.080	.070	.134
	40	.2	.3	.032	.040	.033	.070	.005	.004	.005	.003	.033	.040	.033	.070
			.7	.033	.031	.034	.081	.002	.002	.002	.003	.033	.031	.034	.081
		.5	.3	.032	.026	.033	.064	.004	.003	.003	.003	.032	.026	.033	.064
			.7	.034	.036	.034	.084	.004	.003	.003	.003	.034	.036	.035	.084
		.8	.3	.034	.038	.034	.086	.005	.003	.005	.003	.034	.038	.035	.086
			.7	.033	.041	.033	.082	.004	.003	.004	.003	.033	.041	.033	.082

Table A15. *Bias, SE, and RMSE of* σ_θ^2 *in simulation study 1.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.004	-.002	.004	.004	.066	.066	.067	.067	.067	.066	.067	.067
			.7	-.001	-.005	.000	-.002	.061	.062	.063	.063	.061	.062	.063	.063
		.5	.3	.004	.001	.005	.004	.050	.051	.051	.051	.051	.051	.052	.051
			.7	.016	.011	.017	.016	.055	.054	.055	.054	.057	.055	.057	.056
		.8	.3	.011	.007	.010	.013	.059	.057	.058	.057	.060	.057	.059	.059
			.7	.002	-.001	.001	.006	.055	.054	.054	.054	.055	.054	.054	.054
	40	.2	.3	.000	-.003	.001	.000	.044	.044	.044	.045	.044	.045	.044	.045
			.7	.017	.013	.018	.018	.040	.040	.040	.040	.043	.042	.044	.044
		.5	.3	-.008	-.013	-.008	-.008	.036	.036	.037	.036	.037	.038	.037	.037
			.7	-.002	-.004	-.002	-.002	.041	.042	.041	.041	.041	.042	.041	.041
		.8	.3	.002	.000	.001	.002	.049	.048	.048	.048	.049	.048	.048	.048
			.7	.002	-.001	.003	.002	.043	.043	.044	.043	.044	.043	.044	.043
1000	20	.2	.3	-.006	-.007	-.005	-.006	.047	.047	.047	.047	.047	.048	.047	.048
			.7	.004	.000	.004	.003	.048	.048	.048	.048	.048	.048	.048	.048
		.5	.3	.009	.009	.009	.009	.042	.042	.042	.041	.043	.043	.043	.042
			.7	.005	-.001	.005	.005	.045	.045	.045	.046	.046	.045	.045	.046
		.8	.3	.016	.012	.016	.018	.032	.032	.032	.031	.036	.034	.036	.035
			.7	.008	.005	.008	.010	.047	.047	.048	.048	.048	.047	.048	.049
	40	.2	.3	-.007	-.007	-.007	-.007	.022	.022	.022	.022	.023	.023	.023	.023
			.7	.002	.001	.003	.003	.028	.028	.028	.028	.028	.028	.028	.028
		.5	.3	-.002	-.004	-.002	-.003	.021	.021	.021	.021	.021	.021	.021	.021
			.7	-.007	-.011	-.007	-.007	.035	.035	.035	.035	.036	.036	.035	.036
		.8	.3	.002	.000	.003	.002	.031	.031	.030	.030	.031	.031	.031	.030
			.7	-.001	-.004	-.001	-.002	.023	.023	.023	.023	.023	.023	.023	.023

Table A16. *Bias, SE, and RMSE of $\rho_{\theta\tau}$ in simulation study 1.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.010	.010	.009	.128	.028	.028	.028	.026	.030	.030	.030	.131
			.7	-.010	-.009	-.010	.113	.028	.028	.028	.026	.029	.029	.030	.116
		.5	.3	-.002	-.003	-.002	.098	.015	.015	.015	.013	.015	.015	.015	.099
			.7	.004	.004	.004	.106	.020	.020	.020	.018	.020	.020	.020	.107
		.8	.3	-.010	-.011	-.011	.049	.014	.014	.014	.011	.017	.018	.017	.050
			.7	-.008	-.008	-.008	.050	.011	.011	.011	.008	.014	.013	.014	.050
	40	.2	.3	-.003	-.003	-.002	.109	.020	.020	.020	.020	.021	.020	.021	.111
			.7	-.004	-.004	-.004	.112	.017	.017	.017	.015	.017	.017	.017	.113
		.5	.3	-.001	-.001	-.001	.084	.014	.015	.014	.013	.014	.015	.014	.085
			.7	-.002	-.002	-.002	.086	.019	.019	.019	.016	.019	.019	.019	.088
		.8	.3	-.004	-.004	-.004	.042	.012	.012	.012	.010	.012	.012	.012	.043
			.7	-.010	-.010	-.010	.037	.011	.012	.012	.009	.015	.015	.015	.038
1000	20	.2	.3	.005	.005	.005	.134	.021	.021	.020	.019	.021	.021	.021	.135
			.7	.004	.004	.004	.133	.016	.016	.016	.015	.017	.017	.017	.134
		.5	.3	-.007	-.006	-.007	.093	.014	.015	.015	.012	.016	.016	.016	.094
			.7	-.003	-.003	-.002	.098	.015	.014	.015	.012	.015	.015	.015	.099
		.8	.3	-.003	-.003	-.003	.054	.014	.014	.014	.011	.014	.014	.014	.055
			.7	-.005	-.005	-.005	.055	.012	.012	.012	.009	.013	.013	.013	.056
	40	.2	.3	-.003	-.003	-.003	.107	.011	.011	.011	.010	.012	.012	.012	.107
			.7	-.007	-.007	-.007	.104	.010	.011	.010	.010	.013	.013	.013	.105
		.5	.3	.000	.000	.000	.085	.010	.011	.010	.010	.010	.011	.010	.086
			.7	-.002	-.002	-.002	.087	.011	.010	.010	.009	.011	.010	.010	.087
		.8	.3	-.002	-.002	-.002	.043	.007	.007	.007	.006	.007	.007	.007	.043
			.7	-.004	-.004	-.004	.044	.006	.006	.006	.005	.007	.007	.007	.045

Table A17. *Bias, SE, and RMSE of σ_τ^2 in simulation study 1.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM	JM-RD1	JM-RD2	JM-R	HM
500	20	.2	.3	.002	.002	.003	.015	.005	.005	.005	.006	.006	.006	.006	.016
			.7	.003	.003	.003	.015	.007	.006	.007	.007	.007	.007	.007	.017
		.5	.3	.002	.002	.002	.032	.005	.005	.005	.005	.006	.006	.006	.032
			.7	.002	.002	.002	.033	.006	.006	.006	.007	.006	.006	.006	.034
		.8	.3	.003	.002	.003	.050	.005	.005	.005	.006	.006	.006	.006	.051
			.7	.003	.003	.004	.049	.005	.005	.005	.005	.006	.006	.006	.049
	40	.2	.3	.003	.003	.003	.016	.003	.003	.003	.004	.004	.004	.004	.017
			.7	.002	.002	.002	.016	.003	.004	.003	.004	.004	.004	.004	.016
		.5	.3	.003	.003	.003	.031	.003	.004	.003	.004	.004	.004	.004	.031
			.7	.003	.002	.003	.032	.004	.004	.004	.004	.005	.005	.005	.033
		.8	.3	.003	.003	.003	.047	.003	.004	.003	.004	.005	.005	.005	.047
			.7	.002	.002	.002	.047	.003	.003	.003	.004	.004	.004	.004	.047
1000	20	.2	.3	.001	.001	.001	.015	.003	.003	.003	.004	.003	.003	.003	.016
			.7	.002	.002	.002	.016	.004	.004	.004	.004	.005	.005	.005	.017
		.5	.3	.002	.002	.001	.030	.004	.004	.004	.004	.004	.004	.004	.031
			.7	.002	.002	.002	.031	.003	.003	.003	.004	.004	.004	.004	.031
		.8	.3	.001	.001	.001	.045	.005	.005	.005	.005	.005	.005	.005	.045
			.7	.001	.001	.001	.047	.005	.005	.005	.005	.005	.005	.005	.047
	40	.2	.3	.001	.001	.001	.014	.003	.003	.003	.003	.003	.003	.003	.014
			.7	.001	.001	.001	.013	.003	.003	.003	.003	.003	.003	.003	.014
		.5	.3	.001	.001	.001	.030	.002	.002	.002	.003	.003	.003	.003	.030
			.7	.002	.002	.002	.031	.002	.002	.002	.002	.002	.003	.002	.031
		.8	.3	.001	.001	.001	.043	.002	.002	.002	.003	.002	.002	.002	.044
			.7	.002	.002	.002	.048	.003	.003	.003	.003	.003	.003	.003	.048

Appendix B

Table B1a. *Mean and SD of bias in item difficulty estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.008	.007	.013	.004	.020	.030	.018	.020
			.7	-.003	-.005	.005	.003	.025	.043	.024	.028
		.5	.3	-.003	-.003	.001	-.005	.021	.044	.017	.018
			.7	-.007	-.008	.002	-.004	.027	.043	.030	.030
		.8	.3	.000	.010	.002	.001	.026	.052	.027	.030
			.7	.005	.005	.010	.006	.017	.041	.015	.022
	40	.2	.3	-.002	-.002	.002	.002	.019	.064	.020	.022
			.7	.000	.000	.004	-.002	.025	.047	.026	.025
		.5	.3	.000	.001	.004	.001	.024	.070	.024	.023
			.7	.001	-.005	.007	.000	.021	.041	.021	.025
		.8	.3	.006	.012	.009	.008	.019	.055	.019	.020
			.7	.003	.003	.009	.005	.020	.045	.019	.021
1000	20	.2	.3	-.008	-.007	-.006	-.005	.012	.044	.011	.012
			.7	.001	.003	.006	.004	.015	.025	.014	.015
		.5	.3	-.004	-.002	-.005	-.004	.018	.042	.017	.019
			.7	.003	.005	.007	.002	.016	.072	.014	.015
		.8	.3	-.006	-.002	-.005	-.004	.016	.045	.017	.017
			.7	.003	.003	.006	.006	.012	.041	.012	.013
	40	.2	.3	-.001	.003	-.001	-.001	.015	.053	.015	.013
			.7	.002	.004	.005	.003	.013	.045	.013	.012
		.5	.3	-.001	-.002	.001	.000	.014	.045	.014	.015
			.7	-.003	-.002	.001	-.002	.016	.041	.016	.017
		.8	.3	.001	.000	.003	.003	.013	.051	.013	.015
			.7	-.004	-.007	-.001	-.003	.017	.053	.017	.017

Table B1b. *Mean and SD of SE in item difficulty estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.093	.095	.094	.103	.020	.022	.020	.020
			.7	.093	.093	.093	.099	.021	.023	.021	.017
		.5	.3	.095	.100	.094	.106	.021	.027	.021	.017
			.7	.085	.086	.086	.097	.019	.022	.019	.016
		.8	.3	.089	.087	.089	.102	.021	.018	.020	.013
			.7	.093	.093	.093	.102	.020	.019	.021	.018
	40	.2	.3	.088	.088	.088	.107	.027	.026	.027	.017
			.7	.085	.086	.086	.105	.022	.024	.022	.018
		.5	.3	.086	.091	.086	.104	.028	.036	.028	.017
			.7	.086	.089	.087	.104	.022	.031	.022	.017
		.8	.3	.084	.087	.084	.101	.020	.025	.021	.016
			.7	.085	.087	.085	.102	.023	.026	.023	.018
1000	20	.2	.3	.069	.074	.069	.077	.019	.019	.019	.016
			.7	.064	.065	.064	.071	.012	.012	.013	.009
		.5	.3	.063	.066	.063	.070	.015	.014	.015	.013
			.7	.060	.059	.061	.069	.015	.014	.015	.012
		.8	.3	.062	.069	.062	.070	.021	.031	.020	.019
			.7	.063	.064	.063	.072	.013	.013	.013	.010
	40	.2	.3	.061	.066	.061	.072	.018	.023	.017	.012
			.7	.060	.060	.060	.070	.015	.016	.015	.012
		.5	.3	.061	.063	.061	.073	.015	.017	.016	.011
			.7	.059	.059	.059	.071	.013	.014	.013	.011
		.8	.3	.059	.063	.059	.072	.019	.017	.019	.015
			.7	.061	.062	.061	.074	.017	.019	.017	.013

