ABSTRACT

Dissertation Title:	A MULTILEVEL TESTLET JOINT MODEL OF RESPONSES AND RESPONSE TIME
	Evan D. Olson, Doctor of Philosophy, 2020
Dissertation directed by:	Associate Professor, Hong Jiao Measurement, Statistics, and Evaluation Department of Human Development and Quantitative Methodology

In approaches to joint response and response time (RT) modeling there is an assumption of conditional independence of the responses and the RTs. Further, in IRT modeling of the responses, there is the assumption that the items and the persons have local independence, respectively. In practice, violations of the local item independence results from the bundling of items into testlets. Violation of the person independence are encountered in complex examinee sampling situations.

A multilevel testlet joint responses and RT model is proposed and evaluated in this study that accounts for the dual local item and person dependence due to testlets and complex sampling. A simulation study is performed to investigate parameter recovery for the proposed model and provide comparison to models that do not model dual local dependencies. In addition to the simulation study, a study using empirical data is also conducted to evaluate relative model fit indices.

Generally, results determined by statistical analyses and inspection of graphs developed from descriptive statistics supported the need to model local item dependency and local person dependency. Parameter recovery outcome measures in the simulation study showed interaction of factors included with the model factor when the comparison models were included. When deviance model fit criterion was applied the proposed model was selected as the best-fitting model. For the Bayesian model fit index DIC the proposed model was not selected as best-fitting in for either the simulation or the empirical data analyses. Limitations of the study and opportunities to refine joint response and RT modeling of this dual dependency were elaborated.

A MULTILEVEL TESTLET JOINT MODEL OF RESPONSES AND RESPONSE TIME

by

Evan D. Olson

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
2020

Advisory Committee:
Dr. Hong Jiao, Chair
Dr. Robert W. Lissitz
Dr. Robert J. Mislevy
Dr. Peter M. Steiner
Dr. Yan Li, Dean's Representative

Dedication

To my children, you are now in the beginning of your school journey, and I am at the end of mine. You are my greatest joy.

Acknowledgements

First and foremost, I would like to express my deepest appreciation to Dr. Hong Jiao, a superlative educator and advisor who always met me where I was in terms of conceptions of the subject matter as well as in addressing obstacles that may have thwarted my ultimate success. Her kindness towards me and patience with me were seemingly limitless. As I am a non-traditional student, Dr. Jiao's willingness to talk or meet with me via text, phone, in-person, or in this newer virtual meeting world provided unparalleled advancement of my academic pursuits.

I am thankful for my committee members, some of whom have shaped my understanding since I began pursuing knowledge about psychometrics, and others with whom I have only had recent experience, but who have tremendously assisted on this dissertation path. I want to thank Dr. Robert Mislevy who served as advisor in my master's work and provided me an unexpected, but certainly welcome, graduate research assistantship upon my return to the EDMS program. This assistantship afforded me the opportunity to travel and enjoy his company in a setting away from the classroom. He has been very generous with sharing both his knowledge and his encouragement. I am also grateful to Dr. Robert Lissitz who has provided me advice, which when finally heeded, was instrumental in moving the dissertation forward. I would like to also acknowledge his strong support of student learning when he provided instruction in special topic areas to me and several colleagues via independent study. My gratitude also goes to my other two committee members, Dr. Peter Steiner and Dr. Yan Li, especially for their helpful suggestions on the research

and their remarkable flexibility during these times of uncertainty and global pandemic.

The current and former EDMS faculty and students with whom I have engaged has shown this to be a fantastic community. As I come to the close of this chapter of my academic life, I want to express special thanks to Dr. Jeffrey Harring and to Jannitta Graham for navigating me through the waters of administration.

Finally, I want to share my thanks to my family for their continuous support on this adventure. Though cautioned against the approach, I sought a post-graduate degree after working as an adult, following the footsteps of my mother. Thank you for leading the way. To my father, I thank you for instilling in me a fascination with science and exploration that drives me to this day, and will, with good fortune, be passed along to my children.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The applications of item response theory (IRT) in scoring require a fundamental assumption that item scores are related only underlyingly through latent person variables. This assumption, known as local item independence, requires that the performance on one item should only have a relationship with the underlying latent ability dimension and be conditionally independent of other items given this ability. Violations of local item independence, local item dependence (LID) may have such effects as underestimating the measurement error in the ability parameter estimation. The consequences of underestimation manifest as less precise ability estimates, which can affect scoring decisions and possibly lead to premature stopping for computerized adaptive tests (CATs).

In sampling examinees from identified subsets of a population not selected at random, person clustering issues may arise in statistical analyses. A simple random sample is different from samples that are taken from specific subpopulations because there is the possibility that examinees from the same group have similarities not intended to be accounted for that are not present in other subgroups in modeling. One such grouping unit is a country in an international assessment. Analysis techniques have been developed to address such natural person clustering. Often in social science and education contexts, a hierarchical (or multilevel) structure is used to appropriately model this local person dependence (LPD) structure where the person residuals may be correlated (e.g., Kamata, 1999, 2001; Raudenbush & Bryk, 2000).

Responses such as those obtained and analyzed via IRT modeling, and a form of product data (item responses), are often the only source of student data used to

make an estimation of a student's ability. Some models include other variables, covariates, to inform estimation or provide improved ways to analyze the data. More recently, process data has been available to the assessment professional where previously mainly product data was used for different measurement purposes (Rupp, Gushta, Mislevy, & Shaffer, 2010). A key development is the widespread usage of computers for use in assessment presentation and data collection. Process data, for example response time (RT) needed for a student to indicate a response selection, is becoming better integrated with product data to provide a more comprehensive view of latent trait estimation.

1.1 Statement of the Problem

Many assessments focus primarily on response accuracy (RA; correct response) with a corresponding ability dimension, and have a secondary dimension identified as the speed of the respondent which incorporates time. The relationship of these two dimensions has been of interest to researchers since early in the development of psychometrics (e.g. Luce, 1986; Thurstone, 1937). Only in the theoretical realm are tests conducted without any time limit. Due to the practical restriction of finite time resources, tests include a time component. Hybrid tests (van der Linden, 2007) is the term used for those tests when time is not a key focus of the assessment. While tests have time limits, the time management to complete the entire test within the allotted time is the province of the individual and their decisions regarding time allotment for each item. Often data on the time elapsed for an individual to respond to single item is not captured. More recently, response time, (RT) the time duration from item presentation to examinee response, is captured

without disruption to the assessment context. For example, the Programme for International Student Assessment (PISA; Organisation for Economic Co-operation and Development [OECD], 2017) and the National Assessment of Educational Progress (NAEP; U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2013) reported response time in their public data files.

Models that incorporate RT have historically been proposed and can be categorized into three classes: RT for psychological or cognition processing analysis, RT in assessment collected for separate analysis, RT collected for combined analysis with responses (Schnipke & Scrams, 2002). Psychometric models for non-speededness with a focus on accuracy have been developed using IRT as fundamental measurement models (Lee & Chen, 2011; Schnipke & Scrams, 2002). In this form of modeling there are further differentiators of RT modeled with RA that leads to three classifications. One class is the models that include RT in IRT modeling. Examples of this modeling include using RT as a covariate (e.g. Wang & Hanson, 2005). A second class are the models in RT that incorporate accuracy (e.g., Lee, 2007). Further, a third class are models that jointly consider response accuracy and RT (e.g., Klein Entink, Fox, & van der Linden, 2009; van der Linden, 2007).

The hierarchical model for responses and RTs (van der Linden, 2007, 2006) jointly models the response accuracy and response time. This model estimates RA and RT related latent parameters underlying responses and RT data and is considered to be “the most promising approach within IRT” (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011, p. 351). In a study that compared response and RT

models, Suh (2010) found that the hierarchical model, which is a joint modeling approach to associate the response and RT models, was the best performing when compared to the Thissen (1983) model of RT that includes a RA component, and Wang and Hanson (2005) model for RA that includes a RT component. Contributions to the hierarchical framework have been developed by Klein Entink, Fox, and van der Linden (2009) in the multilevel multivariate accuracy and response time model. This model addresses LPD by adding groups to the multilevel structure.

While items may exhibit LID for various reasons, the presentation method in some assessments to create multiple items based on one common stimulus is a common assessment format that induces LID. A reading passage that has more than one question or a math chart with multiple questions referring to it are just two examples of how items may be intentionally clustered in assessment. These item bundles are known as testlets (Wainer & Kiely, 1987). Various models in IRT have been developed to accommodate the testlet effects (e.g., Bradlow, Wainer, & Wang, 1999; Jiao, Wang, & He, 2013; Jiao, Wang, & Kamata, 2005; Jiao, Kamata, & Xie, 2016; Wainer, Bradlow, & Wang, 2007; Wainer, Bradlow, & Du, 2000; Wang & Hanson, 2005).

Enhancements in modeling for IRT have been addressing the issues of LID and LPD for decades. The joint modeling of IRT and RT is relatively recent; it bears the requirement collection of, or access to, datasets where the requisite product and process data was captured. A challenge facing the practitioner with access to such data may be at the other end of the scale where they have significant amounts of process data, but not a viable approach to analyze it. The sophistication of modeling

in the IRT space has not yet been adopted to joint response and RT modeling accounting for real-world testing scenarios.

1.2 Purpose of the Study

This study models responses and RTs jointly when dual local dependencies are present. The new multilevel testlet joint model (MTJM) extends van der Linden's (2007) hierarchical framework for various situations and item presentations. The proposed work differs from previous models that have been developed on the foundation of the hierarchical framework (e.g., Im, 2015; Klein Entink, et al, 2009), the proposed model addresses the issues of manifest LID and LPD simultaneously. To account for the potential presence of local item dependence due to item clustering, a testlet structure is incorporated into the model. To account for local person dependence due to person clustering, a multilevel model structure for groups and persons is implemented.

The study investigates the following research questions:

1. How do the manipulated study factors (the number of items, the magnitude of the testlet effect, the magnitude of the group effect, and the correlation between person parameters in terms of response accuracy and speed) affect the proposed model parameter estimates?
2. How do violations of local person independence and local item independence affect parameter recovery when fitting the data with standard joint models of response and RT that ignore either the person clustering effects, the item clustering effects, or both of these effects?

3. How does model selection using a Bayesian model fit index perform for the proposed model compared to alternative competing models for joint modeling of responses and RTs when LID and/or LPD is ignored in simulated and empirical data analysis?

The estimation of model parameters is developed within a Bayesian framework using Markov Chain Monte Carlo (MCMC) techniques. There are three comparisons planned. One, to evaluate the effect of the LPD, the proposed model is compared to the hierarchical framework model incorporating testlet effects (Im, 2015), but does not include person grouping. Two, a comparison of the proposed model is made with a multilevel hierarchical framework model (Klein Entink, et al, 2009), a model that does not incorporate testlets for the effect of LID. Three, a model comparison is made to evaluate the proposed model in contrast to the van der Linden (2007) hierarchal framework that does not account for either LPD or LID. An empirical data set from an assessment using testlets with students nested within groups (PISA 2015) is used to demonstrate the application of the proposed model.

1.3 Significance of the Study

This study models dual dependency by proposing a model that more closely approximates the circumstances encountered in several large-scale assessments. There are three benefits that are achieved by the use of the proposed model. The benefits are as follows: one, modeling of the more complex real-world situation, two, statistical estimation improvement, and three, Bayesian modeling methods development.

Regarding advantage one, complex sampling may result from natural groupings as they occur. For example, in a national assessment program, students within schools can be sampled for analysis of student performance nested within schools, or an international study could be conducted for countries and used to look at group level performance across countries. A significant cost-savings may be obtained by sampling these naturally formed groups instead of using simple random sampling. Regarding innovative item developments, testlets (Wainer & Kiely, 1987) as a presentation unit of assessments is widely employed. These item clusters are found in large-scale assessments such as TOEFL, PISA, PIRLS, and NAEP. Testlets can be used in testing across varied subject matter, for example, in language assessment using reading passages; mathematics including data tables; and in science assessment with graphs and figures.

Advantage two is statistical improvement, specifically, parameter estimation and model fit. It is expected that the proposed model will show improved parameter estimation when the assumptions of local item independence and local person independence are violated. The model is expected to reduce the error associated with the parameters of interest through variance partitioning in both the person and the item facets. The model fit refers to the consistency between the data and the fitted model. It is assumed that the proposed more complex model will fit the best when the data that are generated from a similarly complex situation is fitted with the proposed model. Model fit indices that include terms to reward parsimony are anticipated to provide results showing the proposed model better fits the data than the models used for comparison.

Advantage three is in Bayesian estimation methodology. The model proposed, and the accompanying description of its development, may assist practitioners in applying more complex statistical modeling as appropriate to the data structure of their investigations. The planned use of widely available and free statistical programming software R and Bayesian estimator program JAGS support the capability to disseminate the gains from conducting this study.

1.4 Overview of the Chapters

The proposal is organized as follows. Chapter 2 contains the literature review. The chapter includes the discussion of the creation and establishment of joint response and response time modelling. The proposed model is situated within an historical context. Specific models are presented to illustrate key developments in response and RT modeling, the hierarchical framework, and alternative models. Chapter 3 contains the methodology – this chapter presents the proposed model and investigation. The chapter describes the manipulated factors, fixed factors, model constraints, model assumptions. In the chapter, the study design, the simulation of data sets, and the empirical dataset are detailed. The Bayesian methods, prior distributions, hyperprior distributions, and estimation programming are also discussed. A brief conclusion chapter is provided at the end to summarize the proposed study and highlights potential extensions in the future.