Table B1c. *Mean and SD of RMSE in item difficulty estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.095	.099	.096	.105	.021	.024	.021	.020
			.7	.096	.101	.096	.103	.022	.027	.021	.017
		.5	.3	.097	.109	.096	.108	.020	.029	.020	.017
			.7	.089	.095	.090	.101	.020	.025	.022	.017
		.8	.3	.092	.098	.093	.106	.022	.032	.022	.017
			.7	.095	.100	.095	.105	.020	.024	.020	.018
	40	.2	.3	.090	.106	.090	.109	.027	.036	.027	.017
			.7	.088	.096	.089	.108	.024	.029	.024	.021
		.5	.3	.090	.111	.090	.106	.027	.046	.027	.017
			.7	.089	.097	.089	.107	.022	.033	.022	.018
		.8	.3	.087	.102	.087	.104	.020	.031	.021	.016
			.7	.087	.097	.088	.104	.023	.028	.023	.018
1000	20	.2	.3	.071	.085	.070	.078	.018	.024	.018	.016
			.7	.066	.070	.066	.072	.013	.013	.013	.009
		.5	.3	.065	.076	.065	.072	.015	.021	.015	.014
			.7	.062	.084	.063	.071	.016	.040	.016	.013
		.8	.3	.064	.082	.064	.072	.022	.032	.022	.021
			.7	.064	.074	.064	.073	.013	.018	.014	.010
	40	.2	.3	.063	.081	.063	.073	.017	.032	.017	.012
			.7	.061	.071	.061	.071	.016	.028	.016	.012
		.5	.3	.063	.075	.063	.074	.016	.024	.016	.012
			.7	.061	.070	.061	.073	.014	.024	.014	.011
		.8	.3	.061	.078	.061	.073	.019	.027	.019	.015
			.7	.064	.077	.063	.075	.018	.034	.018	.014

Table B2a. *Mean and SD of bias in time discrimination estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	-.011	-.026	-.011	-.052	.011	.026	.011	.042
			.7	-.010	-.021	-.011	-.061	.013	.036	.014	.053
		.5	.3	-.016	-.034	-.016	-.067	.012	.017	.012	.029
			.7	-.011	-.019	-.011	-.059	.012	.024	.011	.033
		.8	.3	-.013	-.026	-.013	-.071	.009	.042	.009	.060
			.7	-.006	-.017	-.007	-.052	.014	.031	.014	.043
	40	.2	.3	-.013	-.032	-.013	-.070	.013	.032	.013	.051
			.7	-.011	-.023	-.011	-.055	.013	.025	.013	.034
		.5	.3	-.011	-.033	-.011	-.065	.011	.044	.011	.062
			.7	-.009	-.018	-.009	-.057	.013	.023	.013	.042
		.8	.3	-.011	-.028	-.011	-.059	.011	.024	.011	.037
			.7	-.012	-.023	-.012	-.057	.013	.029	.013	.041
1000	20	.2	.3	-.007	-.023	-.007	-.060	.012	.040	.012	.049
			.7	-.004	-.015	-.004	-.051	.010	.025	.010	.037
		.5	.3	-.006	-.030	-.006	-.058	.008	.046	.008	.060
			.7	-.002	-.012	-.003	-.052	.011	.023	.011	.038
		.8	.3	-.006	-.028	-.006	-.053	.009	.036	.009	.045
			.7	-.009	-.020	-.009	-.056	.007	.021	.007	.037
	40	.2	.3	-.004	-.027	-.004	-.055	.010	.035	.010	.045
			.7	-.006	-.016	-.006	-.047	.007	.021	.007	.030
		.5	.3	-.006	-.022	-.006	-.051	.009	.038	.009	.051
			.7	-.006	-.015	-.006	-.052	.010	.018	.010	.033
		.8	.3	-.007	-.030	-.007	-.059	.011	.044	.011	.070
			.7	-.006	-.016	-.006	-.054	.011	.019	.011	.033

Table B2b. *Mean and SD of SE in time discrimination estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.064	.063	.064	.061	.009	.009	.009	.009
			.7	.067	.067	.067	.063	.009	.009	.009	.009
		.5	.3	.062	.062	.062	.061	.007	.008	.007	.008
			.7	.066	.064	.066	.062	.007	.007	.007	.008
		.8	.3	.064	.063	.064	.060	.008	.007	.008	.006
			.7	.066	.065	.066	.062	.011	.010	.011	.010
	40	.2	.3	.063	.062	.063	.060	.007	.008	.007	.007
			.7	.061	.060	.061	.058	.009	.009	.009	.009
		.5	.3	.064	.064	.064	.062	.008	.008	.008	.008
			.7	.064	.063	.064	.061	.007	.006	.007	.006
		.8	.3	.063	.062	.063	.060	.009	.008	.009	.007
			.7	.062	.062	.062	.060	.008	.009	.008	.007
1000	20	.2	.3	.046	.045	.047	.043	.005	.005	.005	.006
			.7	.046	.046	.046	.044	.007	.006	.007	.006
		.5	.3	.047	.047	.047	.045	.005	.006	.005	.006
			.7	.046	.046	.046	.044	.006	.006	.006	.006
		.8	.3	.047	.046	.047	.045	.008	.006	.008	.005
			.7	.046	.046	.046	.044	.005	.006	.005	.006
	40	.2	.3	.046	.045	.046	.043	.006	.007	.006	.007
			.7	.044	.044	.044	.043	.006	.006	.006	.006
		.5	.3	.045	.045	.045	.043	.006	.006	.006	.005
			.7	.043	.043	.043	.041	.006	.006	.006	.006
		.8	.3	.044	.043	.044	.042	.005	.005	.005	.005
			.7	.045	.045	.045	.043	.006	.006	.006	.005

Table B2c. *Mean and SD of RMSE in time discrimination estimation in simulation study 2.*

J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	Mean				SD			
				JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.066	.072	.066	.085	.009	.018	.009	.031
			.7	.069	.077	.070	.094	.009	.019	.009	.041
		.5	.3	.065	.073	.065	.093	.008	.011	.008	.021
			.7	.068	.071	.068	.089	.007	.011	.007	.023
		.8	.3	.066	.074	.066	.098	.008	.031	.008	.051
			.7	.068	.074	.068	.087	.010	.014	.011	.029
	40	.2	.3	.065	.074	.065	.097	.008	.019	.008	.041
			.7	.064	.068	.064	.085	.008	.014	.008	.022
		.5	.3	.066	.079	.066	.096	.008	.030	.008	.052
			.7	.066	.070	.066	.088	.007	.011	.007	.031
		.8	.3	.065	.071	.065	.089	.008	.012	.009	.026
			.7	.065	.069	.065	.086	.009	.021	.009	.034
1000	20	.2	.3	.048	.057	.048	.078	.006	.030	.006	.043
			.7	.047	.053	.047	.071	.007	.013	.007	.029
		.5	.3	.048	.064	.049	.080	.005	.033	.005	.051
			.7	.048	.051	.048	.072	.006	.013	.006	.030
		.8	.3	.048	.061	.048	.074	.008	.022	.008	.037
			.7	.048	.053	.048	.075	.005	.010	.005	.029
	40	.2	.3	.047	.058	.047	.074	.006	.023	.006	.038
			.7	.045	.050	.045	.066	.006	.013	.006	.025
		.5	.3	.046	.056	.046	.072	.006	.028	.006	.044
			.7	.045	.048	.045	.070	.006	.010	.006	.026
		.8	.3	.045	.059	.045	.079	.005	.035	.005	.061
			.7	.047	.051	.047	.072	.005	.010	.005	.027

Table B3a. *Mean and SD of bias in time intensity estimation in simulation study 2*.

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	-.014	-.023	-.022	-.187	.056	.130	.076	.222
			.7	.003	.020	-.013	-.189	.017	.083	.049	.176
		.5	.3	.013	.013	.000	-.168	.084	.123	.043	.158
			.7	.002	.009	-.011	-.200	.022	.074	.026	.175
		.8	.3	.009	.023	.005	-.165	.026	.071	.016	.105
			.7	.000	.002	-.013	-.181	.023	.063	.049	.162
	40	.2	.3	.012	.033	.000	-.174	.036	.117	.016	.154
			.7	-.003	-.003	-.012	-.190	.018	.078	.035	.178
		.5	.3	-.005	.012	-.017	-.180	.019	.099	.040	.173
			.7	.004	.024	-.007	-.182	.024	.101	.026	.153
		.8	.3	.003	.008	-.005	-.162	.024	.093	.020	.155
			.7	-.004	-.002	-.014	-.194	.027	.089	.041	.189
1000	20	.2	.3	.003	.021	.001	-.176	.012	.104	.010	.132
			.7	.001	-.002	-.004	-.195	.008	.068	.016	.165
		.5	.3	-.019	-.015	-.009	-.142	.078	.101	.036	.118
			.7	.004	.006	-.008	-.203	.027	.062	.024	.203
		.8	.3	.000	-.014	-.012	-.157	.024	.119	.038	.175
			.7	.001	.006	-.003	-.188	.014	.092	.013	.167
	40	.2	.3	.003	.008	-.001	-.153	.020	.093	.020	.139
			.7	-.004	.001	-.015	-.174	.023	.067	.047	.159
		.5	.3	-.002	.000	-.006	-.167	.017	.122	.029	.172
			.7	-.001	.010	-.010	-.186	.018	.084	.047	.163
		.8	.3	-.002	.007	-.006	-.174	.020	.126	.026	.179
			.7	-.001	.013	-.005	-.192	.012	.093	.018	.163

Table B3b. *Mean and SD of SE in time intensity estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.046	.035	.046	.022	.028	.014	.028	.003
			.7	.043	.035	.042	.021	.024	.016	.021	.002
		.5	.3	.043	.035	.042	.022	.019	.016	.016	.003
			.7	.045	.034	.046	.021	.026	.012	.025	.003
		.8	.3	.044	.035	.042	.022	.029	.018	.023	.003
			.7	.041	.031	.040	.021	.025	.013	.021	.003
	40	.2	.3	.043	.032	.042	.022	.026	.012	.024	.003
			.7	.039	.028	.040	.021	.018	.008	.019	.003
		.5	.3	.044	.033	.043	.022	.029	.013	.027	.003
			.7	.041	.029	.040	.022	.022	.008	.019	.003
		.8	.3	.041	.029	.040	.022	.021	.009	.019	.003
			.7	.037	.028	.038	.021	.018	.008	.018	.003
1000	20	.2	.3	.031	.024	.031	.016	.013	.008	.012	.002
			.7	.031	.023	.031	.015	.016	.007	.016	.002
		.5	.3	.035	.024	.030	.016	.033	.013	.014	.002
			.7	.030	.023	.033	.016	.015	.010	.023	.002
		.8	.3	.030	.027	.028	.016	.021	.025	.015	.002
			.7	.032	.022	.031	.015	.019	.006	.019	.002
	40	.2	.3	.030	.023	.029	.015	.020	.012	.017	.002
			.7	.029	.020	.029	.015	.017	.007	.017	.002
		.5	.3	.029	.020	.029	.015	.018	.006	.017	.002
			.7	.030	.021	.030	.015	.014	.006	.015	.002
		.8	.3	.030	.021	.030	.015	.019	.005	.019	.002
			.7	.029	.020	.029	.015	.013	.005	.013	.002

Table B3c. *Mean and SD of RMSE in time intensity estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.055	.092	.060	.193	.056	.100	.074	.218
			.7	.046	.075	.052	.194	.025	.053	.045	.170
		.5	.3	.059	.092	.050	.174	.076	.090	.036	.153
			.7	.048	.062	.050	.206	.029	.054	.031	.169
		.8	.3	.047	.068	.043	.170	.036	.047	.025	.099
			.7	.045	.061	.049	.188	.028	.035	.046	.156
	40	.2	.3	.048	.091	.044	.181	.040	.087	.026	.147
			.7	.042	.065	.047	.195	.020	.053	.035	.174
		.5	.3	.047	.080	.051	.188	.032	.068	.043	.165
			.7	.045	.076	.045	.186	.027	.076	.027	.150
		.8	.3	.044	.076	.043	.168	.027	.061	.023	.150
			.7	.042	.071	.047	.200	.026	.059	.037	.184
1000	20	.2	.3	.033	.084	.032	.181	.014	.067	.012	.126
			.7	.032	.055	.034	.198	.016	.045	.019	.161
		.5	.3	.047	.085	.036	.146	.081	.060	.034	.114
			.7	.034	.053	.036	.208	.026	.040	.031	.198
		.8	.3	.033	.083	.036	.162	.029	.092	.036	.171
			.7	.034	.067	.033	.191	.021	.066	.020	.163
	40	.2	.3	.032	.075	.032	.158	.025	.061	.023	.134
			.7	.032	.054	.038	.177	.025	.043	.046	.156
		.5	.3	.031	.080	.033	.170	.022	.093	.029	.170
			.7	.032	.064	.037	.189	.019	.059	.045	.159
		.8	.3	.033	.093	.034	.177	.024	.086	.029	.177
			.7	.030	.060	.032	.195	.015	.075	.019	.160

Table B4a. *Mean and SD of bias in ability estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.000	.000	.000	.000	.205	.211	.206	.239
			.7	.000	.000	.000	.000	.213	.226	.211	.236
		.5	.3	.000	.000	.000	.000	.224	.232	.223	.263
			.7	.000	.000	.000	.000	.214	.221	.216	.251
		.8	.3	.000	.000	.000	.000	.232	.251	.232	.260
			.7	.000	.000	.000	.000	.223	.229	.223	.255
	40	.2	.3	.000	.000	.000	.000	.122	.135	.121	.143
			.7	.000	.000	.000	.000	.113	.117	.113	.137
		.5	.3	.000	.000	.000	.000	.148	.152	.148	.169
			.7	.000	.000	.000	.000	.151	.162	.151	.174
		.8	.3	.000	.000	.000	.000	.139	.152	.139	.176
			.7	.000	.000	.000	.000	.146	.152	.147	.176
1000	20	.2	.3	.000	.000	.000	.000	.215	.237	.215	.244
			.7	.000	.000	.000	.000	.218	.224	.219	.250
		.5	.3	.000	.000	.000	.000	.216	.232	.215	.251
			.7	.000	.000	.000	.000	.213	.219	.213	.249
		.8	.3	.000	.000	.000	.000	.221	.233	.221	.254
			.7	.000	.000	.000	.000	.214	.223	.215	.256
	40	.2	.3	.000	.000	.000	.000	.128	.135	.128	.151
			.7	.000	.000	.000	.000	.133	.137	.133	.151
		.5	.3	.000	.000	.000	.000	.133	.141	.133	.162
			.7	.000	.000	.000	.000	.130	.142	.130	.162
		.8	.3	.000	.000	.000	.000	.147	.156	.147	.179
			.7	.000	.000	.000	.000	.148	.159	.148	.183

Table B4b. *Mean and SD of SE in ability estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.364	.374	.364	.390	.057	.056	.056	.054
			.7	.359	.362	.361	.392	.056	.056	.056	.052
		.5	.3	.332	.341	.332	.357	.050	.050	.050	.048
			.7	.333	.336	.333	.357	.052	.053	.052	.045
		.8	.3	.261	.257	.261	.278	.040	.039	.040	.036
			.7	.274	.276	.275	.285	.041	.041	.041	.039
	40	.2	.3	.261	.273	.262	.319	.050	.049	.050	.042
			.7	.273	.281	.273	.315	.052	.050	.052	.043
		.5	.3	.252	.267	.253	.298	.050	.047	.051	.042
			.7	.258	.261	.258	.296	.050	.050	.050	.040
		.8	.3	.220	.228	.221	.250	.037	.037	.037	.034
			.7	.224	.231	.225	.253	.039	.039	.039	.034
1000	20	.2	.3	.358	.363	.358	.390	.059	.058	.059	.053
			.7	.355	.362	.355	.385	.057	.057	.057	.054
		.5	.3	.330	.341	.330	.358	.052	.051	.052	.048
			.7	.332	.336	.332	.356	.053	.053	.053	.048
		.8	.3	.265	.274	.266	.278	.042	.042	.042	.039
			.7	.271	.273	.271	.281	.041	.041	.041	.037
	40	.2	.3	.264	.282	.264	.311	.049	.047	.049	.043
			.7	.275	.283	.275	.314	.049	.049	.050	.043
		.5	.3	.255	.268	.255	.298	.046	.046	.046	.041
			.7	.258	.261	.258	.295	.049	.049	.049	.040
		.8	.3	.221	.232	.221	.251	.042	.041	.042	.033
			.7	.221	.224	.221	.247	.041	.042	.041	.034

Table B4c. *Mean and SD of RMSE in ability estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.410	.422	.410	.450	.096	.097	.096	.100
			.7	.408	.417	.409	.450	.104	.108	.103	.097
		.5	.3	.388	.400	.388	.430	.112	.112	.111	.115
			.7	.385	.391	.386	.424	.105	.107	.105	.111
		.8	.3	.332	.338	.332	.365	.115	.127	.114	.115
			.7	.339	.344	.339	.367	.107	.108	.107	.114
	40	.2	.3	.282	.297	.282	.345	.079	.082	.077	.069
			.7	.292	.301	.292	.340	.067	.066	.067	.062
		.5	.3	.284	.298	.285	.335	.086	.086	.086	.083
			.7	.286	.293	.287	.333	.100	.104	.099	.093
		.8	.3	.254	.266	.255	.298	.068	.073	.068	.073
			.7	.261	.269	.262	.301	.072	.073	.072	.073
1000	20	.2	.3	.405	.419	.405	.450	.118	.125	.117	.111
			.7	.406	.415	.406	.449	.111	.111	.111	.111
		.5	.3	.383	.401	.383	.428	.106	.108	.106	.104
			.7	.385	.392	.386	.425	.098	.099	.098	.099
		.8	.3	.332	.346	.333	.362	.103	.106	.103	.111
			.7	.334	.340	.334	.367	.097	.101	.097	.106
	40	.2	.3	.288	.307	.288	.342	.075	.076	.075	.071
			.7	.299	.307	.299	.342	.080	.081	.081	.076
		.5	.3	.283	.298	.283	.334	.067	.072	.068	.070
			.7	.283	.290	.283	.331	.075	.080	.075	.071
		.8	.3	.258	.272	.258	.302	.074	.075	.074	.073
			.7	.257	.264	.258	.299	.079	.085	.079	.081

Table B5a. *Mean and SD of bias in speed estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.000	.000	.000	.000	.038	.039	.038	.082
			.7	.000	.000	.000	.000	.039	.040	.039	.093
		.5	.3	.000	.000	.000	.000	.045	.045	.045	.093
			.7	.000	.000	.000	.000	.041	.041	.041	.095
		.8	.3	.000	.000	.000	.000	.046	.049	.046	.098
			.7	.000	.000	.000	.000	.047	.047	.047	.089
	40	.2	.3	.000	.000	.000	.000	.023	.023	.023	.092
			.7	.000	.000	.000	.000	.024	.024	.024	.094
		.5	.3	.000	.000	.000	.000	.025	.026	.025	.090
			.7	.000	.000	.000	.000	.024	.025	.024	.090
		.8	.3	.000	.000	.000	.000	.029	.030	.029	.083
			.7	.000	.000	.000	.000	.030	.030	.030	.087
1000	20	.2	.3	.000	.000	.000	.000	.040	.042	.040	.102
			.7	.000	.000	.000	.000	.041	.041	.041	.097
		.5	.3	.000	.000	.000	.000	.042	.044	.042	.091
			.7	.000	.000	.000	.000	.044	.044	.044	.095
		.8	.3	.000	.000	.000	.000	.047	.049	.047	.088
			.7	.000	.000	.000	.000	.046	.047	.046	.090
	40	.2	.3	.000	.000	.000	.000	.022	.023	.022	.088
			.7	.000	.000	.000	.000	.024	.025	.024	.091
		.5	.3	.000	.000	.000	.000	.024	.025	.024	.085
			.7	.000	.000	.000	.000	.025	.027	.025	.097
		.8	.3	.000	.000	.000	.000	.027	.029	.027	.082
			.7	.000	.000	.000	.000	.029	.029	.029	.092

Table B5b. *Mean and SD of SE in speed estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.106	.106	.106	.104	.014	.014	.014	.014
			.7	.106	.106	.106	.103	.015	.015	.015	.014
		.5	.3	.103	.103	.103	.103	.014	.014	.014	.014
			.7	.105	.105	.105	.104	.014	.014	.014	.013
		.8	.3	.095	.094	.095	.100	.013	.013	.013	.013
			.7	.095	.095	.095	.101	.013	.013	.013	.013
	40	.2	.3	.078	.078	.078	.075	.011	.011	.011	.010
			.7	.079	.079	.079	.075	.011	.011	.011	.010
		.5	.3	.077	.077	.077	.075	.010	.010	.010	.009
			.7	.077	.077	.077	.075	.011	.011	.011	.010
		.8	.3	.073	.073	.073	.074	.009	.009	.009	.010
			.7	.071	.071	.071	.073	.009	.009	.009	.009
1000	20	.2	.3	.108	.108	.108	.104	.014	.014	.014	.013
			.7	.107	.107	.107	.104	.014	.014	.014	.014
		.5	.3	.103	.103	.103	.103	.014	.014	.014	.014
			.7	.103	.103	.103	.103	.014	.014	.014	.014
		.8	.3	.095	.095	.095	.101	.013	.013	.013	.013
			.7	.095	.095	.095	.101	.013	.013	.013	.014
	40	.2	.3	.078	.079	.078	.075	.010	.010	.010	.010
			.7	.078	.078	.078	.075	.011	.011	.011	.010
		.5	.3	.077	.077	.077	.075	.010	.010	.010	.010
			.7	.077	.077	.077	.074	.010	.011	.010	.010
		.8	.3	.072	.072	.072	.074	.009	.009	.009	.009
			.7	.073	.072	.073	.074	.009	.009	.009	.009

Table B5c. *Mean and SD of RMSE in speed estimation in simulation study 2.*

				Mean				SD			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.113	.113	.113	.129	.016	.017	.016	.031
			.7	.112	.113	.112	.135	.017	.017	.017	.031
		.5	.3	.112	.112	.112	.135	.018	.018	.018	.032
			.7	.112	.112	.112	.137	.016	.016	.016	.035
		.8	.3	.105	.105	.105	.138	.016	.016	.016	.031
			.7	.105	.105	.105	.132	.018	.018	.018	.028
	40	.2	.3	.081	.081	.081	.114	.012	.012	.012	.032
			.7	.082	.083	.082	.115	.011	.011	.011	.036
		.5	.3	.081	.081	.081	.114	.011	.011	.011	.028
			.7	.080	.081	.080	.112	.012	.012	.012	.034
		.8	.3	.078	.078	.078	.107	.011	.011	.011	.032
			.7	.077	.077	.077	.109	.011	.012	.012	.034
1000	20	.2	.3	.115	.116	.115	.142	.017	.017	.017	.035
			.7	.114	.115	.114	.138	.017	.017	.017	.034
		.5	.3	.111	.112	.111	.135	.017	.017	.017	.030
			.7	.112	.112	.112	.137	.018	.018	.018	.033
		.8	.3	.106	.106	.106	.131	.017	.018	.017	.029
			.7	.105	.105	.105	.132	.018	.018	.018	.033
	40	.2	.3	.081	.082	.081	.113	.011	.011	.011	.028
			.7	.082	.082	.082	.115	.012	.012	.012	.027
		.5	.3	.080	.081	.080	.109	.011	.011	.011	.031
			.7	.081	.081	.081	.118	.011	.012	.011	.032
		.8	.3	.077	.078	.077	.106	.011	.011	.011	.029
			.7	.078	.078	.078	.113	.012	.012	.012	.036