Chapter 2: Literature Review

This chapter provides discussion of well-known and frequently applied models in IRT and RT. For IRT modeling of product data, an incorporation of RT process modeling is presented. Likewise, the converse relationship, RT models that include IRT modeling, are reviewed. Models that jointly estimate the response and RT parameters are next addressed. The chapter concludes with a summary of key Bayesian modeling concepts and implementations.

2.1 Item Response Modeling

2.1.1 Standard IRT Models

The hallmark of IRT, also known as latent trait theory, is the conceptual approach, borne out in statistical modeling, where the unobserved phenomenon of interest is measured using observed responses to a series of items. The latent variable in assessment contexts is often achievement, where it is frequently referred to as the ability parameter in the IRT model. This parameter is not knowable through direct observation, so a statistical model is employed to estimate the value. An advantage of IRT over its predecessor, now referred to as classical test theory (CTT), is that the estimation of the latent ability is not restricted by the requirement to have a fixed set of items, as the term “test” in CTT would imply. Items that are identified for inclusion in an assessment may be calibrated and included without the need to composing a fixed set. This has led to important developments such as computer adaptive testing (CAT) where the items are drawn from an item pool based on their psychometric characteristics and the assessment structural design.

The IRT statistical model is based on linear regression. Specifically, logistic regression is an appropriate form for modeling the response when an item is scored 1 if correct and 0 if incorrect, or the response is identified as a member of discrete categories. As a statistical model, IRT has associated assumptions. Two fundamental assumptions are unidimensionality and local independence (Reckase, 2009). Unidimensionality is the term to identify that the latent trait of interest is the only trait contributing to the measurement of the respondent, That is, there is only one dimension that contributes to the variance of the respondents' performance (Lord & Novick, 1968; Rasch, 1960). In practice, the truth or falsity of this assumption is unknowable. Statistical tests and models have been developed to determine whether explicit violations of the unidimensionality assumption have occurred, and the consequences of its violation (e.g., Bolt, 1999; Camilli, Wang, & Fesq, 1995). Deliberate relaxation of the unidimensionality assumption has been developed in multidimensional IRT models where there is more than one dimension of interest for measurement (e.g., Mulaik, 1972; Reckase, 1972, 2009; Sympson, 1978).

The assumption of local independence requires that the probability of a correct response to an item should have no statistical relationship to that of another item, and no respondent should have an effect on the probability of a correct response for another respondent beyond that accounted for by the parameters. These two elements of the location independence assumption are known respectively as local item independence and local person independence, respectively. The violation of the item element of the assumption is termed local item dependence (LID). The causes of

LID that have been identified include speededness, practice, testlet dependence, and item chaining (Yen, 1984, 1993).

Local person dependence (LPD) addresses the relationship of the respondents. This violation is often encountered in group sampling situations where the respondents are not sampled randomly over person groupings that affect performance. The investigation and assessment of a joint model of responses and RT that addresses LID and LPD is the subject of this study. Specific IRT models are next discussed to highlight the key developments in this univariate modeling; three standard models, Rasch (1960), 2-PL (Birnbaum, 1968) and 3-PL (Birnbaum, 1968) are presented followed by the testlet and multilevel extensions of these standard models.

The Rasch (1960) or the 1-PL model is a foundational IRT model where there is a single item parameter in the modeling relationship. An advantage of the model over earlier models for assessment is that it locates ability and difficulty on the same scale. The model is mathematically expressed as follows:

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

where y_{ij} represents the observed correct response on item *i* for person *j*, θ_j is the latent ability of person *j*, b_i is the item location (difficulty) where, for this logistic function, the success probability is .50.

In the Rasch model, the shape of the logistic curve is the same for all items. To address the slope of the curve as well as location, Birnbaum (1968) developed the 2-PL model which includes a discrimination parameter, a_i. This parameter allows for the scaling of the logistic curve to vary per item. The greater the discrimination value,

the more distinctly an item provides separation for the probability of a correct response around the inflection point for persons of different ability. The 2-PL model is represented as

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

In the IRT models previously discussed, the upper and lower asymptotes of the logistic function probability values tend to 1 for asymptotically increasing ability and to 0 for asymptotically decreasing ability. Education assessments often include multiple-choice items where there is a non-zero probability of correctly responding to an item due to chance. Birnbaum (1968) included a parameter in the IRT model to shift the lower asymptote of the logistic function to account for guessing. Called the pseudo-guessing parameter, c_i raises the lowest probability of correct response to the value of c_i. The item difficulty is no longer the location on the scale where the probability of a correct response is .50, but where the probability of such as response is (1+ c_i) / 2. The 3-PL model is specified as

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

In some IRT modeling contexts, the link function of the probit (normal ogive) is more readily applied than the logistic. The 3-parameter normal ogive (3PNO) and more restricted form (2PNO) are alternatives to the 3-PL and 2-PL IRT model. The latent variables and parameter interpretations in the 3PNO model are the same as that of the 3-PL IRT model. With the discussion of the three standard models of IRT concluded, models that use these foundations to address LID or LPD are discussed in the next section.

2.1.2 IRT Model Extensions – Testlet Models

The IRT Rasch, 2-PL and 3-PL are ubiquitous in unidimensional latent modeling. The modeling of local item dependencies by incorporating a testlet design in the previously discussed standard IRT models followed three decades later with Bradlow, Wainer, and Wang (1999) who proposed the 2-PL testlet response theory (TRT) model. The 3-PL testlet model was further proposed by Wainer, Bradlow, and Du (2000). As a special case of the multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997), the Rasch testlet response model was specified by Wang and Wilson (2005) which is mathematically equivalent to a multilevel model by Jiao, Wang and Kamata (2005) for item clustering effects due to testlets. In each of these approaches, a parameter is specified in the IRT model to account for within-testlet LID. The Rasch testlet model is mathematically presented as

$$P\left(y_{ij}=1|\theta_j,b_i,\gamma_{jd(i)}\right)=\frac{1}{1+\exp\left[-\left(\theta_j-b_i+\gamma_{jd(i)}\right)\right]}, \qquad (2.3)$$

where the testlet parameter $\gamma_{jd(i)}$ denotes the effect of a local effect of dependence for items within a testlet $d(i)$ for person j, θ_j is the latent ability parameter for person j, and b_i is the item difficulty for item i. The LID magnitude is characterized by σ_γ^2. The identification of the variability contributed to the testlet is therefore straightforward with this formulation. The description of the modeling used for incorporating a complex sampling design with LPD is more involved than that for LID, consequently as the next section is on the topic of LPD there will first be a

general example to describe this perspective followed by specific approaches to its modeling.

2.1.3 IRT Model Extensions – Multilevel IRT Models

Generally, a 3-level model may be developed by identifying a level-1 where observations are made, level-2, the cluster in which the level-1 units are nested, and level-3, the group in which the level-2 units are nested. In the example of a school setting, students are nested within a classroom and classrooms are nested within schools (Raudenbush & Bryk, 2002). A model with no predictors, that is, one that is fully unconditional, indicates how variation in the observed measure is dispersed across the three different levels. The following example uses Raudenbush and Bryk (2002) notation for a 3-level multilevel model. In this model, level-1 student achievement is

$$Y_{ijg} = \pi_{0jg} + e_{ijg} \tag{2.4}$$

where Y_{ijg} refers to the achievement of student *i* in classroom *j* in school g, π_{0jg} is the mean of student achievement in classroom *j*, and e_{ijg} is a random effect, the deviation from the classroom mean for student *i* in classroom *j* school g. The level-2 classroom model presents each classroom mean as a random effect of a school mean:

$$\pi_{0jg} = \beta_{00g} + r_{0jg}. \tag{2.5}$$

The mean achievement for school g is β_{00g}, and r_{0jg} is the random effect for classrooms which are assumed normally distributed with mean 0 and variance σ_{π}^{2}. The level-3 model is the school-level model that shows the variability between schools:

$$\beta_{00g} = \gamma_{000} + u_{00g}, \tag{2.6}$$

where γ_{000} is the grand mean across all schools, u_{00g} represents a random effect for schools the effects are assumed normally distributed with mean 0 and variance σ_{β}^{2}.

Multilevel models in IRT that can account for LPD due to person grouping were developed by Fox and Glas (2001), Kamata (2001), and Maier (2001). These models employ the hierarchical linear model (Bryk & Raudenbush, 1992, Goldstein, 1987) or a generalization (Raudenbush, 1995) as the modeling framework. The Fox and Glas (2001) model is first presented followed by Kamata (2001) to highlight different approaches in the area of IRT multilevel modeling and parameter specification.

Fox and Glas (2001) developed the multilevel IRT model, which has a focus on the structural, not the measurement model (Fox, 2007). The measurement model, level-1 is a 2PNO with a link function of the standard normal cumulative distribution function Φ, the model parameters are the same as those of the 2PL IRT with the addition of *j* students nested in g groups. The level-1 model is

$$P\left(y_{ijg} = 1|\theta_{jg}, a_i\ b_i\right)\ = \Phi\left(a_i\theta_{jg} - b_i\right). \qquad (2.7)$$

Level-2 is the person level of the structural model, where the ability parameter θ_{jg} is the dependent variable. The intercept β_{0g} represents the student mean within a cluster, covariates **x** are included in the model with the associated coefficients $\beta_{Qk}$where Q is the number of covariates. The model is presented as:

$$\theta_{jg}\ = \beta_{0g} + \ \beta_1 x_{1jg} + \cdots + \beta_{qg} x_{qjg} + \ \cdots + \ \beta_{Qg} x_{Qjg} + \ e_{jg}. \qquad (2.8)$$

where $e_{jg} \sim N(0, \sigma^2)$.

The Level-3 model is the group level of the structural model. Each of the coefficients in the Level-2 model are treated as dependent variables with their own

intercepts and covariates. The intercepts are the student population mean, denoted as $\boldsymbol{\gamma}_{Q0}$, the covariates **w** at this level have associated coefficients $\gamma_{Qk}$where *S* is the number of covariates. There may be several models at Level 3, a representation of each type, intercept, slope, and sum, respectively, are as follows:

$$\begin{aligned} \beta_{0g} &= \gamma_{00} + \gamma_{01}w_{1g} + \cdots + \gamma_{0S}w_{Sg} + u_{jg} \\ \beta_{1g} &= \gamma_{10} + \gamma_{10}w_{1g} + \cdots + \gamma_{1S}w_{Sg} + u_{1g} \\ &\vdots = \vdots \\ \beta_{Qg} &= \gamma_{Q0} + \gamma_{Q1}w_{1g} + \cdots + \gamma_{QS}w_{Sg} + u_{Qg}, \end{aligned} \tag{2.9}$$

where $\mathbf{u}_j \sim N(0, \mathrm{T})$. The multilevel modeling approach has the advantage of simultaneously estimating all model parameters, unlike methods that use fixed parameters, the uncertainty in measurements is included in the multilevel IRT modeling estimation (Fox, 2004). Another framework developed for incorporating LPD, albeit one that gives equal status to the development of the measurement model, was developed by Kamata (2001).

For the framework of the hierarchical generalized linear model (HGLM; Kamata, 2001) various link functions are afforded such as probit, complementary log-log, and logit. The logit link function is predominantly used in IRT modeling and was detailed in Kamata (2001) to show that for models that include levels 1 and 2, they are equivalent to the Rasch (1960) IRT model.

Kamata (2001) provides two forms of notation for the model to include covariates. As it is similar in form to the notation used by Fox and Glas (2001), where a choice is possible, the more compact notation is described. The model assumes

Bernoulli sampling for item i and person j within group g. Therefore, the item responses y_{ijg} have expected mean and variance for a correct response:

$$E(y_{ijg} = 1|p_{ijg}) = p_{ijg} \text{ and} \tag{2.10}$$

$$Var(y_{ijg} = 1|p_{ijg}) = p_{ijg}(1 - p_{ijg}),$$

where p_{ijg} is the probability of a correct response. By employing this probability, level-1 is the item-level model where the log-odds (logit) of p_{ijg} is η_{ijk} and

$$\log\left(\frac{p_{ijg}}{1 - p_{ijg}}\right) = \eta_{ijg} = \beta_{0jg} + \sum_{q=1}^{k-1} \beta_{qjg} x_{qijg}\,. \tag{2.11}$$

The Kamata (2001) model includes an intercept β_{0jk} and linear predictors with associated coefficients, β_{qjk}. These predictors are dummy variables where x_{qijk} is the qth dummy variable for person j in group g, with values for $q = i$, 1 and $q \neq i$, 0 for item i. To achieve full rank, an item dummy variable, usually the last one, is selected to be dropped from the equation. Due to this, the number of dummy variables q is indexed q = 1, …, k-1. Interpretation of the intercept differs from other multilevel models as the dropped coefficient is now associated as a reference item. The individual item effect β_{qjk} is therefore established as the deviation from the effect β_{0jk}. The simplified model for a correct response of person j in group g on item i is

$$P(y_{ijg} = 1) = \frac{1}{1 + \exp(-\eta_{ijg})}. \tag{2.12}$$

Level-2 is a person-level model where the item effects β_{qjg} are decomposed into a fixed component and a random component; the random component u_{0jg} is the person effect. This level closely aligns with the structural level-3 of the multilevel model previously described in this section. To emphasize the combined model

presentation, the treatment will not include additional specific predictors at level-2 or level-3. The Level-2 representations of the model are

$$\beta_{0jg} = \gamma_{00g} + u_{0jg} \tag{2.13}$$

$$\beta_{1jg} = \gamma_{10g}$$

$$\vdots = \vdots$$

$$\beta_{(k-1)jg} = \gamma_{(k-1)0g},$$

where $u_{0jg} \sim N(0, \sigma_w^2)$ and w is the index for the ability distribution. The combined model for level 1 and 2 for a person j correctly responding to item i in group g with a probability p_{ijg} of a correct response is

$$p_{ijg} = \frac{1}{1 + \exp\{-[u_{0jg} - (-\gamma_{q0g} - \gamma_{00g})]\}}, \tag{2.14}$$

where $i = q$. This equation is mathematically equivalent to the Rasch model. The relationships are as follows: the ability parameter and the random effect for persons have a one-to-one correspondence, $\theta_{jg} = u_{0jg}$, and the item difficulty parameter b_i has been parsed into components. That is, $b_i = -\gamma_{q0g} - \gamma_{00g}$, where each is a fixed parameter.