Table B6. *Bias, SE, and RMSE of ω_0 in simulation study 2.*

				Bias			SE			RMSE		
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	JM-DD1	JM-DD2	JM-D	JM-DD1	JM-DD2	JM-D
500	20	.2	.3	.024	.041	.018	.036	.036	.036	.043	.054	.040
			.7	.002	-.019	.000	.024	.025	.025	.024	.032	.025
		.5	.3	-.004	-.002	-.002	.034	.029	.033	.034	.029	.033
			.7	-.007	-.015	-.016	.029	.027	.031	.030	.031	.035
		.8	.3	-.021	-.080	-.020	.025	.021	.023	.033	.083	.031
			.7	-.009	-.009	-.005	.027	.024	.027	.028	.026	.027
	40	.2	.3	-.014	-.032	-.012	.016	.016	.016	.022	.036	.020
			.7	.012	.011	-.002	.020	.020	.019	.024	.023	.019
		.5	.3	.005	-.010	.011	.018	.012	.017	.019	.016	.020
			.7	-.005	-.029	-.007	.020	.019	.019	.021	.034	.021
		.8	.3	-.004	-.017	.001	.020	.018	.020	.021	.025	.020
			.7	.004	.003	-.003	.014	.013	.014	.014	.014	.014
1000	20	.2	.3	.009	-.037	-.003	.018	.021	.018	.020	.043	.018
			.7	-.008	-.015	-.015	.019	.021	.019	.020	.026	.024
		.5	.3	.018	.008	.009	.019	.016	.017	.026	.018	.019
			.7	-.004	.003	-.014	.017	.019	.019	.018	.019	.023
		.8	.3	.000	.013	.011	.017	.017	.017	.017	.021	.020
			.7	-.006	-.013	-.014	.018	.016	.018	.019	.021	.023
	40	.2	.3	-.003	-.003	-.001	.010	.013	.011	.011	.013	.011
			.7	.003	-.005	.009	.012	.010	.012	.012	.011	.015
		.5	.3	.000	-.002	.001	.013	.012	.014	.013	.012	.014
			.7	-.004	-.030	-.006	.015	.014	.014	.015	.033	.016
		.8	.3	.005	.014	.005	.014	.013	.015	.015	.019	.016
			.7	-.003	-.016	-.012	.013	.012	.012	.013	.020	.017

Table B7. *Bias, SE, and RMSE of ω_1 in simulation study 2.*

				Bias		SE		RMSE	
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-DD1	JM-DD2	JM-DD1	JM-DD2
500	20	.2	.3	.072	.060	.034	.030	.080	.067
			.7	.017	.029	.033	.037	.037	.047
		.5	.3	.049	.055	.033	.042	.059	.069
			.7	.022	.031	.024	.025	.032	.040
		.8	.3	-.018	-.063	.034	.040	.038	.075
			.7	-.035	-.031	.030	.028	.046	.042
	40	.2	.3	.044	.090	.023	.026	.049	.094
			.7	.023	.015	.018	.019	.030	.024
		.5	.3	.043	.085	.022	.028	.048	.089
			.7	-.034	-.007	.018	.016	.039	.017
		.8	.3	.006	.000	.021	.023	.022	.023
			.7	.002	.009	.020	.020	.020	.022
1000	20	.2	.3	.027	-.003	.028	.037	.039	.037
			.7	-.006	-.032	.021	.024	.022	.040
		.5	.3	-.131	-.116	.039	.027	.136	.119
			.7	.074	.103	.022	.025	.077	.106
		.8	.3	.018	-.020	.025	.036	.031	.041
			.7	-.022	-.022	.024	.019	.032	.029
	40	.2	.3	-.028	-.005	.015	.024	.032	.025
			.7	-.046	-.041	.015	.014	.048	.043
		.5	.3	-.001	.003	.017	.015	.017	.015
			.7	-.018	-.034	.012	.013	.021	.036
		.8	.3	.015	.076	.013	.013	.020	.077
			.7	-.020	.005	.015	.012	.025	.014

Table B8. *Bias, SE, and RMSE of $\rho_{b\lambda}$ in simulation study 2.*

				Bias	SE	RMSE
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD1	JM-DD1
500	20	.2	.3	.097	.100	.139
			.7	-.240	.077	.252
		.5	.3	-.010	.087	.088
			.7	-.176	.102	.203
		.8	.3	-.188	.090	.208
			.7	-.353	.095	.366
	40	.2	.3	-.040	.055	.068
			.7	-.205	.054	.212
		.5	.3	-.043	.048	.064
			.7	-.348	.067	.355
		.8	.3	-.127	.052	.137
			.7	-.257	.057	.263
1000	20	.2	.3	-.071	.068	.098
			.7	-.257	.065	.265
		.5	.3	-.482	.097	.492
			.7	-.090	.049	.102
		.8	.3	-.098	.063	.117
			.7	-.326	.063	.332
	40	.2	.3	-.219	.039	.222
			.7	-.370	.047	.373
		.5	.3	-.140	.045	.147
			.7	-.278	.041	.281
		.8	.3	-.097	.034	.103
			.7	-.316	.052	.320

Table B9. *Bias, SE, and RMSE of σ_{ϕ}^{2} in simulation study 2.*

				Bias		SE		RMSE	
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-D	JM-DD1	JM-D	JM-DD1	JM-D
500	20	.2	.3	-.005	.005	.008	.010	.009	.011
			.7	.009	.005	.006	.011	.011	.012
		.5	.3	.003	.006	.006	.007	.007	.009
			.7	.003	.001	.010	.011	.010	.011
		.8	.3	.008	.003	.010	.010	.013	.010
			.7	.005	-.009	.006	.006	.008	.011
	40	.2	.3	.005	.008	.006	.007	.007	.011
			.7	.003	-.002	.004	.005	.005	.006
		.5	.3	.006	.008	.004	.005	.007	.010
			.7	.002	-.010	.004	.004	.005	.010
		.8	.3	.003	.002	.005	.006	.006	.006
			.7	.003	-.003	.003	.005	.004	.006
1000	20	.2	.3	.007	.006	.005	.007	.009	.009
			.7	.004	-.001	.006	.009	.007	.009
		.5	.3	.009	.004	.004	.004	.010	.005
			.7	.003	.012	.004	.009	.005	.015
		.8	.3	.010	.008	.005	.006	.011	.009
			.7	.007	-.004	.004	.007	.008	.008
	40	.2	.3	.006	.002	.004	.004	.007	.004
			.7	.000	-.015	.002	.003	.002	.015
		.5	.3	.002	.000	.004	.004	.005	.004
			.7	-.001	-.012	.002	.003	.002	.012
		.8	.3	.001	.001	.003	.003	.003	.003
			.7	.003	-.006	.002	.002	.004	.007

Table B10. *Bias, SE, and RMSE of μ_b in simulation study 2.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.007	.006	.012	.005	.024	.024	.023	.025	.025	.025	.026	.025
			.7	-.003	-.005	.005	.003	.024	.025	.024	.025	.025	.026	.024	.025
		.5	.3	-.003	-.003	.001	-.004	.018	.022	.018	.021	.018	.022	.018	.021
			.7	-.007	-.008	.002	-.002	.017	.017	.017	.019	.018	.019	.017	.019
		.8	.3	.000	.010	.002	.002	.017	.016	.017	.019	.017	.019	.017	.019
			.7	.005	.005	.010	.007	.025	.025	.024	.027	.025	.026	.026	.028
	40	.2	.3	-.002	-.002	.003	.002	.013	.014	.013	.017	.013	.014	.013	.017
			.7	.000	.000	.004	-.001	.013	.014	.013	.014	.013	.014	.014	.014
		.5	.3	.000	.000	.004	.001	.015	.015	.014	.016	.015	.015	.015	.016
			.7	.001	-.005	.008	.000	.013	.013	.014	.015	.013	.014	.016	.015
		.8	.3	.007	.012	.009	.008	.013	.013	.013	.016	.015	.017	.016	.018
			.7	.003	.004	.009	.006	.016	.018	.017	.019	.016	.018	.019	.020
1000	20	.2	.3	-.007	-.007	-.005	-.003	.020	.022	.020	.022	.021	.023	.020	.022
			.7	.000	.004	.006	.006	.015	.014	.015	.014	.015	.015	.016	.015
		.5	.3	-.004	-.002	-.004	-.003	.012	.013	.011	.012	.012	.013	.012	.013
			.7	.003	.005	.007	.003	.014	.015	.015	.016	.015	.015	.017	.016
		.8	.3	-.005	-.002	-.005	-.003	.015	.015	.016	.016	.016	.015	.016	.016
			.7	.003	.003	.007	.007	.016	.015	.014	.014	.016	.015	.016	.016
	40	.2	.3	-.001	.002	.000	-.001	.011	.012	.012	.012	.011	.012	.012	.012
			.7	.002	.004	.005	.004	.009	.008	.009	.009	.009	.009	.010	.009
		.5	.3	-.001	-.003	.000	.000	.011	.011	.010	.012	.011	.011	.010	.012
			.7	-.003	-.001	.001	-.002	.010	.011	.010	.010	.010	.011	.010	.010
		.8	.3	.001	.000	.003	.003	.011	.012	.011	.014	.011	.012	.012	.014
			.7	-.004	-.008	.000	-.003	.012	.012	.013	.012	.012	.014	.013	.013

Table B11. *Bias, SE, and RMSE of μ_β in simulation study 2.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	-.014	-.023	-.022	-.185	.021	.020	.021	.006	.025	.031	.030	.185
			.7	.003	.019	-.013	-.187	.017	.017	.016	.005	.017	.025	.021	.187
		.5	.3	.013	.013	.000	-.166	.018	.015	.016	.006	.022	.020	.016	.166
			.7	.002	.008	-.011	-.198	.016	.015	.016	.004	.016	.017	.019	.198
		.8	.3	.009	.023	.005	-.164	.017	.016	.014	.004	.019	.027	.015	.164
			.7	.000	.001	-.012	-.180	.015	.015	.014	.006	.015	.015	.019	.180
	40	.2	.3	.012	.033	.000	-.174	.012	.012	.010	.004	.018	.035	.010	.174
			.7	-.004	-.003	-.012	-.189	.009	.010	.010	.004	.010	.010	.015	.189
		.5	.3	-.005	.012	-.017	-.179	.013	.011	.012	.003	.014	.016	.020	.179
			.7	.004	.024	-.007	-.181	.013	.010	.012	.003	.014	.026	.014	.181
		.8	.3	.003	.007	-.005	-.161	.011	.010	.010	.003	.011	.013	.011	.161
			.7	-.004	-.001	-.014	-.193	.009	.010	.009	.005	.010	.010	.016	.194
1000	20	.2	.3	.002	.021	.001	-.174	.011	.011	.010	.004	.011	.024	.010	.174
			.7	.001	-.002	-.004	-.192	.011	.011	.010	.004	.011	.011	.011	.192
		.5	.3	-.019	-.015	-.008	-.140	.015	.012	.011	.003	.024	.020	.014	.140
			.7	.004	.007	-.008	-.201	.010	.010	.011	.004	.011	.012	.014	.201
		.8	.3	.001	-.013	-.012	-.155	.012	.014	.010	.003	.012	.019	.015	.155
			.7	.001	.006	-.003	-.186	.011	.010	.011	.003	.011	.012	.011	.186
	40	.2	.3	.003	.008	-.001	-.153	.008	.010	.007	.003	.008	.012	.007	.153
			.7	-.004	.001	-.015	-.173	.007	.006	.007	.003	.008	.006	.016	.173
		.5	.3	-.002	.000	-.006	-.166	.007	.007	.007	.002	.007	.007	.009	.166
			.7	-.001	.010	-.010	-.185	.008	.008	.008	.003	.008	.013	.013	.185
		.8	.3	-.002	.007	-.006	-.173	.009	.008	.009	.003	.009	.010	.011	.173
			.7	-.001	.013	-.005	-.191	.006	.007	.007	.003	.007	.014	.008	.191

Table B12. *Bias, SE, and RMSE of σ_b^2 in simulation study 2.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.107	.103	.108	.113	.053	.052	.052	.053	.120	.115	.120	.125
			.7	.107	.117	.103	.097	.068	.068	.063	.068	.126	.135	.121	.118
		.5	.3	.112	.109	.109	.108	.071	.069	.069	.070	.132	.129	.129	.129
			.7	.106	.105	.105	.103	.049	.047	.050	.052	.117	.115	.116	.115
		.8	.3	.103	.140	.103	.096	.058	.053	.056	.055	.118	.150	.117	.110
			.7	.124	.126	.126	.123	.060	.060	.059	.061	.138	.139	.139	.138
	40	.2	.3	.065	.074	.061	.054	.041	.041	.040	.040	.077	.084	.073	.067
			.7	.043	.042	.042	.044	.048	.047	.046	.047	.064	.063	.062	.064
		.5	.3	.061	.063	.059	.058	.045	.042	.043	.046	.075	.075	.073	.074
			.7	.066	.083	.064	.058	.040	.039	.040	.042	.077	.091	.075	.071
		.8	.3	.068	.078	.067	.064	.040	.040	.041	.042	.079	.088	.079	.077
			.7	.053	.054	.052	.050	.040	.038	.039	.044	.066	.066	.065	.066
1000	20	.2	.3	.125	.146	.124	.121	.042	.044	.041	.042	.132	.152	.131	.129
			.7	.106	.106	.107	.105	.032	.034	.032	.033	.111	.111	.111	.110
		.5	.3	.100	.107	.102	.098	.046	.045	.043	.045	.110	.116	.111	.108
			.7	.113	.109	.112	.114	.038	.039	.040	.040	.120	.116	.119	.121
		.8	.3	.103	.096	.102	.098	.058	.057	.059	.057	.118	.111	.118	.113
			.7	.120	.125	.123	.117	.035	.036	.035	.036	.125	.130	.128	.123
	40	.2	.3	.056	.052	.055	.055	.032	.033	.032	.034	.065	.062	.064	.065
			.7	.051	.055	.051	.052	.031	.030	.031	.032	.060	.063	.059	.061
		.5	.3	.042	.047	.044	.043	.025	.024	.026	.026	.049	.053	.051	.051
			.7	.052	.068	.051	.051	.029	.031	.028	.028	.060	.075	.059	.058
		.8	.3	.047	.047	.048	.044	.031	.031	.031	.033	.056	.056	.057	.055
			.7	.044	.058	.047	.046	.025	.025	.026	.024	.051	.063	.054	.052

Table B13. *Bias, SE, and RMSE of $\rho_{b\beta}$ in simulation study 2.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	-.005	.016	.018	.190	.033	.030	.034	.021	.034	.034	.038	.191
			.7	-.044	-.078	.010	.192	.044	.044	.037	.018	.062	.089	.039	.193
		.5	.3	-.105	-.098	-.054	.119	.040	.036	.034	.019	.112	.105	.064	.120
			.7	-.048	-.068	-.007	.182	.032	.029	.030	.019	.058	.074	.031	.183
		.8	.3	-.062	-.087	-.049	.076	.052	.046	.040	.020	.081	.098	.063	.079
			.7	-.029	-.044	.015	.186	.041	.040	.033	.015	.050	.059	.036	.186
	40	.2	.3	-.050	-.147	-.009	.161	.034	.034	.025	.017	.061	.151	.027	.162
			.7	-.018	-.018	.010	.200	.021	.021	.020	.012	.027	.028	.022	.200
		.5	.3	-.003	-.048	.035	.199	.035	.032	.027	.012	.035	.058	.044	.199
			.7	-.034	-.092	.004	.181	.032	.025	.025	.016	.047	.096	.026	.182
		.8	.3	-.027	-.028	.000	.168	.028	.029	.021	.014	.039	.041	.021	.169
			.7	-.009	-.023	.025	.220	.023	.025	.020	.012	.025	.034	.032	.220
1000	20	.2	.3	-.042	-.087	-.037	.097	.020	.018	.021	.015	.047	.089	.043	.099
			.7	-.037	-.038	-.021	.160	.026	.026	.026	.017	.046	.046	.034	.161
		.5	.3	.065	.034	.013	.097	.046	.025	.022	.016	.079	.043	.025	.098
			.7	-.053	-.057	-.013	.196	.023	.023	.028	.012	.058	.061	.031	.196
		.8	.3	-.050	-.008	.012	.121	.037	.032	.023	.015	.063	.033	.026	.122
			.7	-.035	-.062	-.022	.149	.018	.019	.016	.013	.039	.065	.027	.150
	40	.2	.3	-.032	-.042	-.018	.134	.023	.024	.019	.012	.039	.049	.026	.134
			.7	-.011	-.010	.034	.189	.024	.021	.019	.010	.027	.023	.039	.190
		.5	.3	-.012	-.019	.007	.171	.023	.019	.019	.011	.026	.027	.020	.171
			.7	-.015	-.037	.021	.193	.020	.018	.018	.010	.025	.041	.028	.193
		.8	.3	-.015	-.055	-.003	.166	.021	.015	.019	.010	.025	.057	.020	.166
			.7	-.017	-.082	-.001	.178	.020	.019	.017	.009	.026	.084	.017	.178

Table B14. *Bias, SE, and RMSE of σ_{β}^{2} in simulation study 2.*

J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	Bias				SE				RMSE			
				JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.087	.105	.097	.190	.017	.011	.017	.006	.089	.105	.099	.190
			.7	.070	.051	.079	.129	.011	.007	.011	.005	.070	.052	.080	.129
		.5	.3	.054	.063	.065	.117	.009	.009	.009	.007	.054	.064	.066	.117
			.7	.074	.084	.081	.166	.011	.008	.013	.006	.075	.084	.082	.167
		.8	.3	.070	.091	.070	.127	.010	.006	.010	.006	.071	.091	.071	.127
			.7	.067	.057	.075	.115	.009	.006	.008	.005	.068	.058	.075	.115
	40	.2	.3	.024	.023	.032	.085	.009	.004	.010	.003	.026	.023	.034	.085
			.7	.034	.038	.039	.091	.006	.004	.007	.004	.035	.038	.040	.091
		.5	.3	.036	.015	.045	.088	.008	.007	.008	.004	.037	.017	.045	.088
			.7	.031	.021	.036	.084	.007	.005	.007	.004	.032	.022	.037	.084
		.8	.3	.032	.040	.037	.096	.010	.006	.010	.004	.034	.041	.039	.096
			.7	.036	.030	.040	.094	.006	.003	.006	.004	.037	.030	.041	.095
1000	20	.2	.3	.067	.068	.068	.117	.008	.009	.008	.004	.067	.069	.068	.117
			.7	.072	.098	.076	.162	.007	.007	.007	.005	.072	.098	.076	.162
		.5	.3	.066	.057	.065	.090	.007	.005	.006	.005	.067	.057	.065	.091
			.7	.060	.061	.083	.191	.011	.008	.016	.005	.061	.061	.084	.191
		.8	.3	.078	.135	.086	.185	.008	.014	.007	.004	.078	.135	.086	.185
			.7	.071	.091	.073	.161	.010	.005	.010	.005	.071	.091	.073	.161
	40	.2	.3	.031	.033	.036	.081	.008	.008	.007	.003	.032	.034	.036	.081
			.7	.036	.031	.045	.098	.005	.004	.005	.003	.037	.031	.045	.098
		.5	.3	.033	.030	.036	.083	.005	.004	.005	.003	.034	.031	.036	.084
			.7	.037	.038	.046	.095	.006	.003	.007	.003	.038	.038	.047	.095
		.8	.3	.035	.036	.039	.108	.007	.005	.007	.003	.035	.037	.040	.108
			.7	.033	.051	.033	.103	.005	.004	.005	.003	.034	.051	.033	.103

Table B15. *Bias, SE, and RMSE of σ_{θ}^{2} in simulation study 2.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.019	.024	.019	.011	.053	.053	.053	.053	.056	.058	.056	.054
			.7	.009	-.021	.017	.020	.060	.061	.062	.061	.061	.065	.064	.065
		.5	.3	-.004	-.002	-.002	-.004	.068	.071	.068	.072	.068	.071	.068	.073
			.7	.009	.001	.009	.006	.052	.051	.054	.056	.053	.051	.055	.056
		.8	.3	-.002	-.058	.001	.006	.049	.055	.049	.050	.050	.080	.049	.050
			.7	.015	.012	.018	.017	.060	.062	.060	.063	.062	.063	.063	.065
	40	.2	.3	.010	-.013	.014	.023	.041	.040	.041	.048	.042	.042	.044	.053
			.7	.030	.027	.032	.026	.033	.034	.035	.036	.045	.043	.047	.045
		.5	.3	-.008	-.024	-.004	-.003	.042	.041	.042	.041	.043	.047	.042	.041
			.7	-.009	-.039	-.007	-.002	.042	.043	.043	.041	.043	.058	.044	.041
		.8	.3	-.002	-.021	.000	.002	.036	.039	.036	.041	.036	.045	.036	.041
			.7	-.003	-.008	-.002	-.002	.045	.042	.045	.045	.045	.043	.045	.045
1000	20	.2	.3	.009	-.032	.008	.009	.053	.054	.054	.054	.054	.063	.054	.055
			.7	-.015	-.014	-.016	-.013	.033	.035	.033	.035	.036	.038	.037	.038
		.5	.3	.002	-.013	.003	.001	.031	.033	.031	.035	.031	.036	.031	.035
			.7	.012	.009	.015	.011	.044	.044	.044	.043	.045	.045	.046	.045
		.8	.3	.006	.008	.007	.002	.044	.046	.045	.044	.044	.046	.045	.044
			.7	.012	.004	.011	.007	.043	.041	.043	.043	.045	.041	.044	.044
	40	.2	.3	-.009	-.015	-.009	-.008	.025	.024	.025	.025	.027	.029	.027	.026
			.7	-.001	-.013	.001	-.001	.029	.029	.029	.031	.029	.032	.029	.031
		.5	.3	-.006	-.015	-.005	-.008	.029	.030	.030	.030	.030	.033	.031	.030
			.7	-.001	-.029	.000	-.005	.030	.030	.030	.031	.030	.042	.030	.031
		.8	.3	.001	-.010	.001	.003	.031	.031	.031	.031	.031	.032	.031	.031
			.7	-.005	-.030	-.006	-.009	.028	.030	.028	.028	.028	.042	.028	.030