The level-3 model is the group-level model. The specification is the same as that of the level-2 model with exception of the addition of parameters to identify group effects. The level-2 person intercept γ_{00g} is modeled at the group-level with a group intercept and a random effect r_{00g} for variability between groups. The $r_{00g} \sim N(0, \sigma_\pi^2)$ and π is the index for the group ability distribution. The models for groups g are

$$\gamma_{00g} = \pi_{000} + r_{00g} \tag{2.15}$$

$$\gamma_{10g} = \pi_{100}$$

$$\vdots = \vdots$$

$$\gamma_{(k-1)0g} = \pi_{(k-1)00}.$$

When the combined model is presented following dummy coding at level-1,

$$p_{ijg} = \frac{1}{1 + \exp\{-[(r_{00g} + u_{0jg}) - (-\pi_{q00} - \pi_{000})]\}}, \qquad (2.16)$$

the similarities with the Rasch model remain, with some caveats. As in the 2-level model, there are two components for the item difficulty. In the level-3 model, $-\pi_{q00} - \pi_{000}$ expresses the item difficulty for items (excepting the reference item where the difficulty is π_{000}). Ability is also provided by two components at this level of the model, r_{00g} and u_{0jg}. The random effect for groups r_{00g} may be considered the average ability of persons in group g. The random effect for persons u_{0jg} is interpreted as a person-specific ability of persons j in group g. This model therefore provides the advantage of separating contributions of ability effects by groups and by persons.

An alternative representation of this model that emphasizes the correspondence with the Rasch model was described by Jiao, Kamata, Wang, and Jin (2010) as

$$P(y_{ijg} = 1|\theta_{jg}, \theta_g, b_i) = \frac{1}{1 + \exp[-(\theta_{jg} + \theta_g - b_i)]} \qquad (2.17)$$

where θ_{jg} denotes the ability of persons j in group g, and θ_g represents the ability for group g.

2.1.4 IRT Model Extensions – Multilevel Testlet Models

There are a few testlet models that include a consideration for person clustering. Such a model was developed by Jiao, Kamata, Wang, and Jin (2010, 2012). In this model a multilevel structure for groups and a testlet model for items were incorporated in the RA model to address both LID and LPD simultaneously. The multilevel structure they described was adapted from the multilevel model (Kamata, 2001) and the model extension for testlets (Jiao, Wang, & Kamata, 2005).

In this dual local dependence model, four levels are specified (Jiao, Kamata, Wang, & Jin, 2010). The level-1 model provides an expression of the relationship in a linear regression with intercept term of persons j with items i, for these items nested in testlet d. The level-2 model includes a testlet effect where the random effect is analogous to the testlet parameter included in testlet IRT models (e.g., Bradlow et al, 1999; Wainer et al, 2000). The level-3 model presents a person effect where the random effect is the person ability as in the previously discussed IRT models. The level-4 model establishes a person group effect where the variability of the group ability is a measure of the impact of the person clustering, a graphical depiction of the model is represented in Figure 1.

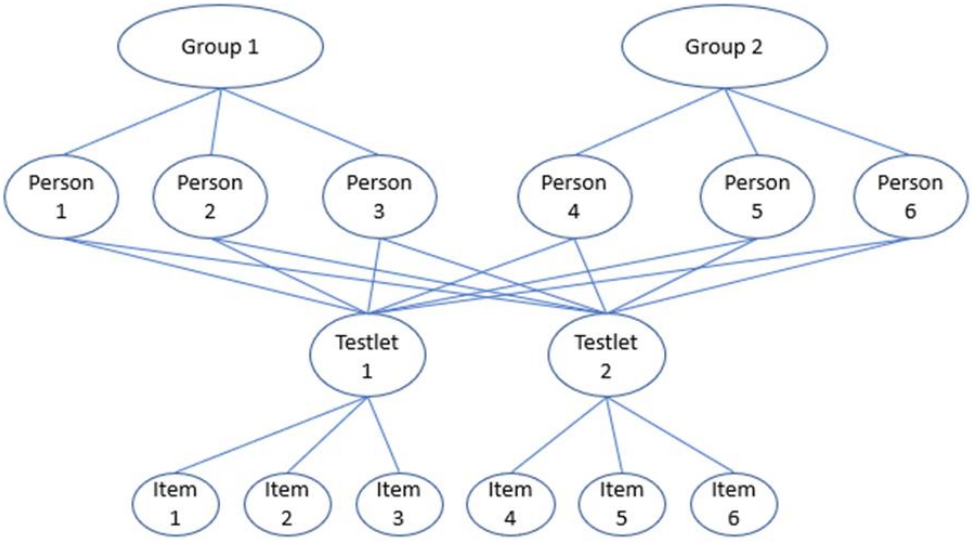

Figure 1. A hierarchical structure of the joint modeling of the multilevel testlet joint model (MTJM) of responses and response time (Jiao, Kamata, Wang, & Jin, 2012).

As the level-2 model is introduced here into the multilevel framework, it is described in more detail. The other levels of the model follow closely with Kamata (2001). The item cluster effect is modeled for person j in group g responding to item i in testlet d as

$$\beta_{0djg} = \gamma_{00jg} + u_{0djg}, \text{and} \tag{2.18}$$

$$\beta_{qdjg} = \gamma_{q0jg},$$

where γ_{00jg} is the fixed effect of the level-1 intercept, and u_{0djg} is a random effect of the level-1 intercept. This random effect may be considered an interaction between testlet and ability. There is an assumption that $u_{0djg} \sim N(0, \sigma_u^2)$. The combined model for all levels is

$$P\left(y_{ijg} = 1 | \theta_{jg}, \theta_g, b_i, \gamma_{jd(i)}\right) = \frac{1}{1 + \exp\left[-\left(\theta_j + \theta_g - b_i + \gamma_{jd(i)}\right)\right]} \tag{2.19}$$

where, as in previously discussed testlet models, the testlet effect has its own parameter, $\gamma_{jd(i)}$, and all other parameters follow the IRT conventions where θ_j is the person-specific ability, $\theta_g$is the group-specific ability, and b_i is item difficulty.

2.2 RT Modeling

The duration of an examinee time elapsed prior to responding has been used as a proxy for ability in early RT models. As conceptualizations of ability and RT have generally become more nuanced, these measures have been disentangled. More recently research has investigated the use of RT in estimation of ability accuracy (e.g., Ferrando & Lorenzo-Seva, 2007; Meng, Tao, & Chang, 2015). Models for measurement of RT and some applications are next discussed.

2.2.1 Standard RT Models

Many observations of RT distributions positively skew (skew to the right) This is good fit for lognormal distribution (e.g., Schnipke & Scrams, 1999; Thissen, 1983), other RT distributions have been observed including symmetrical and negatively skewed. To accommodate these observances, alternative distributions have been incorporated into the modeling of RT.

van der Linden (2006) described the lognormal RT model

$$\ln(t_{ij}) = \beta_i - \tau_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_i^{-2}) \tag{2.20}$$

where τ is person speed parameter, β_i is item intensity parameter, α_i is discrimination parameter. The value of α_i is $1/\sigma$ of the log RT error distribution. The β_i item speed and α_i item discrimination parameters of this RT model are analogues to the *b*, difficulty (location) parameter and the *a*, discrimination, and in the 2-PL IRT model.

The Box-Cox normal model (Klein Entink, van der Linden, and Fox, 2009) identifies a 2PNO for the RA model. The formulation of the RT model is

$$t_{ij}{}^{(v)} = \beta_i - \tau_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim (0, \alpha_i^{-2}). \tag{2.21}$$

T^v is the Box-Cox transformed time with that distribution's shape parameter v, τ_j is person speed, β_i is item intensity, α_i is discrimination parameter (1/σ of log RT). In Box and Cox (1964) T^v is transformed from T^v-1 / ^{v} for v ≠ 0 and log T for v=0. T is the original time, therefore v=0 is a lognormal transform as in van der Linden (2006). The shape parameter may be applied with values differing by item, or by setting the value to be the same for all items. The fit for each item may be improved by modeling the differing distribution's shapes, but this improvement in fit comes with the disadvantage that the parameter interpretation may no longer be the same across items. The shape parameterization provides a family of RT shape approximating data from Weibull, gamma, and exponential models (Klein Entink, van der Linden, and Fox, 2009).

The Semiparametric Cox proportional hazards (PH) model (Ranger & Ortner, 2012; Wang, Fan, Chang, & Douglas, 2013) is an alternative model that addresses the concerns raised for the Box-Cox RT model regarding interpretability. The model includes a hazard function, which is often represented as *h((t))*, that specifies an event's instantaneous rate of occurrence. In testing contexts, the hazard rate is the conditional probability of finishing a task in the next moment. The model is

$$h_i(t_{ij}|\tau_j) = h_{0j}(t_{ij}) \exp(\gamma_i \tau_j) \tag{2.22}$$

where t_{ij} is RT, $h0j$ $(\cdot)$ identifies the baseline hazard function, τ_j denotes the speed parameter for person j, γ_i is an item slope parameter that determines the increase in hazard rate.

When the distributions for the normal, lognormal, gamma, and Weibull were evaluated for fit, Schnipke and Scrams (1999, 2002) found the lognormal distribution to be best fitting in exploratory and confirmatory settings

2.2.2 Incorporating RT for Modeling RA

Following the introduction to standard RT models, the topic of interest regarding the phenomenon of the speed-accuracy tradeoff is presented. The discussion will first include RA models that incorporate RT.

The Rasch RT model, (Roskam, 1987, 1997) implemented the observed RT as collateral information for insertion into the IRT model. The model is specified as

$$P\left(y_{ij} = 1 \middle| \theta_j, t_{ij}, b_i\right) = \frac{\theta_j t_{ij}}{\theta_j t_{ij} + b_i} = \frac{\exp\left(\theta_j * + t_{ij} * - b_i *\right)}{1 + \exp\left(\theta_j * + t_{ij} * - b_i *\right)} \qquad (2.23)$$

where θ_j is person ability, here termed mental speed, b_i is item difficulty, and t_{ij} is the RT, the * parameters are the respective logarithms of the person ability, item difficulty, and RT; the product of mental speed and response time is termed effective ability. The speed-accuracy tradeoff is specified by the positive relationship where with increasing time the probability of a correct response tends to 1.0. This model is therefore appropriate for speed tests where the assumption is with infinite time respondents will correctly respond to all items.

In a Rasch-like model, where speed is included as a latent parameter, Verhelst, Verstralen, and Jansen (1997) introduced a model where the distribution of

the RT could vary. The p_i item parameter allows the shape of RT distribution to change; the model is

$$P(y_{ij} = 1|\theta_j, \tau_j, b_i) = \left\{\frac{1}{1 + \exp[-(\theta_j + \tau_j - b_i)]}\right\}^{-p_i} \quad (2.24)$$

where θ_j is person ability, b_i is item difficulty, and τ_j person speed. The model includes a latent variable for RT, speed, as compared to an observed *t*. The distribution of RT is a combination of a generalized extreme-value distribution and a gamma distribution.

Wang and Hanson (2005) developed a 4-PL model that includes the 3PL IRT and 1 new parameter, $-\rho_j d_i / t_{ij}$ where $\rho_j$is a person slowness parameter, d_i, is an item slowness parameter and t_{ij} is RT

$$P(y_{ij} = 1|\theta_j, a_i\, b_i, c_i) \;=\; c_i + \frac{1 - \; c_i}{1 + \exp\left[-a_i\left(\theta_j\left(\frac{\rho_j d_i}{t_{ij}}\right) - b_i\right)\right]}. \quad (2.25)$$

The contribution of the parameters for "slowness" each has the same assumed effect on the probability of a correct response. With increasing time, the probability of a correct response approaches that found in the 3-PL IRT model. It is therefore identified as a model that is appropriate for hybrid tests.