Table B16. *Bias, SE, and RMSE of $\rho_{\theta\tau}$ in simulation study 2.*

				Bias				SE				RMSE			
J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.004	.003	.002	.138	.030	.029	.029	.024	.030	.029	.029	.140
			.7	-.002	-.008	-.002	.153	.025	.024	.025	.028	.025	.025	.026	.155
		.5	.3	.004	.000	.003	.119	.022	.026	.023	.022	.023	.026	.023	.121
			.7	-.006	-.008	-.007	.111	.028	.028	.027	.020	.028	.030	.028	.113
		.8	.3	-.008	-.013	-.008	.045	.017	.018	.017	.014	.018	.023	.019	.047
			.7	-.010	-.010	-.010	.044	.015	.015	.015	.014	.018	.018	.018	.046
	40	.2	.3	-.008	-.011	-.008	.155	.020	.020	.020	.019	.021	.023	.022	.156
			.7	.011	.010	.010	.162	.019	.019	.019	.014	.022	.021	.021	.163
		.5	.3	.005	-.001	.005	.113	.013	.015	.013	.014	.014	.015	.014	.114
			.7	.001	-.001	.001	.115	.020	.020	.020	.016	.020	.020	.020	.116
		.8	.3	-.007	-.010	-.007	.042	.011	.011	.011	.010	.013	.015	.014	.043
			.7	-.006	-.007	-.007	.040	.008	.009	.008	.008	.010	.011	.010	.041
1000	20	.2	.3	.004	-.006	.004	.168	.016	.017	.016	.015	.017	.018	.017	.168
			.7	.001	-.003	-.001	.156	.019	.019	.018	.014	.019	.019	.018	.157
		.5	.3	-.002	-.002	-.002	.111	.017	.018	.017	.015	.017	.018	.017	.112
			.7	-.001	-.002	-.003	.116	.019	.020	.019	.015	.020	.020	.020	.117
		.8	.3	-.005	-.006	-.005	.046	.010	.012	.010	.010	.012	.014	.012	.047
			.7	-.009	-.008	-.009	.048	.010	.012	.010	.010	.013	.014	.014	.049
	40	.2	.3	-.001	-.006	-.001	.149	.012	.012	.012	.010	.012	.014	.012	.150
			.7	-.006	-.007	-.006	.149	.013	.014	.013	.013	.014	.016	.014	.149
		.5	.3	-.002	-.004	-.002	.107	.013	.013	.013	.011	.013	.014	.013	.107
			.7	-.001	-.004	-.002	.119	.010	.011	.011	.008	.010	.012	.011	.119
		.8	.3	-.003	-.004	-.003	.044	.008	.009	.008	.007	.009	.010	.009	.045
			.7	-.002	-.003	-.002	.048	.008	.008	.008	.006	.008	.009	.008	.049

Table B17. *Bias, SE, and RMSE of σ_{τ}^{2} in simulation study 2.*

J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	Bias				SE				RMSE			
				JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM	JM-DD1	JM-DD2	JM-D	HM
500	20	.2	.3	.004	.003	.003	.023	.007	.007	.007	.006	.008	.007	.007	.024
			.7	.002	.002	.002	.027	.006	.006	.006	.005	.006	.006	.006	.028
		.5	.3	.002	.002	.002	.052	.007	.007	.007	.006	.008	.007	.007	.052
			.7	.005	.004	.004	.058	.006	.006	.006	.006	.008	.007	.007	.058
		.8	.3	.001	-.003	.001	.091	.006	.006	.006	.005	.006	.007	.006	.091
			.7	.003	.002	.003	.080	.007	.006	.007	.006	.007	.006	.007	.080
	40	.2	.3	.001	.001	.001	.026	.004	.004	.004	.004	.004	.004	.004	.027
			.7	.003	.003	.003	.029	.004	.004	.004	.004	.005	.005	.005	.030
		.5	.3	.004	.003	.003	.052	.004	.004	.005	.004	.006	.005	.006	.053
			.7	.003	.002	.003	.054	.005	.005	.005	.005	.006	.005	.006	.054
		.8	.3	.004	.002	.004	.073	.006	.006	.006	.004	.007	.006	.007	.074
			.7	.002	.001	.001	.076	.006	.005	.006	.004	.006	.006	.006	.076
1000	20	.2	.3	.002	.001	.002	.031	.003	.003	.003	.003	.004	.003	.004	.031
			.7	.001	.001	.001	.028	.005	.005	.005	.004	.005	.005	.005	.028
		.5	.3	.001	.000	.001	.050	.004	.004	.004	.004	.004	.004	.004	.051
			.7	.001	.000	.000	.052	.004	.004	.004	.004	.004	.004	.004	.052
		.8	.3	.001	.000	.001	.076	.005	.004	.005	.004	.005	.004	.005	.076
			.7	.002	.000	.001	.079	.004	.004	.004	.003	.004	.004	.004	.080
	40	.2	.3	.001	.001	.001	.025	.003	.003	.003	.003	.003	.003	.003	.025
			.7	.002	.001	.002	.025	.002	.002	.002	.002	.003	.003	.003	.025
		.5	.3	.001	.000	.001	.049	.004	.004	.004	.003	.004	.004	.004	.049
			.7	.001	-.001	.001	.055	.004	.004	.004	.003	.004	.004	.004	.055
		.8	.3	.002	.003	.002	.070	.004	.004	.004	.003	.004	.005	.004	.071
			.7	.001	.000	.001	.081	.003	.003	.003	.002	.003	.003	.003	.081

Appendix C

Table C1. *Mean and SD of bias, SE, and RMSE in item difficulty estimation in simulation study 3.*

						Mean							SD						
Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	-.003	-.009	-.002	.005	.002	.006	-.003	.023	.108	.021	.031	.027	.030	.021
		500	40	.2	.3	-.001	.002	-.002	-.003	-.002	-.001	-.002	.025	.094	.024	.024	.026	.024	.025
	SE	500	20	.2	.3	.101	.098	.101	.103	.102	.103	.102	.016	.017	.016	.015	.017	.014	.016
		500	40	.2	.3	.103	.103	.103	.103	.103	.104	.103	.015	.015	.015	.015	.015	.015	.015
	RMSE	500	20	.2	.3	.103	.135	.103	.107	.105	.106	.104	.016	.054	.016	.018	.018	.018	.016
		500	40	.2	.3	.106	.134	.105	.106	.106	.106	.106	.016	.039	.016	.016	.016	.016	.016
JM-DD1	Bias	500	20	.2	.3	.004	.005	.004	.008	.007	.013	.004	.021	.023	.020	.020	.030	.018	.020
		500	40	.2	.3	.002	.002	.002	-.002	-.002	.002	.002	.022	.023	.022	.019	.064	.020	.022
	SE	500	20	.2	.3	.103	.103	.103	.093	.095	.094	.103	.020	.020	.020	.020	.022	.020	.020
		500	40	.2	.3	.106	.107	.107	.088	.088	.088	.107	.016	.016	.016	.027	.026	.027	.017
	RMSE	500	20	.2	.3	.106	.106	.105	.095	.099	.096	.105	.020	.021	.020	.021	.024	.021	.020
		500	40	.2	.3	.109	.109	.109	.090	.106	.090	.109	.017	.017	.017	.027	.036	.027	.017

Table C2. *Mean and SD of bias, SE, and RMSE in time discrimination estimation in simulation study 3.*

						Mean							SD						
Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	-.009	-.028	-.009	-.081	-.088	-.081	-.092	.013	.028	.013	.077	.081	.079	.084
		500	40	.2	.3	-.009	-.034	-.009	-.091	-.096	-.091	-.099	.013	.035	.013	.079	.081	.079	.083
	SE	500	20	.2	.3	.064	.063	.064	.064	.062	.064	.062	.010	.010	.010	.010	.010	.010	.010
		500	40	.2	.3	.063	.062	.063	.061	.060	.061	.060	.008	.008	.008	.009	.009	.009	.009
	RMSE	500	20	.2	.3	.066	.074	.066	.112	.117	.113	.119	.009	.015	.009	.063	.068	.064	.071
		500	40	.2	.3	.065	.076	.065	.119	.122	.119	.125	.008	.022	.008	.063	.067	.063	.069
JM-DD1	Bias	500	20	.2	.3	-.050	-.052	-.050	-.011	-.026	-.011	-.052	.040	.041	.040	.011	.026	.011	.042
		500	40	.2	.3	-.067	-.070	-.067	-.013	-.032	-.013	-.070	.049	.051	.049	.013	.032	.013	.051
	SE	500	20	.2	.3	.061	.061	.061	.064	.063	.064	.061	.009	.009	.009	.009	.009	.009	.009
		500	40	.2	.3	.060	.060	.061	.063	.062	.063	.060	.007	.007	.007	.007	.008	.007	.007
	RMSE	500	20	.2	.3	.084	.085	.084	.066	.072	.066	.085	.029	.030	.029	.009	.018	.009	.031
		500	40	.2	.3	.095	.097	.095	.065	.074	.065	.097	.039	.041	.039	.008	.019	.008	.041

Table C3. *Mean and SD of bias, SE, and RMSE in time intensity estimation in simulation study 3.*

						Mean							SD						
Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.000	-.002	-.001	-.105	-.123	-.116	-.173	.009	.096	.009	.070	.133	.090	.181
		500	40	.2	.3	.000	.008	-.001	-.103	-.105	-.117	-.167	.010	.094	.008	.074	.098	.078	.128
	SE	500	20	.2	.3	.033	.026	.033	.042	.029	.044	.022	.008	.006	.008	.023	.008	.028	.004
		500	40	.2	.3	.033	.025	.033	.036	.027	.036	.024	.007	.004	.007	.014	.008	.013	.003
	RMSE	500	20	.2	.3	.034	.074	.034	.115	.138	.126	.177	.009	.065	.009	.071	.120	.093	.178
		500	40	.2	.3	.034	.078	.034	.117	.121	.128	.175	.009	.058	.008	.063	.083	.070	.118
JM-DD1	Bias	500	20	.2	.3	-.178	-.178	-.181	-.014	-.023	-.022	-.187	.204	.207	.211	.056	.130	.076	.222
		500	40	.2	.3	-.167	-.165	-.167	.012	.033	.000	-.174	.140	.146	.142	.036	.117	.016	.154
	SE	500	20	.2	.3	.028	.024	.028	.046	.035	.046	.022	.006	.004	.005	.028	.014	.028	.003
		500	40	.2	.3	.029	.023	.028	.043	.032	.042	.022	.006	.004	.006	.026	.012	.024	.003
	RMSE	500	20	.2	.3	.184	.185	.186	.055	.092	.060	.193	.201	.203	.208	.056	.100	.074	.218
		500	40	.2	.3	.172	.172	.173	.048	.091	.044	.181	.136	.140	.138	.040	.087	.026	.147

Table C4. *Mean and SD of bias, SE, and RMSE in ability estimation in simulation study 3.*

						Mean							SD						
Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.000	.000	.000	.000	.000	.000	.000	.239	.241	.239	.241	.243	.240	.245
		500	40	.2	.3	.000	.000	.000	.000	.000	.000	.000	.146	.147	.146	.150	.149	.149	.148
	SE	500	20	.2	.3	.392	.391	.392	.402	.395	.402	.390	.054	.054	.054	.056	.056	.055	.054
		500	40	.2	.3	.318	.318	.318	.325	.321	.324	.317	.046	.046	.046	.046	.046	.045	.045
	RMSE	500	20	.2	.3	.450	.450	.450	.460	.454	.459	.451	.104	.106	.104	.106	.107	.105	.107
		500	40	.2	.3	.346	.346	.346	.354	.350	.353	.346	.070	.071	.070	.071	.071	.070	.070
JM-DD1	Bias	500	20	.2	.3	.000	.000	.000	.000	.000	.000	.000	.238	.238	.238	.205	.211	.206	.239
		500	40	.2	.3	.000	.000	.000	.000	.000	.000	.000	.143	.143	.143	.122	.135	.121	.143
	SE	500	20	.2	.3	.390	.390	.391	.364	.374	.364	.390	.054	.054	.054	.057	.056	.056	.054
		500	40	.2	.3	.319	.319	.319	.261	.273	.262	.319	.042	.042	.042	.050	.049	.050	.042
	RMSE	500	20	.2	.3	.450	.450	.450	.410	.422	.410	.450	.100	.100	.100	.096	.097	.096	.100
		500	40	.2	.3	.345	.345	.345	.282	.297	.282	.345	.069	.069	.069	.079	.082	.077	.069

Table C5. *Mean and SD of bias, SE, and RMSE in speed estimation in simulation study 3.*

						Mean							SD						
Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.000	.000	.000	.000	.000	.000	.000	.032	.033	.032	.052	.054	.051	.061
		500	40	.2	.3	.000	.000	.000	.000	.000	.000	.000	.019	.019	.019	.040	.040	.039	.057
	SE	500	20	.2	.3	.104	.105	.104	.107	.107	.107	.108	.013	.013	.013	.014	.014	.014	.014
		500	40	.2	.3	.074	.075	.074	.077	.077	.077	.078	.010	.010	.010	.011	.011	.011	.011
	RMSE	500	20	.2	.3	.109	.110	.109	.118	.119	.118	.123	.014	.014	.014	.019	.020	.019	.022
		500	40	.2	.3	.077	.078	.077	.086	.086	.086	.095	.010	.010	.010	.015	.015	.014	.020
JM-DD1	Bias	500	20	.2	.3	.000	.000	.000	.000	.000	.000	.000	.081	.081	.081	.038	.039	.038	.082
		500	40	.2	.3	.000	.000	.000	.000	.000	.000	.000	.090	.090	.090	.023	.023	.023	.092
	SE	500	20	.2	.3	.104	.104	.104	.106	.106	.106	.104	.014	.014	.014	.014	.014	.014	.014
		500	40	.2	.3	.075	.075	.075	.078	.078	.078	.075	.010	.010	.010	.011	.011	.011	.010
	RMSE	500	20	.2	.3	.129	.128	.129	.113	.113	.113	.129	.030	.031	.031	.016	.017	.016	.031
		500	40	.2	.3	.113	.113	.113	.081	.081	.081	.114	.031	.031	.031	.012	.012	.012	.032

Table C6. *Bias, SE, and RMSE of ω_0 in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D
JM-RD1	Bias	500	20	.2	.3	.001	.012	.000	.226	.256	.220
		500	40	.2	.3	.000	-.008	.000	.216	.219	.218
	SE	500	20	.2	.3	.009	.009	.010	.031	.028	.030
		500	40	.2	.3	.009	.009	.008	.019	.017	.019
	RMSE	500	20	.2	.3	.009	.015	.010	.228	.257	.222
		500	40	.2	.3	.009	.012	.008	.217	.220	.219
JM-DD1	Bias	500	20	.2	.3	.292	.292	.292	.024	.041	.018
		500	40	.2	.3	.289	.287	.289	-.014	-.032	-.012
	SE	500	20	.2	.3	.013	.013	.013	.036	.036	.036
		500	40	.2	.3	.008	.008	.008	.016	.016	.016
	RMSE	500	20	.2	.3	.292	.292	.292	.043	.054	.040
		500	40	.2	.3	.289	.287	.289	.022	.036	.020

Table C7. *Bias, SE, and RMSE of ω_1 in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-DD1	JM-DD2
JM-RD1	Bias	500	20	.2	.3	.078	.088	.055	.035
		500	40	.2	.3	.033	.059	.018	.018
	SE	500	20	.2	.3	.016	.019	.024	.023
		500	40	.2	.3	.012	.014	.020	.022
	RMSE	500	20	.2	.3	.079	.090	.060	.042
		500	40	.2	.3	.035	.061	.027	.029
JM-DD1	Bias	500	20	.2	.3	-.030	-.031	.072	.060
		500	40	.2	.3	-.047	-.045	.044	.090
	SE	500	20	.2	.3	.014	.013	.034	.030
		500	40	.2	.3	.010	.010	.023	.026
	RMSE	500	20	.2	.3	.033	.033	.080	.067
		500	40	.2	.3	.048	.046	.049	.094

Table C8. *Bias, SE, and RMSE of $\rho_{b\lambda}$ in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-DD1
JM-RD1	Bias	500	20	.2	.3	.076	.155
		500	40	.2	.3	-.056	.059
	SE	500	20	.2	.3	.041	.093
		500	40	.2	.3	.031	.089
	RMSE	500	20	.2	.3	.086	.181
		500	40	.2	.3	.064	.107
JM-DD1	Bias	500	20	.2	.3	.038	.097
		500	40	.2	.3	-.199	-.040
	SE	500	20	.2	.3	.158	.100
		500	40	.2	.3	.078	.055
	RMSE	500	20	.2	.3	.163	.139
		500	40	.2	.3	.214	.068

Table C9. *Bias, SE, and RMSE of σ_{ϕ}^{2} in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-R	JM-DD1	JM-D
JM-RD1	Bias	500	20	.2	.3	.001	.018	-.017	-.013
		500	40	.2	.3	.002	.006	-.025	-.026
	SE	500	20	.2	.3	.005	.007	.009	.012
		500	40	.2	.3	.003	.004	.003	.004
	RMSE	500	20	.2	.3	.005	.019	.019	.017
		500	40	.2	.3	.004	.007	.025	.026
JM-DD1	Bias	500	20	.2	.3	-.033	-.036	-.005	.005
		500	40	.2	.3	-.031	-.035	.005	.008
	SE	500	20	.2	.3	.001	.002	.008	.010
		500	40	.2	.3	.001	.001	.006	.007
	RMSE	500	20	.2	.3	.033	.036	.009	.011
		500	40	.2	.3	.031	.035	.007	.011

Table C10. *Bias, SE, and RMSE of μ_b in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	-.003	-.009	-.002	.005	.002	.005	-.002
		500	40	.2	.3	-.002	.002	-.002	-.003	-.001	-.001	-.001
	SE	500	20	.2	.3	.018	.019	.019	.018	.018	.017	.018
		500	40	.2	.3	.018	.017	.018	.018	.018	.019	.018
	RMSE	500	20	.2	.3	.018	.021	.019	.018	.018	.018	.018
		500	40	.2	.3	.018	.018	.019	.019	.018	.019	.018
JM-DD1	Bias	500	20	.2	.3	.006	.006	.006	.007	.006	.012	.005
		500	40	.2	.3	.002	.003	.002	-.002	-.002	.003	.002
	SE	500	20	.2	.3	.025	.026	.026	.024	.024	.023	.025
		500	40	.2	.3	.016	.016	.017	.013	.014	.013	.017
	RMSE	500	20	.2	.3	.026	.026	.026	.025	.025	.026	.025
		500	40	.2	.3	.017	.017	.017	.013	.014	.013	.017

Table C11. *Bias, SE, and RMSE of* μ_β *in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.000	-.002	-.001	-.104	-.123	-.116	-.171
		500	40	.2	.3	.000	.008	-.001	-.103	-.105	-.117	-.166
	SE	500	20	.2	.3	.008	.008	.008	.017	.016	.017	.006
		500	40	.2	.3	.005	.005	.005	.011	.010	.010	.003
	RMSE	500	20	.2	.3	.008	.008	.008	.105	.124	.117	.171
		500	40	.2	.3	.005	.009	.005	.103	.105	.117	.166
JM-DD1	Bias	500	20	.2	.3	-.176	-.176	-.179	-.014	-.023	-.022	-.185
		500	40	.2	.3	-.166	-.164	-.167	.012	.033	.000	-.174
	SE	500	20	.2	.3	.010	.010	.009	.021	.020	.021	.006
		500	40	.2	.3	.006	.005	.005	.012	.012	.010	.004
	RMSE	500	20	.2	.3	.177	.177	.179	.025	.031	.030	.185
		500	40	.2	.3	.166	.164	.167	.018	.035	.010	.174

Table C12. *Bias, SE, and RMSE of σ_b^2 in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.116	.109	.118	.100	.106	.101	.117
		500	40	.2	.3	.054	.037	.057	.051	.054	.052	.055
	SE	500	20	.2	.3	.048	.049	.047	.047	.047	.048	.047
		500	40	.2	.3	.043	.042	.043	.042	.043	.042	.042
	RMSE	500	20	.2	.3	.125	.119	.127	.110	.116	.112	.126
		500	40	.2	.3	.069	.056	.071	.067	.069	.067	.069
JM-DD1	Bias	500	20	.2	.3	.110	.112	.113	.107	.103	.108	.113
		500	40	.2	.3	.053	.051	.054	.065	.074	.061	.054
	SE	500	20	.2	.3	.054	.054	.053	.053	.052	.052	.053
		500	40	.2	.3	.041	.040	.040	.041	.041	.040	.040
	RMSE	500	20	.2	.3	.123	.124	.125	.120	.115	.120	.125
		500	40	.2	.3	.067	.065	.067	.077	.084	.073	.067

Table C13. *Bias, SE, and RMSE of $\rho_{b\beta}$ in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	-.036	-.007	-.032	.037	.070	.067	.149
		500	40	.2	.3	-.023	-.072	-.015	.033	.030	.078	.135
	SE	500	20	.2	.3	.027	.025	.026	.030	.026	.031	.021
		500	40	.2	.3	.018	.023	.018	.028	.027	.021	.014
	RMSE	500	20	.2	.3	.044	.026	.041	.048	.075	.074	.150
		500	40	.2	.3	.029	.076	.023	.043	.041	.080	.135
JM-DD1	Bias	500	20	.2	.3	.172	.173	.181	-.005	.016	.018	.190
		500	40	.2	.3	.157	.146	.160	-.050	-.147	-.009	.161
	SE	500	20	.2	.3	.024	.023	.021	.033	.030	.034	.021
		500	40	.2	.3	.019	.019	.017	.034	.034	.025	.017
	RMSE	500	20	.2	.3	.174	.174	.182	.034	.034	.038	.191
		500	40	.2	.3	.158	.148	.161	.061	.151	.027	.162

Table C14. *Bias, SE, and RMSE of σ_{β}^{2} in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.072	.085	.073	.095	.117	.109	.157
		500	40	.2	.3	.032	.017	.033	.031	.034	.042	.061
	SE	500	20	.2	.3	.006	.006	.006	.014	.010	.018	.006
		500	40	.2	.3	.005	.004	.006	.007	.006	.007	.005
	RMSE	500	20	.2	.3	.072	.086	.073	.096	.117	.110	.157
		500	40	.2	.3	.032	.017	.034	.032	.035	.042	.061
JM-DD1	Bias	500	20	.2	.3	.179	.177	.185	.087	.105	.097	.190
		500	40	.2	.3	.081	.079	.083	.024	.023	.032	.085
	SE	500	20	.2	.3	.008	.008	.007	.017	.011	.017	.006
		500	40	.2	.3	.006	.004	.005	.009	.004	.010	.003
	RMSE	500	20	.2	.3	.179	.177	.185	.089	.105	.099	.190
		500	40	.2	.3	.081	.079	.083	.026	.023	.034	.085

Table C15. *Bias, SE, and RMSE of σ_θ^2 in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.004	-.002	.004	.015	.011	.017	.004
		500	40	.2	.3	.000	-.003	.001	-.001	.000	.001	.000
	SE	500	20	.2	.3	.066	.066	.067	.068	.066	.068	.067
		500	40	.2	.3	.044	.044	.044	.044	.044	.044	.045
	RMSE	500	20	.2	.3	.067	.066	.067	.070	.067	.070	.067
		500	40	.2	.3	.044	.045	.044	.044	.044	.044	.045
JM-DD1	Bias	500	20	.2	.3	.012	.012	.012	.019	.024	.019	.011
		500	40	.2	.3	.023	.023	.022	.010	-.013	.014	.023
	SE	500	20	.2	.3	.052	.052	.053	.053	.053	.053	.053
		500	40	.2	.3	.048	.047	.047	.041	.040	.041	.048
	RMSE	500	20	.2	.3	.054	.053	.054	.056	.058	.056	.054
		500	40	.2	.3	.053	.053	.052	.042	.042	.044	.053