2.2.3 Incorporating RA for Modeling RT

For the focus on the RT distributions, RA has been used to improve estimation. Early development was the Thissen (1983) model

$$\ln(t_{ij}) \;= \mu + \tau_j + \beta_i - \rho(a_i\theta_j - b_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad (2.26)$$

where μ overall mean log RT, θ_j "effective ability", b_i is item difficulty, a_i is item discrimination, τ_j is person "slowness", β_i is item "slowness", ρ is a regression slope

parameter of log RT on the IRT model (2-PL). Interpretation of the "slowness" parameters in Wang and Hanson (2005) differs from this model. The two models reflect different goals, where Wang and Hanson (2005) are modeling response accuracy probability and Thissen (1983) is modeling RT.

The Ferrando and Lorenzo-Seva (2007) model is similar to Thissen (1983), the regression is on square root of squared IRT parameters modeled $\sqrt{a_i^2(\theta_j - b_i)^2}$. Developed in personality measures context, informed by distance-difficulty hypothesis, Ferrando and Lorenzo-Seva (2007) and Thissen (1983) each models a speed-accuracy tradeoff. The correlation parameter ρ represents the direction of the tradeoff, with a positive value for the parameter indicating an increase in τ_j associated with an increase of θ_j, and a negative value indicating the reverse relationship.

Gaviria (2005) presented the double log-normal distribution model that is represented as

$$ln\left(\frac{t_{ij} - T_0}{A}\right) = -a_i(\theta_j - b_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim lognormal(0, \sigma_i^2). \qquad (2.27)$$

T_0 is time taken by the person on an infinitely easy item, A is a scaling constant for RT where a respondent's ability coincides with the item's difficulty, θ_j is person ability, b_i is item difficulty, a_i is item discrimination, ε_{ij} is a residual. The model is posited for correct responses.

2.3 Joint Modeling of RT and RA

In the standard IRT models presented, the assumption was made that the level of a person's ability did not vary during the assessment. For the conditionally independent joint response and RT models, this assumption is also held for person

speed (e.g., Goldhammer & Kroehne, 2014; Meng et al., 2015; van der Linden, 2009). Based on these assumptions, the within-subject relationship of the respondent with the items does not vary during the test.

The joint distribution of RT and RA can be expressed as follows:

$$f(y_{ij}, t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) \tag{2.28}$$

where y_{ij} is the response of person j for item i, t_{ij} is the RT associated with response y_{ij}, θ_j and τ_j are respectively the latent ability and the latent speed parameters, β_i represents the item parameters in the IRT model and λ_i represents the item parameters in the RT model.

This section addresses three different approaches to modeling the joint distribution of RT and RA (Ranger & Ortner, 2012). In this their classification, one group of models are conditionally independent, a second group of models have dependency within the response model, and a third group of models have dependency within the RT models. In the first classification, there are two marginals, here the assumption is the responses and RT are conditionally independent (e.g., Thissen, 1983); van der Linden, 2007)

$$f(y_{ij}, t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) = f(y_{ij}|\theta_j, \tau_j, \beta_i) f(t_{ij}|\theta_j, \tau_j, \lambda_i). \tag{2.29}$$

In the second classification, a conditional and a marginal probability relationship is specified; here for an observed response y_i a dependency exists on the RT t_{ij} spent on this item. In the third classification, a conditional and a marginal are again specified, in this case, where the reverse dependency is modeled; an observed RT t_{ij} depends on the response y_i on this item.

Thissen (1983) and van der Linden (2007) each chose to model the relationship of response and RT to reflect that they are conditionally independent of each other

$$f(y_{ij}, t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) = f(y_{ij}|\theta_j, \beta_i) f(t_{ij}|\, \tau_j, \lambda_i). \quad (2.30)$$

This assumption indicates that responses are dependent on the IRT ability and item parameters, and RTs are dependent on the RT model speed and item parameters. With these being the only dependencies, the responses and RT are therefore conditionally independent of each other. Other models that require this assumption include, for example, Klein Entink, van der Linden, and Fox (2009) and Wang, Fan, Chang, and Douglas (2013).

van der Linden's (2007) hierarchical framework is often applied for joint response and RT modeling. Level-1 for measurement models, one for IRT and one for RT. At this level, the observed response or RT is associated with latent model parameters. Level-2 is the population modeling level where a covariance structure allows the modeling of relationship for person and item parameters rather than having it dictated by a specific speed-accuracy tradeoff function.

The IRT model chosen was the three-parameter normal-ogive (3PNO), the selection of the RT model was the lognormal model (van der Linden, 2006). The second level specified a multivariate normal distribution for the person parameters and another multivariate normal distribution for the item parameters. For the person parameters, the mean vector and covariance matrix are identified with subscript *P* and specified as:

$$\boldsymbol{\mu}_P = (\mu_\theta, \mu_\tau), \quad (2.31)$$

$$\boldsymbol{\Sigma}_P = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\tau\theta} & \sigma_\tau^2 \end{pmatrix}, \tag{2.32}$$

The mean vector and covariance matrix for item parameters are subscripted *I*:

$$\boldsymbol{\mu}_I = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta) \tag{2.33}$$

$$\boldsymbol{\Sigma}_I = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix} \tag{2.34}$$

Built upon this framework, several response and RT models are developed by changing the RT model for another distribution (Klein Entink, van der Linden, & Fox, 2009) or making a level-1 change for another measurement model, such as, the Cox PH model (Ranger & Kuhn, 2014a; Wang, Fan, et al., 2013), multilevel for person groups (Klein Entink, Fox, & van der Linden, 2009), multivariate (Fox, Klein Entink, & Timmers, 2014), and item clustering (Im, 2015).

A model of the first version of conditional independence not assumed is

$$f(y_{ij}, t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) = f(y_{ij}|t_{ij}, \theta_j, \tau_j, \beta_i, \lambda_i) f(t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i). \tag{2.35}$$

In its simplified form, it is represented as

$$f(y_{ij}, t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) = f(y_{ij}|t_{ij}, \theta_j, \tau_j, \beta_i) f(t_{ij}|\, \tau_j, \lambda_i). \tag{2.36}$$

Models that employ this dependency address item characteristics such as item difficulty (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2017). The model developed in that research incorporated a median split to identify a fast or slow response. They found that slower responding was associated with greater probability of a correct response on difficult items, but that slower responding was associated with lower probability of a correct response on easy items.

To investigate variable ability and speed, Partchev and De Boeck (2012) used a branching model for evaluation of fast or slow intelligence. A mixture model was applied for high-stakes and low-stakes assessments when pace may differ due to the nature of the assessment (rapid-guessing, Wang & Xu, 2015). Lee and Wollack (2017) used the information in their model to determine whether the respondents had item pre-knowledge.

The second form of conditional independence not assumed in its full version is

$$f(y_{ij}, t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) = f(y_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) f(t_{ij}|y_{ij}, \theta_j, \tau_j, \beta_i, \lambda_i). \quad (2.37)$$

This model may also be simplified where it takes the form

$$f(y_{ij}, t_{ij}|\theta_j, \tau_j, \beta_i, \lambda_i) = f(y_{ij}|\theta_j, \beta_i) f(t_{ij}|y_{ij}, \tau_j, \lambda_i). \quad (2.38)$$

This conception applies to differing models for the RTs associated with correct and incorrect responses (e.g., Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; Glas & van der Linden, 2010; and van der Linden & Glas, 2010). Results from these studies often encounter issues of conditional independence violations.

2.3.1 Multilevel Models for Person Clustering

The multivariate multilevel model for response and RT (Klein Entink, Fox, & van der Linden, 2009) provides three significant changes that extend the model of van der Linden (2007). First, the model introduces a time discrimination parameter within the RT measurement model. Second, covariates are enabled for inclusion by the specification of the regression equations. Third, the model includes a third level so that person grouping variance may be accounted for. The model at level one applies a 3PNO for responses and the RT measurement model:

$$t_{ijg} = \beta_i - \varphi_i \tau_{jg} + \varepsilon_{\tau ijg}, \quad \varepsilon_{\tau ijg} \sim N(0, \sigma_I^2) \tag{2.39}$$

the additional subscript g is used as there is a need to identify the group membership of persons, τ is person speed parameter, β is item intensity parameter, φ is the time discrimination parameter.

In level-2 of the model, as in van der Linden (2007), two covariate matrices are employed to model the relationships of the person parameter with other person parameters; this model also supports the use of additional covariates. The regression equations written in matrix notation are:

$$\theta_{jg} = \mathbf{x}_{jg}^t \boldsymbol{\beta}_{1g} + e_{\theta jg}, \tag{2.40}$$

$$\tau_{jg} = \mathbf{x}_{jg}^t \boldsymbol{\beta}_{2g} + e_{\tau jg}, \tag{2.41}$$

where $\mathbf{x}_{jg}^t$ known covariate vector for all persons *j* in group g, $\boldsymbol{\beta}_g$ vector of regression coefficients for each group g, the error terms, $e_{\theta jg}$ and $e_{\tau jg}$ can correlate. Error terms are assumed to follow a bivariate normal distribution with means and covariance matrix specified as follows:

$$\boldsymbol{\mu}_P = (\mu_\theta, \mu_\tau), \tag{2.42}$$

$$\boldsymbol{\Sigma}_p = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\tau\theta} & \sigma_\tau^2 \end{pmatrix}. \tag{2.43}$$

For the items, the parameter distributions are assumed to follow are multivariate normal distribution with means and covariance matrix specified as follows:

$$\boldsymbol{\mu}_I = \left(\mu_a, \mu_b, \mu_\varphi, \mu_\beta\right), \tag{2.44}$$

$$\boldsymbol{\Sigma}_I = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\varphi} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{b\varphi} & \sigma_{b\beta} \\ \sigma_{\varphi a} & \sigma_{\varphi b} & \sigma_\varphi^2 & \sigma_{\varphi\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta\varphi} & \sigma_\beta^2 \end{pmatrix}, \tag{2.45}$$

where the vector

$$\boldsymbol{\xi}_I = \boldsymbol{\mu}_I + \boldsymbol{e}_I, \boldsymbol{e}_I \sim N(\mathbf{0}, \boldsymbol{\Sigma}_I). \quad (2.46)$$

As there are no covariates in this presentation, it is simplified, where the four equations for the item parameters are condensed. Therefore, the value of the parameter is the mean of the parameter with random error.

The c_i parameter is not included in the multivariate normal distribution shown as there is not an analogous parameter in the RT model, therefore an independent Beta prior distribution was described.

For the level-3 a group relationship structure is presented:

$$\boldsymbol{\beta}_{1g} = \mathrm{w}_g \gamma_1 + u_{1g}, \quad (2.47)$$

$$\boldsymbol{\beta}_{2g} = \mathrm{w}_g \gamma_2 + u_{2g}. \quad (2.48)$$

Parameters $\boldsymbol{\beta}_{1g}$ and $\boldsymbol{\beta}_{2g}$ are the random effects, where $\mathbf{w}_g$ is a known covariate vector for all groups g. The group-level error terms, (u_{1g}, u_{2g}), are assumed to be multivariate normally distributed with means of zero and covariance matrix **V**. With the assumption of restricting the covariance matrix to block diagonal, the IRT model random effects are independent of RT model random effects.

2.3.2 Joint Testlet Models for RA and RT

Im (2015) described a measurement model for the RA that is the testlet response model with the 2-PL for parameters to adjust the shape of the logistic curve. All parameters are interpreted as in the 2-PL IRT model with the addition of the testlet parameter $\gamma_{jd(i)}$. The model is represented as

$$P\left(y_{ij} = 1 | \theta_j, a_i, b_i, \gamma_{jd(i)}\right) = \frac{1}{1 + \exp\left[-a_i\left(\theta_j - b_i + \gamma_{jd(i)}\right)\right]}. \quad (2.49)$$

The measurement model for the RT follows van der Linden (2007) hierarchical framework with an additional parameter for testlet $\delta_{jd(i)}$. The RT model is

$$f\left(t_{ij}|\tau_j, \alpha_i, \beta_i, \delta_{jd(i)}\right) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_{ij} - \left(\beta_i - \tau_j + \delta_{jd(i)}\right)\right)\right]^2\right\}. \quad (2.50)$$

Priors used were similar to those in van der Linden (2007),

$$\boldsymbol{\mu}_P = (\mu_\theta, \mu_\tau) = (0, 0),$$

$$\boldsymbol{\Sigma}_P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\boldsymbol{\mu}_I = \left(\mu_a, \mu_b, \mu_\alpha, \mu_\beta\right) = (1, 0, 1, 0),$$

$$\boldsymbol{\Sigma}_I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\gamma_d \sim N\left(0, \sigma_\gamma^2\right),$$

$$\delta_d \sim N\left(0, \sigma_\delta^2\right).$$

Hyperpriors for the testlet parameters:

$$\sigma_\gamma^2 \sim U(0, 10),$$

$$\sigma_\delta^2 \sim U(0, 10).$$

This model is the only known one to explicitly address LID using a testlet parameter in joint IRT response and RT modeling.

In their model, which can account for item dependencies, Klotzke and Fox (2019) used a Bayesian covariance structure modeling approach. An additive covariance matrix models dependencies directly through the covariance parameters. An advantage of this method is that the modeling of random effects is not required. By taking this approach, unlike when using random effects, covariances can be

negative or positive. Also, tests for local independence do not encounter boundary conditions that can impede effective measurement. This affords the testing of local within-testlet independence, and the development of more parsimonious models compared to models requiring random effects.

2.4 Model Estimation

Two approaches to model estimation in statistics are the frequentist and the Bayesian. In the frequentist perspective, data sampling is based a hypothetical infinite number of draws from a sample. Parameters are fixed for distributions and data are randomly sampled. In the Bayesian perspective, the parameters are random and described probabilistically. Prior beliefs about the parameters that describe the data are included in the estimation of the values of the priors after data have been observed. The estimation of models in the Bayesian perspective may be more complex and more demanding computationally, compared with the frequentist modeling, due to the ability to model posterior distributions that do not have a closed form.

Software using a frequentist perspective that employs maximum likelihood estimation (MLE) include Mplus (Muthén & Muthén, 2007) and LatentGOLD (Vermunt & Magidson, 2013). Software in Bayesian modeling includes Markov chain Monte Carlo (MCMC) methods and are implemented in JAGS (Plummer, 2015), WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003), OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009), and Stan (Gelman, Lee, & Guo, 2015), to list a few.

A Bayesian modeling approach was selected for this effort for three reasons. One, Bayesian inferences via MCMC methods could be used to sample the complex parameter space of the proposed model that is not readily available with frequentist modeling tools. Two, the parameters each have a distribution and are not assumed to be without error. This better aligns with the investigative goals of appropriately modeling and estimating sources of error. Three, for a complex modeling design, the sample sizes are those that may be expected in real-life scenarios and are not asymptotically large.

2.4.1 Introduction to Bayesian Inference

Bayes' theorem provides the relationship of a prior probability distribution $P(\boldsymbol{\theta})$, the likelihood of the data given the parameters $P(\boldsymbol{X}|\boldsymbol{\theta})$, the marginal probability of the data $P(\boldsymbol{X})$, to provide the posterior distribution of the parameters, given the data. This model is

$$P(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{X})}. \tag{2.51}$$

In the Bayesian modeling approach, data are fixed and parameters are random. $P(\boldsymbol{X})$ is a normalizing constant, so the probability density function integrates to *1*. The contribution of these two facts is that the Bayes' fully specified model may be simplified to

$$Posterior \propto Likelihood \times Prior. \tag{2.52}$$

With this specification, the parameters' posterior distribution given the data is proportional to the product of the likelihood and the prior.

2.4.2 Markov Chain Monte Carlo Methods

Bayesian inference is well-suited to problems where the closed-form solution, which otherwise could have been addressed analytically, does not exist. The Bayesian approach is appropriate for modeling in high-dimensional parameter space. Sampling in Bayesian estimation often uses MCMC methods to build a Markov chain from a probability distribution; this approximates the joint posterior distribution. As the number of iterations increases, the approximation to the posterior distribution improves (Gelman, Carlin, Stern, & Rubin, 2003). The probability of an event following the Markov process is dependent only on the state of the immediately preceding event.

There are three commonly employed MCMC sampling methods used in Bayesian inference. These methods are: the Gibbs sampler (Geman & Geman, 1984), the Metropolis sampler (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), and the Metropolis-Hastings sampler (Hastings, 1970). The Gibbs sampler has the requirement that the model be conditionally conjugate. That is, there is a model structure for known distributions for which the posterior distribution of interest may be parsed. The Metropolis sampler may be used to estimate models that are not conditionally conjugate. The Metropolis-Hastings sampler is the most flexible of the three; it employs a proposal distribution to determine whether the new proposed state should be rejected or accepted.

2.4.3 Convergence Diagnosis

A stationary distribution should be obtained for the Markov chain providing sufficient iterations have been afforded. This expectation is based on the theory for this method; in practice the stationarity may not be achieved. The rate of convergence

can be affected by autocorrelation, choice of sampling algorithm, and model identification issues (Kim & Bolt, 2007). The lack of convergence can be determined by visual and diagnostic means. Visual inspection is the observation of plots such as history, running mean, density, quantiles. The practitioner reviews the graphs for these plots to determine if the properties such as a lack of variability, or for density, strongly unimodal distributions, are observed.

Automated tools have been developed to analytically diagnose whether acceptable levels of variation have been met. Frequently applied measures are the z-score, (Geweke, 1992) and the potential scale reduction factor, also called *R*hat (Gelman and Rubin, 1992). For the z-score method, the z-score for the difference between the first 10% of the iterations after burn-in and last 50% of these iterations is calculated. The z-score is then tested for a significance where values that are within ± 1.96 provide evidence for convergence. The R-hat is a measure that compares the between-chain and within-chain variance. When R-hat is approximately 1.0 convergence is acceptable. In practice, R-hat smaller than 1.2 is considered as acceptable convergence.

In conclusion, this chapter introduces foundational concepts in response and RT modeling including models most-often applied in IRT and extensions that address LID, LPD or both. Discussion of RT modeling was provided for the speed-accuracy tradeoff which in some models is specified directly and in other models is permitted to be estimated. Statistical relationships were presented, for example, conditionally independent, where dependency is associated only within the modeling of the IRT and within that of the RT, respectively. Advancements in joint IRT and RT modeling

of responses and RT were presented, followed by an overview of Bayesian estimation methods. The next chapter details a proposed joint response and RT model and the methods for investigation of the model.

Chapter 3: Methods

The previous chapter presents a review of the landscape for research in development, evaluation, and parameterization of models with dual dependency. The current chapter highlights the relevant factors that determine the scope and challenges of driving the proposed study. The discussion addresses prior similar efforts and the factors that must be addressed related to the proposed model, estimation, evaluation of the proposed model against multiple competing models in a simulation study and an empirical study.

3.1 A Multilevel Testlet Joint Response and Response Time Model

The van der Linden (2007) joint response and response time model allows speed and accuracy parameters to covary. That is, unlike most previous models in the history of response time modeling, the model does not a priori determine the strength or direction of the relationship. The van der Linden model's flexibility has led to its use in the development of several extensions (e.g., Im, 2015; Klein Entink, Fox, & van der Linden, 2009; Klotzke & Fox, 2019). Presented in Figure 2 is the hierarchical model framework.

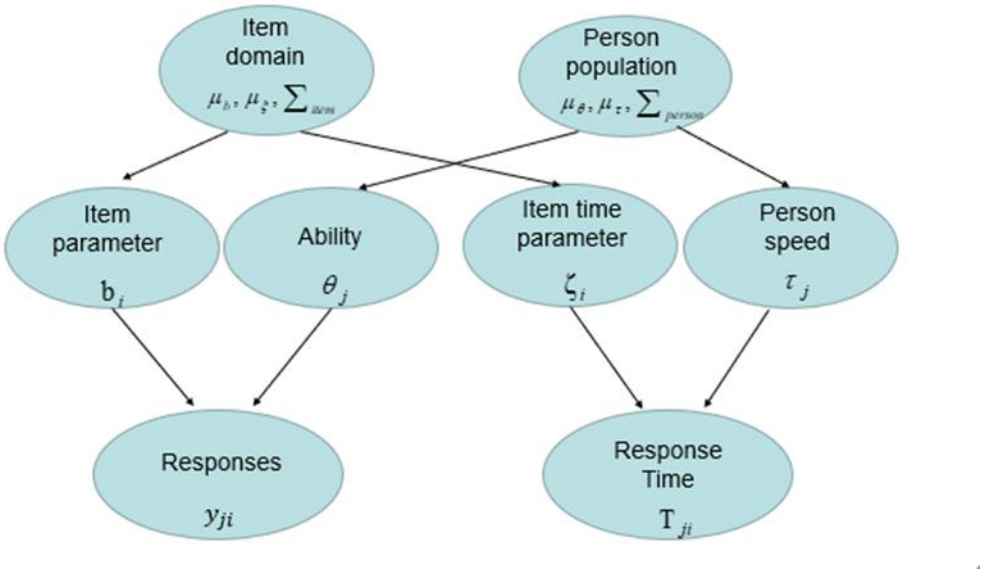

Figure 2. A hierarchical structure of the joint modeling of responses and response time (van der Linden, 2007).

In some testing programs, items are being developed not as stand-alone items rather associated with item clusters or testlets. Departing from the simple structure of one item being associated with one response makes it likely one item response might be dependent of another response. There is opportunity for more sophisticated modeling to capture the testlet effects. Currently, many testing programs (e.g. PISA, NAEP) use testlets. While these programs use complex sampling designs inducing person clustering, which in the proposed model is accounted for by multilevel modeling by group, there has not been a thorough simulation study to assess a model that addresses the dual dependencies (violations of independence) for both LID and LPD in a joint response and response time model. The proposed model, the multilevel testlet joint model (MTJM) takes into account the dual dependence due to person and

item clustering. This research study intends to develop a Bayesian estimation method for model parameter recovery for the proposed new model.

In general, the proposed model extends the hierarchical framework for joint modeling of responses and response time (van der Linden, 2007), the multilevel models for person clustering effects (Fox & Glas, 2001; Kamata, 1998, 2001), the testlet (Bradlow, Wainer, & Wang, 1999), and the joint modeling person and item clustering effects (Jiao et al., 2012). Figure 3 graphically represents the proposed MTJM. To investigate the impact of ignoring the person and item clustering effects in the joint modeling of responses and response time, this study referred to the above models as the HM – the hierarchical model, MJM – the multilevel joint model, and TJM – testlet joint model. The HM provides the joint estimation of latent parameters for both a response and response time model and assumes the conditional independence for responses and RTs, respectively. The MJM accounts for the group effects within a population as an extension of the hierarchical model. The TJM is also an extension of the HM. The TJM model enables the estimation of a testlet effect due to item clustering. In the TJM, the testlet effect is modeled for the response accuracy and the response time concurrently. The proposed MTJM account for both person and item clustering effects in responses and response time data due to complex sampling as in the MJM and item clustering for the response model.

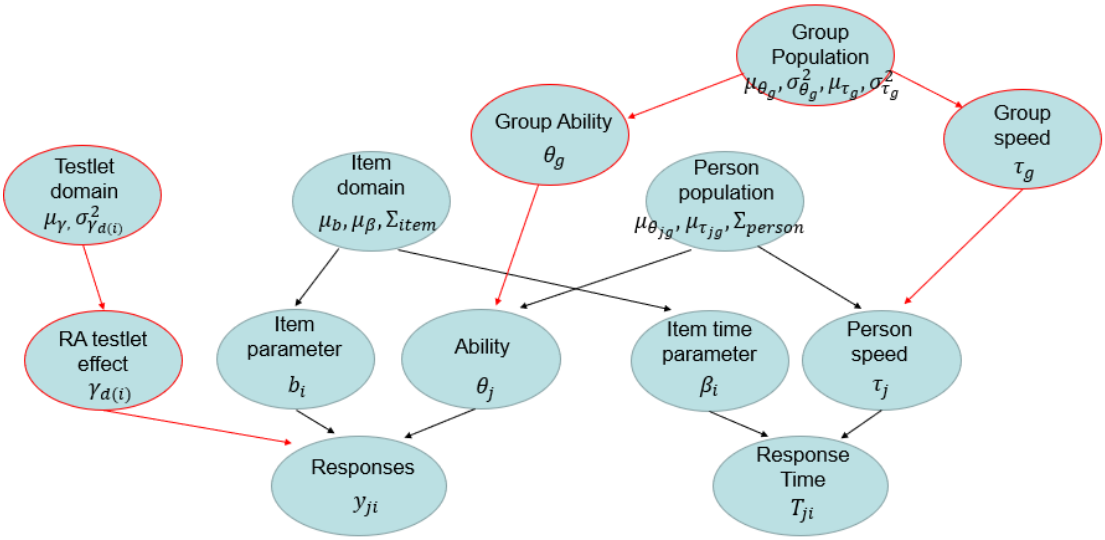

Figure 3. A hierarchical structure of the joint modeling of the multilevel testlet joint model (MTJM) of responses and response time.
Note. Extensions to the hierarchical model (van der Linden, 2007) are identified in red.

The MTJM is a multi-level model that is intended to be used to investigate person clustering and item clustering effects. The IRT model used for modeling item responses is the traditional Rasch model (Rasch 1960). As the proposed model incorporates testlets (Im, 2015), the Rasch testlet model (Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005) is adopted for dichotomous responses. On the other hand, according to Gelman and Hill (2007), there are five approaches to modeling a multi-level structure. Several of these have been addressed in multilevel response and RT studies (Klein Entink, Fox, & van der Linden, 2009; van der Linden, 2007) and in the testlet literature (Im, 2015; Jiao et al., 2012). This proposed research models the dual clustering effects simultaneously using the combined approach as in Jiao et al. (2012). Combining the multilevel structures in responses, the RA model can be presented as follows:

$$P\left(y_{ijg}=1|\theta_{jg},\theta_g,b_i,\gamma_{jd(i)}\right)=\frac{1}{1+\exp-\left(\theta_{jg}+\theta_g-b_i+\gamma_{jd(i)}\right)}. \quad (3.1)$$

In this formulation, the probability of a correct response $y_{ijg}=1$ by examinee *j* within group g with group ability θ_g and person-specific latent ability θ_{jg} to item *i* within testlet *d* with testlet-specific ability $\gamma_{jd(i)}$ is presented as $P\left(y_{ijg}=1|\theta_{jg},\theta_g,b_i,\gamma_{jd(i)}\right)$. Following the simplified formulation as in Jiao, Kamata, Wang, and Jin (2012), item difficulty is b_i, and the testlet effect parameter, $\gamma_{jd(i)}$, where *d* indicates the specific testlet items *i* associated with. The group ability is the same for all persons in a group but differs for persons from different groups. The variance of the ability of the groups is the indication of the person clustering effects.