Table C16. *Bias, SE, and RMSE of $\rho_{\theta\tau}$ in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.010	.010	.009	.097	.105	.095	.128
		500	40	.2	.3	-.003	-.003	-.002	.069	.070	.068	.109
	SE	500	20	.2	.3	.028	.028	.028	.037	.036	.036	.026
		500	40	.2	.3	.020	.020	.020	.020	.021	.020	.020
	RMSE	500	20	.2	.3	.030	.030	.030	.103	.111	.101	.131
		500	40	.2	.3	.021	.020	.021	.072	.073	.071	.111
JM-DD1	Bias	500	20	.2	.3	.135	.135	.135	.004	.003	.002	.138
		500	40	.2	.3	.151	.151	.151	-.008	-.011	-.008	.155
	SE	500	20	.2	.3	.026	.026	.026	.030	.029	.029	.024
		500	40	.2	.3	.019	.019	.019	.020	.020	.020	.019
	RMSE	500	20	.2	.3	.138	.138	.138	.030	.029	.029	.140
		500	40	.2	.3	.152	.152	.152	.021	.023	.022	.156

Table C17. *Bias, SE, and RMSE of σ_{τ}^{2} in simulation study 3.*

Data Generating Model	Fit Indices	J	I	$\rho_{\theta\tau}$	$\rho_{b\lambda}$	JM-RD1	JM-RD2	JM-R	JM-DD1	JM-DD2	JM-D	HM
JM-RD1	Bias	500	20	.2	.3	.002	.002	.003	.011	.012	.010	.015
		500	40	.2	.3	.003	.003	.003	.010	.010	.010	.016
	SE	500	20	.2	.3	.005	.005	.005	.007	.006	.006	.006
		500	40	.2	.3	.003	.003	.003	.004	.004	.004	.004
	RMSE	500	20	.2	.3	.006	.006	.006	.013	.013	.012	.016
		500	40	.2	.3	.004	.004	.004	.011	.011	.011	.017
JM-DD1	Bias	500	20	.2	.3	.023	.023	.023	.004	.003	.003	.023
		500	40	.2	.3	.025	.025	.025	.001	.001	.001	.026
	SE	500	20	.2	.3	.005	.005	.005	.007	.007	.007	.006
		500	40	.2	.3	.004	.004	.004	.004	.004	.004	.004
	RMSE	500	20	.2	.3	.023	.023	.023	.008	.007	.007	.024
		500	40	.2	.3	.026	.026	.026	.004	.004	.004	.027

References

Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: OECD.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317-332.

Beerli, P. (2005). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, *22*(3), 341-345.

Berger, J. O., Brunero, L., & Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, *14*, 1-28.

Bergstrom, B., Gershon, R., & Lunz, M. E. (1994). *Computerized adaptive testing: Exploring examinee response time using hierarchical liner modeling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores* (chapter 17-29). Reading, MA: Addison-Wesley.

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika*, *50*(4), 383-397.

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., & 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. (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.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education*, *12*(4), 383-407.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3-62.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153-168.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd edition). New York, NY: Springer.

Bustamante, C. D., Nielsen, R., & Hartl, D. L. (2003). Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theoretical Population Biology*, *63*(2), 91-103.

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement*, *32*(1), 79-96.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404-431.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (2007). The patient-Reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, *45*(5 Suppl 1), S3-S11.

Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, *33*(2), 181-201.

Chang, S.-R. (2007). *Computerized adaptive test item response times for correct and incorrect pretest and operational items: Testing fairness and test-taking strategies* (Doctoral dissertation). Retrieved from https://search.proquest.com/docview/304842739.

Cizek, G. J. & Wollack, J. A. (Eds.) (2017). *Handbook of Quantitative Methods for Detecting Cheating on Tests*. New York, NY: Routledge.

Clauser, B., Margolis, M., & von Davier, M. (2017). *Timing issues in simulations, games, and other performance assessments.* Paper presented at the Timing Impact on Measurement in Education conference, Philadelphia, PA.

Clinton, J., Jackman, S., & Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review*, *98*(2), 355-370.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Dennis, I., & Evans, J. S. B. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology*, *87*(1), 105-129.

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.

Ebel, R. L. (1953). The use of item response time measurements in the construction of educational achievement tests. *Educational and Psychological Measurement*, *13*(3), 391-401.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380-396.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists* (Multivariate applications book series). Mahwah, N.J.: L. Erlbaum Associates.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525-543.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359-374.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*(3), 396-413.

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

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*(7), 1-14.

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., & 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.

Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, 56, 1-106.

Gaviria, J. L. (2005). Increase in precision when estimating parameters in computer assisted testing using response time. *Quality & Quantity, 39*(1), 45-69.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515-534.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. New York, NY: Chapman & Hall.

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.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733-807.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457-511.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721-741.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics* (Vol. 4, pp. 169-193). Oxford, UK: Oxford University Press.

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.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237-288.

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., & 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., 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., 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.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. *Annales de I'Institut Henri Poincaré*, *5*(2), 115-158.

Gumbel, E. J. (1941). The return period of flood flows. *The Annals of Mathematical Statistics*, *12*(2), 163-190.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Nijhof.

Harik, P. (2017). *Timing and examinee pacing on a test of physician licensure: Experimental findings*. Paper presented at the Timing Impact on Measurement in Education conference, Philadelphia, PA.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186-201.

Holden, R. R., & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment*, *4*(2), 170-173.

Hornke, L.F. (2000) Item response time in computerized adaptive testing. *Psychologia-Revista de Metodologia y Psycologia Experimental*, *21*, 175-189.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*(3), 261-277.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, *1*(1), 69-82.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188-196.

IBM Corp. (2017). *IBM SPSS Statistics for Windows, Version 25.0.* Armonk, NY: IBM Corp.

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.

Ip, E. H. S. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*(1), 73-91.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*(1), 1-21.

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/KANG-DISSERTATION-2016.pdf?sequence=1&isAllowed=y.

Kennedy, M. (1930). Speed as a personality trait. *The Journal of Social Psychology*, *1*(2), 286-299.

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*(1), 54-75.

Klein Entink, R. H., van der Linden, W. J., & Fox, J. P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 621-640.

Kim, J., & Bolt, D. M. (2007). An NCME instructional module on estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*, 38-51.

Lee, S. Y., & Wollack, J. (2017). *Use of response time for detecting security threats and other anomalous behaviors.* Paper presented at the Timing Impact on Measurement in Education conference, Philadelphia, PA.

Lee, Y.-H. (2007). *Contributions to the statistical analysis of item response time in educational testing* (Doctoral dissertation). Retrieved from https://search.proquest.com/openview/6f55d58f2fa0647fadda566f916728b9/1?pq-origsite=gscholar&cbl=18750&diss=y.

Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden, J., & Gisle, L. (2017). Work organization and mental health problems in PhD students. *Research Policy*, *46*(4), 868-879.

Li, T. (2014). *Different approaches to covariate inclusion in the mixture Rasch model* (Doctoral dissertation). Retrieved from http://journals.sagepub.com/doi/pdf/1.1177/001316441561038.

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, *76*(3), 487-503.

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice reaction time responses: a model and a method. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 276-291.

Lord, F. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2), 157-162.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores* (Addison-Wesley series in behavioral science. quantitative methods). Reading, Mass: Addison-Wesley Pub.

Luce, R. D. (1986). *Response times: Their roles in inferring elementary mental organization.* Oxford, UK: Oxford University Press.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine, 28*(25), 3049-3067.

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*. doi:1.1177/1073191117740953

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*(6), 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, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615-633.

Martin, A. D., & Quinn, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953-1999. *Political Analysis*, *10*(2), 134-153.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics, 11*(2), 204-209.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*(2), 121-137.

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.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087-1092.

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. Retrieved from http://members.home.nl/jeroenvermunt/molenaar2016.pdf.

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., 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.

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*(3), 291-315.

Monica, R. S. (2008). Exploring the extension of item response theory models to the economic and social measurement. In *Proceedings of the 12th WSEAS international conference on Computers* (pp. 247-251). Retrieved from http://www.wseas.us/e-library/conferences/2008/crete/Computers/035-computers.pdf

Mulaik, S. A. (1972). *A mathematical investigation of some multidimensional Rasch models for psychological tests.* Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*(2), 252-284.

Muthén, L. K., & Muthén, B. O. (2007). Mplus. *Statistical analysis with latent variables. Version 3*.

O'neill, T. R., Marks, C. M., & Reynolds, M. (2005). Re-evaluating the NCLEX-RN® passing standard. *Journal of Nursing Measurement*, *13*(2), 147-168.

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report.* Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

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.

Plummer, M (2015). *JAGS Version 4.0.0 User Manual.* Lyon, France. URL http://sourceforge.net/projects/mcmc-jags/.

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.R-project.org/.

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. (2012). 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*(1), 61-80.

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). Analysing 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. (2012a). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 334-349.

Ranger, J., & Ortner, T. (2012b). 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: Danish Institute for Educational Research.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Rechase, M. D. (1972). *Development and application of a multivariate logistic latent trait model.* Unpublished doctoral dissertation, Syracuse University, Syracuse NY.

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York, NY: Springer.

Robert, C., & Casella, G. (1999). *Monte Carlo statistical methods* (Springer texts in statistics). New York: Springer.

Roberts, R. D., & Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: Toward a taxonomic model. *Learning and Individual Differences*, *11*(1), 1-12.

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). New York: Springer.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68*(4), 589-606.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172.

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*(1), 18-38.

Schnipke, D. L., & Scrams, D. J. (1999). Representing Response-Time Information in Item Banks. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626-633.

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.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual (Version 1.4.3)*. Cambridge, UK: MRC Biostatistics Unit.

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)*.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics-Theory and Methods*, *7*(1), 13-26.

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.

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, examine characteristics, and response times on USMLE Step 1. *Academic Medicine*, *76*(10), S114-S116.

Swanson, D. B., Featherman, C. M., Case, S. M., Luecht, R. M., & Nungester, R. (1999). *Relationship of response latency to test design, examinee proficiency and item difficulty in computer-based test administration*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* (pp. 82-98). Minneapolis, MN: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement, 8*(3-1), 353-374.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. (pp. 179-203). New York: Academic Press.

Tractenberg, R. E. (2010). Classical and modern measurement theories, patient reports, and clinical outcomes. *Contemporary Clinical Trials*, *31*(1), 1-3.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629-65.

Tukey, J. W. (1953). The problem of multiple comparisons. In H. Braun (Ed.), *The collected works of John W. Tukey volume VIII, multiple comparisons: 1948-1983* (pp. 1-300). New York: Chapman & Hall.

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.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287-308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247-272.

van der Linden, W. J., & Fox, J.-P. (2015). Joint hierarchical modeling of responses and response times. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol 1. Models.* Boca Raton: FL: Chapman & Hall/CRC.

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., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365-384.

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.

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 418-438.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251-265.

van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141-177.

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.

van der Maas, H. L., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, *118*(1), 29-60.

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.

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.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44-62.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. (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). New York: Springer.

Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. *Belmont, MA: Statistical Innovations Inc*.

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, 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, 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*(4), 381-417.

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, 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, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323-339.

Wang, T., & Zhang, J. (2006). Optimal partitioning of testing time: theoretical properties and practical implications. *Psychometrika*, *71*(1), 105-12.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126-149.

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*(4), 708-719.

Xie, C. (2014). *Cross-classified modeling of dual local item dependence* (Doctoral dissertation). Retrieved from https://drum.lib.umd.edu/handle/1903/15142.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, *30*(3), 187-213.