On the other hand, combining the multilevel structure in RT, the RT model can be presented as follows:

$$f\left(t_{ijg}|\tau_j,\beta_i\right)=\frac{1}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\ln t_{ijg}-\left(\beta_i-\tau_{jg}\right)\right)^2\right\}. \quad (3.2)$$

Expressed a bit differently, the formulation is

$$t_{ijg}=\beta_i-\tau_{jg}+\varepsilon_{\tau_{ijg}}\ ;\ \varepsilon_{\tau_{ijg}}\sim N(0,1) \quad (3.3)$$

$t_{ijg}$was transformed via log-normal function. The RT is a continuous measure, t_{ijg} and is modeled for examinee *j* within group g with the lognormal distribution using the person speed parameter, τ_{jg}, and item parameters for item intensity β_i. As each examinee is nested within one group, the joint model provides indexing of parameters that represents the examinee's group membership. The group speed parameter variance is the indicator of the between-group effects for speed. The proposed RT

model is a simplified HM where the discrimination parameter is assumed constant and set to unity.

For this study, the effects of LID and LPD are a key motivation. Consequently, the proposed model is compared to the alternative models using the same parameterizations and formulations when applicable. Table 1 provides the response model parameters and RT model parameters in the proposed model, which is useful to highlight the key differences compared to the alternative models.

Table 1

The Proposed and the Alternative Models

Model	Abbreviation	Response Formulation	RT Formulation
Multilevel testlet joint model	MTJM	$\theta_{jg} + \theta_g - b_i + \gamma_{jd(i)}$	$\beta_i - \tau_{jg}$
Testlet joint model	TJM	$\theta_j - b_i + \gamma_{jd(i)}$	$\beta_i - \tau_j$
Multilevel joint model	MJM	$\theta_{jg} + \theta_g - b_i$	$\beta_i - \tau_{jg}$
Hierarchical model	HM	$\theta_j - b_i$	$\beta_i - \tau_j$

3.2 Model Parameter Estimation

The estimation of the model parameters is conducted in R using the package R2Jags with calls to the Bayesian estimation program JAGS (Just Another Gibbs Sampler; Plummer, 2017). Jags uses Markov Chain Monte Carlo (MCMC) simulation to generate a posterior distribution for each stochastic node, that is, each node follows a distribution with different moments, instead of a known fixed value.

In Bayesian estimation, the distribution of a parameter before being updated based on data via the MCMC process is known as the prior distribution. The choice of

priors has influence on the posterior distribution. This effect diminishes when the prior has a diffuse distribution, that is greater variability, and when more information is available from data during the updating process (Gelman & Hill, 2007). The selection of the priors for this study is guided by preceding research. The prior distributions for the related model parameters are presented as follows. Across models, the same prior distributions are set for the equivalent model parameters.

$y \sim$ Bernoulli distribution,

$t \sim N$, after log transform of response time, t,

$$\boldsymbol{\mu}_P = (\mu_\theta, \mu_\tau) = (0,0)^{\mathrm{T}},$$

$$\boldsymbol{\Sigma}_P = \begin{pmatrix} 1 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_I = \begin{pmatrix} \sigma_b^2 & \sigma_{b\beta} \\ \sigma_{b\beta} & \sigma_\beta^2 \end{pmatrix},$$

$$\gamma_d \sim N\left(0, \sigma_\gamma^2\right),$$

$$\theta_g \sim N\left(0, \sigma_{\theta g}^2\right)$$

$$\tau_g \sim N\left(0, \sigma_{\tau g}^2\right)$$

Hyperpriors

$$\boldsymbol{\Sigma}_I \sim InvWishart(\boldsymbol{I}_2, 2)$$

$$\sigma_\gamma^2 \sim InvGamma(1,1)$$

$$\sigma_{\theta g}^2 \sim InvGamma(1,1)$$

$$\sigma_{\tau g}^2 \sim InvGamma(1,1)$$

$\boldsymbol{I}_2$ is a 2-dimensional identity matrix

The item parameters are assumed to follow a bivariate distribution, whereas van der Linden (2007) applied the multivariate normal distribution due to the addition of time discrimination parameters to be estimated. Liao (2018) found that there were a ten-fold gain in efficiency, in terms of time for one iteration to be processed, by using the bivariate estimation approach. As the partitioning is similar for this study, the Liao (2018) approach is adopted here. The model is identified by fixing the scale of the person parameters.

3.3 Simulation Design

To determine the effects of the manipulated factors on the estimation of the proposed model, a simulation study was designed. The effects on parameter estimates as a function of the test length, testlet effects, person clustering effects, and the correlation between the ability and the speed parameters were evaluated when compared to competing models. The selection of the data generating model which is the proposed model when compared to the alternative models using relative model fit measures are presented. In addition, an empirical study using a PISA (2015) international dataset is presented to illustrate the application of the proposed model to real test data.

3.3.1 Manipulated Factors

This section details the factors and the levels manipulated for the simulation study. Table 2 provides a summary of the values of each level of these factors. Each of the factors includes two levels. Fully crossing the levels of the manipulated factors results in 16 experimental conditions. The conditions are summarized in Table 3.

Table 2

Summary of Manipulated Factors

	Manipulated Factors			
Levels	Test Length	$\sigma^2_{\gamma i(d)}$	Group Variance	$\rho_{\theta\tau}$
1	24	0.25	0.25	.30
2	48	1.0	1.0	.70

Table 3

Summary of the Simulation Conditions

	Manipulated Factors			
Condition No.	Test Length	$\sigma^2_{\gamma i(d)}$	Group Variance	$\rho_{\theta\tau}$
1	24	0.25	0.25	.30
2	24	0.25	0.25	.70
3	24	0.25	1.0	.30
4	24	0.25	1.0	.70
5	24	1.0	0.25	.30
6	24	1.0	0.25	.70
7	24	1.0	1.0	.30
8	24	1.0	1.0	.70
9	48	0.25	0.25	.30
10	48	0.25	0.25	.70
11	48	0.25	1.0	.30
12	48	0.25	1.0	.70
13	48	1.0	0.25	.30
14	48	1.0	0.25	.70
15	48	1.0	1.0	.30
16	48	1.0	1.0	.70

In studies related to response and RT modeling, the number of items (test length) has been a manipulated factor. Researchers have included as few as 15 items (Man, Harring, Jiao, & Zhan, 2019) and as many as 68 items (Im, 2015) in their investigations. The most frequently paired levels for items is 20 and 40; this two-level manipulation is found in the studies by Liao (2018), Molenaar, Tuerlinckx, and van der Maas (2015) and Wang, Fan, Chang, and Douglas (2013). Other choices for test

length include 30 and 60 items, (e.g., Suh, 2010), 20 and 60 items (e.g., Wang & Hanson, 2005), and 25 and 49 items (e.g., Bolsinova, de Boeck, & Tijmstra, 2017). In multilevel modeling for person clustering effects, Klein Entink, Fox, and van der Linden (2009) included 20 items but did not manipulate the number of items as a factor. A testlet was included in the response and RT joint modeling by Im (2015) which had 54 to 68 items in the simulation and 33 items in the empirical dataset. Other studies on polytomous testlet models but not RT models included 20 items (e.g., Huang & Wang, 2014) and 36 items (e.g., Jiao & Zhang, 2015). For this proposed study, two levels are manipulated for test length at 24 items and 48 items.

The number of testlets is 3 for the proposed 24-item test and 6 for the proposed 48-item test. This is a relatively low number which is motivated by the interest in evaluating the variance of the testlet estimate. Studies of item clustering have included as few as 2 in a condition (e.g., Huang & Wang, 2014) and 10 or more (DeMars, 2006, 2012; Ra, 2012). Tests described with 4 testlets include those by Wang and Wilson (2005), Wainer, Bradlow, and Du (2000), Jiao, Kamata, Wang and Jin (2012), and with 6 testlets (Im, 2015), Jiao and Zhang (2015), Jiao, Wang, and He (2013), Jiao, Kamata, Wang, and Jin (2012).

Items per testlet is often a consideration along with the number of testlets as the product provides the overall test length. At the lower end, studies such as Wang and Wilson (2005) used 5 items. On the other hand, the study by Im (2015) included a range of 8 to 14 items per testlet. Bradlow, Wainer, and Wang (1999), and Wainer, Bradlow, and Du (2000), included 10. Nine items per testlet were in the studies by

Jiao, Kamata, Wang and Jin (2012), and Jiao, Wang, and He (2013); Jiao and Zhang (2015) incorporated 6 items per testlet.

The magnitude of testlet effects has often been simulated by manipulating the testlet variance at different levels. Researchers have included a zero-variance condition as a control, (e.g., Bradlow, Wainer, & Wang, 1999; Jiao, Wang, & He, 2013; DeMars, 2012; Murphy, Dodd, & Vaughn, 2010). The variance of 0.25 has been selected as the value for a "small" testlet effect by several studies (e.g., Glas, Wainer, & Bradlow, 2000; Im, 2015; Jiao, Wang, & He, 2013; Jiao & Zhang, 2015; Wang & Wilson, 2005). This testlet variance magnitude was also simulated in Jiao, Kamata, Wang, and Jin (2012), which modeled LID and LPD jointly. When a study has included a level for a "moderate" effect of a testlet, the value tends to be 0.50 (e.g., Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005; Im, 2015). For a "large" effect size, the variance chosen is rarely greater than 1.0 (e.g., Bradlow, Wainer, & Wang, 1999; DeMars, 2012 Murphy, Dodd, & Vaughn, 2010), and 1.0 is very frequently the magnitude for testlet variance (e.g. Bradlow, Wainer, & Wang, 1999; Glas, Wainer, & Bradlow, 2000; Im, 2015; Jiao, et al, 2012; Jiao, Wang, & He, 2013; Jiao & Zhang, 2015; Wang & Wilson, 2005). Studies that model examinee responses and RT using testlets for LID are not well-represented in the literature. In a testlet partial credit model (PCM), Jiao and Zhang (2015) considered the LPD, and in Im (2015) testlets were applied but without the consideration of the LPD. In the study proposed, the "small" and "large" testlet variances included have values, 0.25 and 1.0, respectively.

For joint modeling of response and RT, often a correlation of person parameters is explicitly simulated. Values of the correlation between latent person ability and person speed have ranged from -0.9 to 0.9 (Suh, 2010). Some researchers have studied a correlation of 0.50 representing moderate strength (e.g. Klein Entink, Fox, & van der Linden, 2009; Molenaar, Tuerlinckx, & van der Maas, 2015). The same value was also used by Liao (2018) and Zhan, Jiao, and Liao (2017). Studies with low strength include absolute correlations values of 0.20 (e.g., Liao, 2018) and 0.30 (e.g., Suh, 2010; Zhan, Jiao, & Liao, 2017; Man, Harring, Jiao, & Zhan, 2019). Strong correlations between the ability and speed parameters have been 0.80 (e.g., Sen, 2012; Liao, 2018), and 0.90 (Suh, 2010). In their empirical data study, Zhan, Jiao, and Liao (2017) reported the absolute value of correlation 0.507. For this study, the levels manipulated are 0.30 and 0.70 which respectively represent weak and strong association between the ability and speed parameters.

The group clustering effect is quantified in terms of the group variance (Raudenbush & Bryk, 2002). When a testlet effect was modeled (Jiao et al, 2012) in IRT, group variances were manipulated at two levels of 0.25 and 1.0. These same values were used when the multilevel effects were simulated in the polytomous IRT study by Jiao and Zhang (2015). For this study, the group clustering effect is manipulated at two levels of group variances, 0.25 and 1.0. These two levels are applied to both the ability and speed person clustering parameters.

3.3.2 Fixed Factors

The simulation design includes consideration of factors reflecting the real application situations. Some factors are manipulated as described in the above section

while other factors are fixed as described in this section. Table 4 provides a summary of factors that are not manipulated.

Table 4

Summary of Fixed Factors

Factor	Fixed Value
Number of person clusters	50
Cluster group size	25
Total sample size	1,250
Number of items per testlet	8
Correlation between item difficulty and item intensity	0.30
Time discrimination	1

For simulations with fixed correlations between item parameters (here, difficulty and intensity), recent research (e.g., Liao, 2018; Man et al, 2019) has set the parameter to .30. This correlation strength is also used in the proposed study.

In the 2015 paper by Jiao and Zhang, the sample of 1,000 students was split among 40 groups with each group having the same number of students, therefore, 25 students per group. In the proposed study, as in Jiao, Kamata, Wang, and Jin (2012) supported by Binici (2007), the cluster number for this design is 50. Therefore, 50 groups of equal size with 25 respondents per group result in a sample size of 1,250 respondents. The number of respondents proposed is in line with Im (2015) who evaluated a TJM with sample sizes ranging from 888 to 1,378 respondents, but for which there was no person clustering. Other studies in RT have used similar sample sizes, for example, van der Linden (2007) included 1,104 respondents and Klein Entink, Fox, and van der Linden (2009) modeled the response and response times of 1,000 respondents. The proposed model has additional complexity compared to the comparison models, so the number of respondents is set at least as large as those

studies that have one fixed sample size and near the upper range of those studies that included multiple samples.

As the design is to have all levels fully crossed, the total number of conditions is 2 × 2 × 2 × 2 = 16. Originally, 25 replications were proposed for each condition. Due to difficulties of convergence, there were a resulting 10 converged datasets for all models simulated for each condition. This provides 160 total simulated datasets. The generating model is the proposed new model. Each dataset is fitted with the proposed model and the three comparison models, respectively. The individual-specific ability, individual-specific speed, and item difficulty each follow standard normal distributions. The time intensity has distribution N(4,1). To account for homogeneity of variance the within-group variance was fixed. R version 3.5.3 (R Core Team, 2019) is used to generate data, and interface with JAGS using the R2jags package (Su & Yajima, 2015). It is also used for analyses to evaluate model parameter recovery evaluation.

3.3.3 Evaluation Criteria

For this study, two chains are used for model parameter estimation. In MCMC, the initial iterations that are not kept for the estimation of the posterior distribution are known as burn-in iterations. For this study, a minimum of 10,000 iterations are identified as burn-in number of iterations. The number of burn-in iterations depends on the model. The convergence is evaluated by the Gelman-Rubin convergence diagnostic $\hat{R} < 1.20$ (Gelman & Rubin, 1992). A stopping rule is performed automatically in the JAGS program when this criterion is met for all sampled nodes. There are 10,000 iterations per chain kept for the estimation post-burn

in resulting in an estimation sample of 20,000. Visual inspection of the parameters of interest showed good mixing.

To evaluate the model parameter recovery, two error indices are computed: the bias and the root mean squared error (RMSE). The formulations for these indices are:

$$Bias(\hat{\boldsymbol{\eta}}) = \frac{\sum_{r=1}^{R}(\widehat{\eta_r} - \eta)}{R}, \tag{3.4}$$

$$RMSE(\hat{\boldsymbol{\eta}}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\eta_r} - \eta)^2}\ , \tag{3.5}$$

The true parameter of interest is η and the parameter estimate $\hat{\eta}$ represents η at 1 replication. The R term indicates the number of replications for a specific study condition. For these calculations, the effective number of replications in this study was 1, as each simulation run was aggregated within a condition. According to Lord (1986), the increase in the bias of a Bayesian estimate in IRT is the trade-off for achieving a lower overall mean squared error, which relates to the random error in model parameter estimates.

To investigate the effects of the manipulated factors, an analysis of variance (ANOVA) is performed for each of the outcome measures. Prior to conducting the ANOVA analyses, the assumptions are checked so that the right procedure is identified. For example, when the sphericity assumption is violated, the Huynh-Feldt correction (Huynh & Feldt, 1976) is applied. A summary of parameters of interest is provide in Table 5. This table also identifies the method of analysis used for reporting in the results section.

Table 5

Methods for Summarizing Model Parameters

No.	Symbol	Variable Description	Analysis
1	$\theta_{j(g)}$	Individual-specific ability	ANOVA
2	$\tau_{j(g)}$	Individual-specific speed	ANOVA
3	b_i	Item difficulty	ANOVA
4	β_i	Time intensity	ANOVA
5	$\sigma^2_{\theta j(g)}$	Variance of individual-specific ability	Descriptive
6	$\sigma^2_{\theta g}$	Variance of group-specific ability	Descriptive
7	$\sigma^2\tau_{j(g)}$	Variance of individual-specific speed	Descriptive
8	$\sigma^2\tau_g$	Variance of group specific speed	Descriptive
9	$\sigma^2_{\gamma d(i)}$	Variance of testlet	Descriptive
10	$\rho_{\theta\tau}$	Correlation between ability and speed	Descriptive
11	σ^2_b	Variance of item difficulty	Descriptive
12	σ^2_β	Variance of time intensity	Descriptive
13	$\rho_{b\beta}$	Correlation between item difficulty and time intensity	Descriptive

Some factors are repeated measures in ANOVA. The repeated factor in ANOVA in this study is the factor: model, including the proposed model and the competing models (namely 4 models: MTJM –multilevel testlet joint model, MJM – multilevel joint model, TJM – testlet joint model, and HM – hierarchical model). Other non-repeated factors in the ANOVA include: the number of items (test length), the testlet effects, the correlation of ability and speed, and person clustering effects. The dependent variables are the bias and RMSE for each of the following parameters: individual-specific ability, individual-specific speed, item difficulty, and time intensity.

The proposed model is also compared with competing models on measures of relative model fit to identify which of the model fit indices perform best on

identifying the true model. The model fit indices investigated are the deviance and the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) which is derived from deviance and an estimated number of effective parameters. Its formulation is:

$$DIC = \overline{D(S)} + p_D, \tag{3.6}$$

where S indicates the sample space of all model parameters, $\overline{D(S)}$ is the posterior mean of the deviance, calculated as -2log likelihood, $p_D$is a complexity term that is calculated as the posterior mean deviance at each iteration minus the deviance at the posterior means of the parameters.

3.4 Empirical Data Analysis

An empirical data analysis is conducted to demonstrate the application of the proposed model and for comparison with competing models. A set of mathematics items from the Program of International Student Assessment (PISA) 2015 was selected; the data are publicly available and include response and response time data for each item. The set of items are from the M1 and M2 testing clusters. There are 17 dichotomously scored items on the subtest, included are 4 testlets of 2 items each and 9 independent items. Response times are provided in milliseconds. They were transformed to seconds then set on a logarithmic scale prior to estimation using the proposed model. Cases with missing values were deleted listwise. This selection process resulted in a dataset of 8,606 students from 58 countries. This initial dataset is the same as that used by Zhan, Liao, and Bian (2018). The empirical analysis uses a random sample of 20% from the PISA study. Countries that had 20 or more students were included. The sample cluster sizes per country were expected to be proportional

to the number of participants from a given country. This selection processes resulted in a final dataset with G=44 countries and N=1,478 respondents. There was an average of 34.37 students per country with SD=19.60, the maximum number of students was 119 from a single country, and as previously stated the minimum cluster size was 20.

Unlike a simulation, the true values of the parameters are not known, therefore the empirical data analysis is limited to relative comparisons. The proposed model and the three alternative models included in the simulation design are compared for model fit. For these model comparisons, the DIC (Spiegelhalter, Best, Carlin, & van der Linde, 2002) and deviance are compared.

Chapter 4: Results

The previous chapter introduced the proposed model and research study design. In the present chapter the findings from the simulation study and the empirical study are presented. The chapter is organized in the following manner: Section 4.1, the results of the simulation study for the measured error for parameter recovery and the effects of manipulated factors; Section 4.2, the performance of DIC in selecting the MTJM as the best-fitting model for data with LID and LPD; and Section 4.3, the application of the proposed model using empirical data.

4.1 Results of Simulation Study

The proposed and comparison models were fitted to each generated data set to estimate parameters from simulated datasets to evaluate model parameter recovery. The MCMC Bayesian estimation approach was used for model parameter estimation. Among the compared models: the MTJM, the multilevel joint model (MJM), the testlet joint model (TJM) and the hierarchical model (HM), not all models included the same parameters as some models may ignore the effects of LID or LPD or both. An overview of the model specification is included in Table 6. For parameters where sample size was adequate, the evaluation criteria were summarized in terms of descriptive and inferential statistics methods.

Table 6

Overview of the Model Specifications of the Estimation Model in the Simulation Study

Model	Model addresses presence of local dependence	
	Local item dependence	Local person dependence
HM	×	×
MJM	×	√
TJM	√	×
MTJM	√	√

Note. HM= Hierarchical Model, MJM= Multilevel Joint Model, TJM= Testlet Joint Model, MTJM= Multilevel Testlet Joint Model; × represents absence of parameter in the model, √ represents presence of parameter in the model.

The simulation study design includes the manipulation of four factors including the test length, testlet variance, group variance, and person ability and speed correlation. As each factor included two levels, in this fully crossed design there are 16 simulation conditions. The parameters of interest for recovery determination are listed in Table 5 (previous chapter). Estimated covariances were converted to correlations using the estimated variances for the respective parameters.

Convergence for the Bayesian MCMC parameter estimation was empirically evaluated by checking if $\hat{R} < 1.20$ for all model parameters. The number of burn-in iterations differed by model and test length. For the models that did not include a multilevel component (TJM and HM), burn-in was set to 10,000 iterations; for the shorter test length condition with models that did include a multilevel component (MTJM and MJM) burn-in was 50,000 iterations, and in the longer test length conditions for the MTJM and MJM burn-in was 100,000 iterations. If convergence was not met in the post-burn in sample, then for the 10,000 burn-in scenario autojags was run until the convergence criteria was met; this occurred for all estimation runs in five or fewer updates. For the scenarios with burn-in of 50,00 and 100,000, autojags was not used as the time to run an estimation replication could be considerable. These replications were re-run on a separate computing instance to allow the script to continue with the estimation of additional replications.

Visual inspection of diagnostic plots and trace plots was also used in convergence assessment. There were two MCMC chains run, in the two scenarios

with longer burn-in, these were computed using parallel processing. The data sample included 5,000 iterations post-burn in, and as there were 2 chains per MCMC run, this resulted in 10,000 iterations used for the estimates of the model parameters. In the two scenarios with longer burn-in, the 5,000 iterations were secured by running 25,000 iterations and applying a thinning of 5 in order to conserve computer memory. Ten replications with converged results were obtained within each simulation condition. All estimations were performed using cloud computing on virtual machines using an Amazon Machine Image (AMI) that was pre-built with R, RStudio, and JAGS included (Aslett, 2019). Specifications for these computing resources include virtual machines for the short burn-in scenario each with 2 cores, 2.8 GHz speed, 4 GB memory, and virtual machines for the longer burn-in scenarios each with 4 cores, 2.3 GHz speed, and 8 GB memory. A replication took approximately 4 hours for the short burn-in scenario and 48 hours for the longest burn-in scenario.

Two dependent measures were computed based on the estimates of the parameters. The bias and RMSE were calculated for person-specific ability, speed, item difficulty, and time intensity parameters. The detailed bias and RMSE for each simulation condition is reported for each parameter identified in the ANOVA analysis of Table 5 in Appendix A. Mixed-effect ANOVA was performed to investigate the effects of the manipulated factors on the estimation error measures. This analysis was conducted using the repeated measures ANOVA in SPSS (version 26.0, IBM Corp, 2019). The mixed-effect ANOVA included the estimation model as a within-subjects

factor, and the four manipulated variables as the between-subjects factors in the simulation design.

The assumptions for the mixed-effect ANOVA are checked prior to the ANOVA and reporting. The assumption of normality for the dependent variables was addressed by inspection of P-P and Q-Q plots; these were determined to be within acceptable levels based on robustness to moderate deviations from normality (e.g., Glass, Peckham, & Sanders, 1972).

Following the recommendation of Maxwell and Delaney (1990), the simulation was designed to have equal sample sizes to address the violation of the assumption. This implementation with equal sample sizes in study conditions with the same test length makes the procedure robust to the normality assumption. Sphericity, the variances of the differences between the related groups, was checked using the Mauchly's Test of Sphericity (Mauchly, 1940). SPSS provides a null hypothesis test in the output of the ANOVA. For example, in the RMSE of the speed parameter the Mauchly's W=.000 (df=5, p<.001), the Huynh-Feldt epsilon=.340 which was used to adjust to degrees of freedom for the tests of within-subject effects. In all cases the sphericity assumption was violated, therefore the results using the Huynh-Feldt correction (Huynh & Feldt, 1976) are reported.

In addition to statistical testing for significance (evaluated at the alpha level of 0.05), an effect size measure is used for evaluation of practical significance. The effect size applied in this study is the partial η^2 (Cohen, 1965). The range of effect size was used as suggested by Cohen (1988). That is, a small effect: 0.01≤ partial η^2 <0.06, a medium effect: 0.06≤ partial η^2 <0.14, and a large effect: partial η^2 ≥0.14.

Only the effects which are both statistically significant with at least a small effect size are reported and discussed in this study. To make interpretation manageable, the higher-order interaction reported for the mixed-ANOVA in this study is up to a three-way interaction.

The mixed-effect ANOVA is conducted for (a) the person ability parameters , θ_j, (estimated in the IRT model), (b) person speed parameter, τ_j (estimated in the RT model), (c) the item difficulty parameters, b_i, (estimated in the IRT model), and (d) the item time intensity parameters, β_i, (estimated in the RT model). The results are summarized as follows.

4.1.1 Person Parameters

For the person parameters, the mixed-effect ANOVA results indicate that the interaction of the study factors: the estimation model and the person group variance, has significant impact with a moderate or large effect size on the RMSE of the person ability and speed parameter recovery respectively. For bias, none of the factors and their interactions were identified as significant. A summary of the higher-order interaction for the person parameters is provided in Table 7.

Table 7

Summary of the ANOVA results on the Person Ability Parameter Recovery

Effect	Person Ability (θ_j)		Person Speed (τ_i)	
	Bias	RMSE	Bias	RMSE
Model*group_var	-	.066	-	.158

Effect Size	Small (0.01≤partial η^2<0.06)	Medium (0.06≤partial η^2<0.14)	Large (partial η^2≥0.14)

Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2); The values in the cells are partial η^2.

Person ability. In general, no factors or their interactions had significant effects on the bias of the ability parameter estimates for the mixed-effect ANOVA. However, to better understand bias in model parameter recovery, the conditional bias is plotted. Figure 4 provides a comparison of bias conditional on the true ability parameter between two conditions that differ only in the terms of the test length factor (conditions 4 and 12 respectively). In both conditions the same pattern results where the mean bias is greatest in the extremes near 3 logits in absolute value and decreases to nearly zero as the true ability decreases in absolute units. In consideration of the pattern observed and to support the analysis of the effects, the bias and RMSE were calculated using all the estimates for a parameter within a condition and not by aggregating these values per replication.

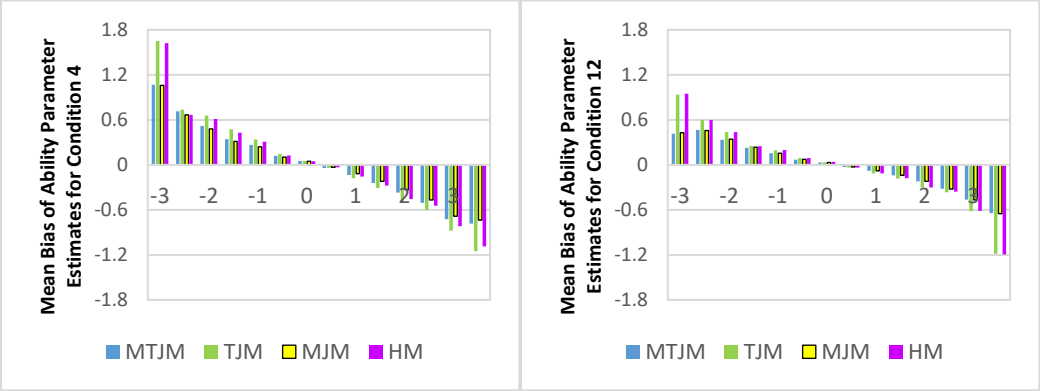

Figure 4. Mean Bias of Ability Parameter Estimates, θ_j, for Representative Conditions 4 and 12, which have the same factor levels with the exception of test length where condition 4 is I=24 and 12 is I=48.
Note. X-axis values are grouped by the true ability parameter with bin size of 0.5. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

There is a significant model and group variance two-way interaction on the RMSE of the θ_j with medium effect size (F=14183.43, p<.001, partial η^2=.066), there is also a significant two-way model and test length interaction with small effect

size (F=2030.59, p<.001, partial η^2=.010). These effects are reported in Table 8. The Model main effect is large (F=56338.34, p<.001, partial η^2=.220). The group variance factor has a small effect (F=7651.87, p=<.001, partial η^2=.037).

Table 8

The ANOVA Results of the RMSE of the Person Ability Estimates

Source	RMSE of θ_j		
	F Statistics	p-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	56338.34	<0.001	.220
Model*test_length	2030.59	<0.001	.010
Model*group_var	14183.43	<0.001	.066
Between-Subject Effects			
group_var	7651.87	<0.001	.037

Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2).

To better understand the two-way interaction between the two factors: model and test length, a mean plot is presented in Figure 5. Longer test lengths are associated with smaller mean RMSEs for θ_j. The two-way interaction between model and group variance is provided in Figure 6. In both of these figures the two models that take into account person clustering effects yielded mean RMSEs for θ_j smaller than those that do not include a group parameter in the model largely due to the fact that person clustering is an effect related to person parameters. Thus, models that take into account of such effect are expected to produce lower total estimation error. Figure 6 shows that for the TJM and HM, a larger group variance, σ_g^2, is associated with larger θ_j mean RMSE, compared to the level with smaller σ_g^2.

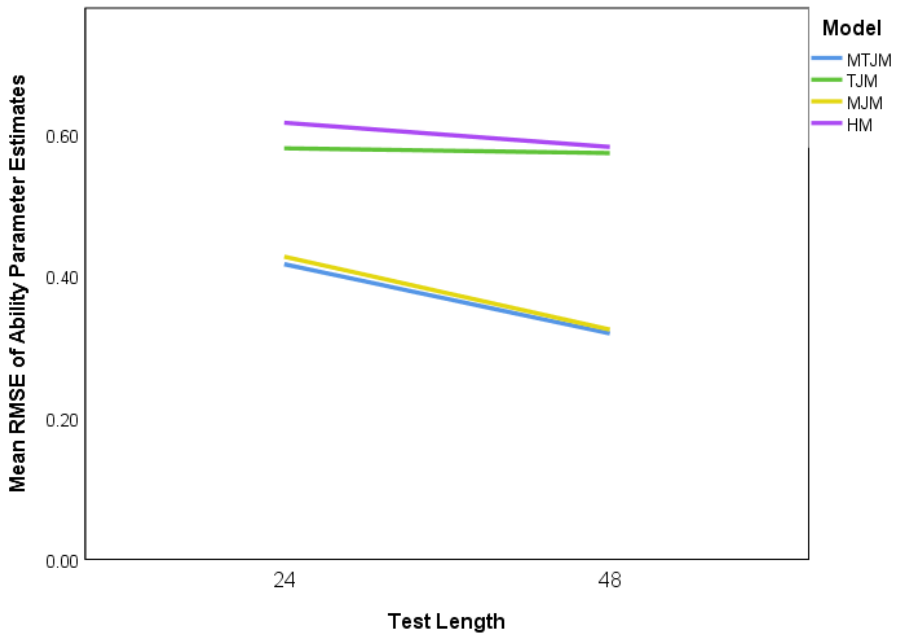

Figure 5. Significant two-way interaction of Model*test_length on the RMSE of the person ability parameter estimates, θ_j.
[Note. Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2).]

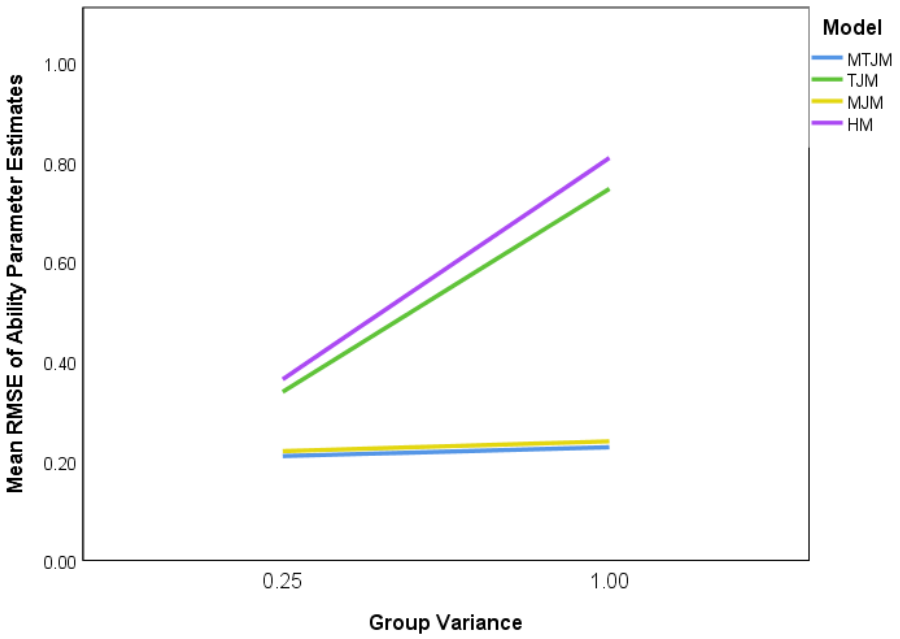

Figure 6. Significant two-way interaction of Model*group_var on the RMSE of the person ability parameter estimates, θ_j.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2).

Person speed. For the individual-specific speed parameter, τ_j, in the RT model, no factors have statistically significant effects on the bias. The model and the two-way interaction between the model and the group variance had significant effects on the RMSE of the speed parameter, τ_j. The two-way interaction between the model and the group variance has large effect size (F=37559.59, p<.001, partial η^2=.158) while the model has a large effect size (F=185938.24, p<.001, partial η^2=.482). The group variance factor has a large effect size (F=30467.07, p=<.001, partial η^2=.132). The ANOVA results are summarized in Table 9.

Table 9

The ANOVA Results of the RMSE of the Person Speed Estimates

Source	RMSE of τ_j		
	F Statistics	p-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	185938.24	<0.001	.482
Model*group_var	37559.59	<0.001	.158
Between-Subject Effects			
group_var	30467.07	<0.001	.132

Note. Model=Estimation model type; test_length=Number of test items (I); group_var=Group variance magnitude (σ_g^2).

Figure 7 depicts the two-way interaction between the model and the group variance. The mean RMSE for two of the four competing models, TJM and HM are almost overlapped; these are at the top of the figure. A similar pattern is found for the MTJM and the MJM; these models are at the bottom of the figure indicating smaller mean RMSE τ_j, These two models, which share the characteristic that they do have a grouping parameter, are showing smaller mean RMSE of τ_j than the TJM and HM. The magnitude of this difference is greater for the larger group variance.

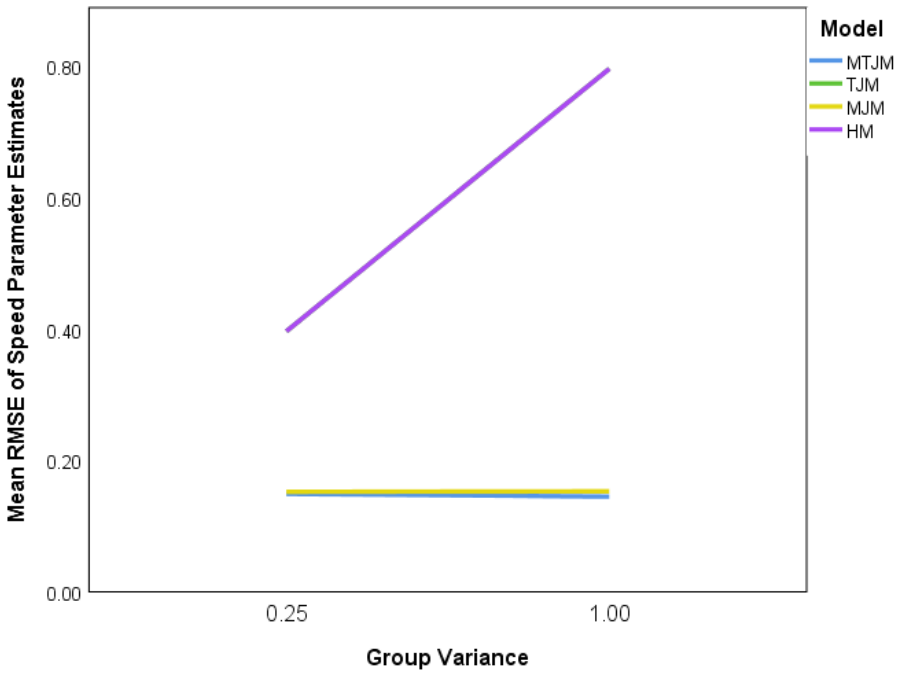

Figure 7. Significant two-way interaction of Model*group_var on the RMSE of the person ability parameter estimates, τ_j.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2).

4.1.2 Item Parameters

The ANOVA results of the bias and RMSE for the item parameters are presented in Table 10. Only the significant effects with at least a small effect size are presented in the table.

Table 10

The ANOVA Results for the Item Parameter Recovery

I	Effect	Item Difficulty Parameter (b_i)		Item Time Intensity Parameter (β_i)	
		Bias	RMSE	Bias	RMSE
24	Model*group_var	.013	.051	.022	
	Model*testlet_var		.051		
	Model*testlet_var*group_var				.020
48	Model*testlet_var*theta_tau_corr	-	.018	.016	.017

Effect Size	Small (0.01≤partial η^2<0.06)	Medium (0.06≤partial η^2<0.14)	Large (partial η^2≥0.14)

Note. Number of test items (I); Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$). The values in the cells are partial η^2.

Item difficulty. Two test lengths were investigated in this study. For the shorter test length (I=24), the two-way interaction between the model and the group variance has a statistically significant and small effect (F=25.01, p<.001, partial η^2=.013) on the bias of the item difficulty, b_i, parameter estimate between model and group variance. The mixture-effect ANOVA results are presented in Table 11. The factor, Model, has a significant and small effect size on the bias of the item difficulty parameter estimates (F=21.38, p<.001, partial η^2=.011). The two-way interaction is presented graphically in Figure 8. The mean estimation bias due to the estimation models is larger in absolute magnitude for the level with greater group variance, σ_g^2. There is a significant testlet variance effect on the bias of the $b_i$with small effect size (F=25.03, p<.001, partial η^2=.013) and a group variance effect (F=163.26, p<.001, partial η^2=.079) with medium effect size.

Table 11

The ANOVA Results of the Bias of the Item Difficulty Estimates (I=24)

Source	Bias of b_i		
	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	21.38	<0.001	.011
Model*group_var	25.01	<0.001	.013
Between-Subject Effects			
testlet_var	25.03	<0.001	.013
group_var	163.26	<0.001	.079

Note. Number of test items (I); Model=Estimation model type; group_var=Group variance magnitude (σ_g^2); testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

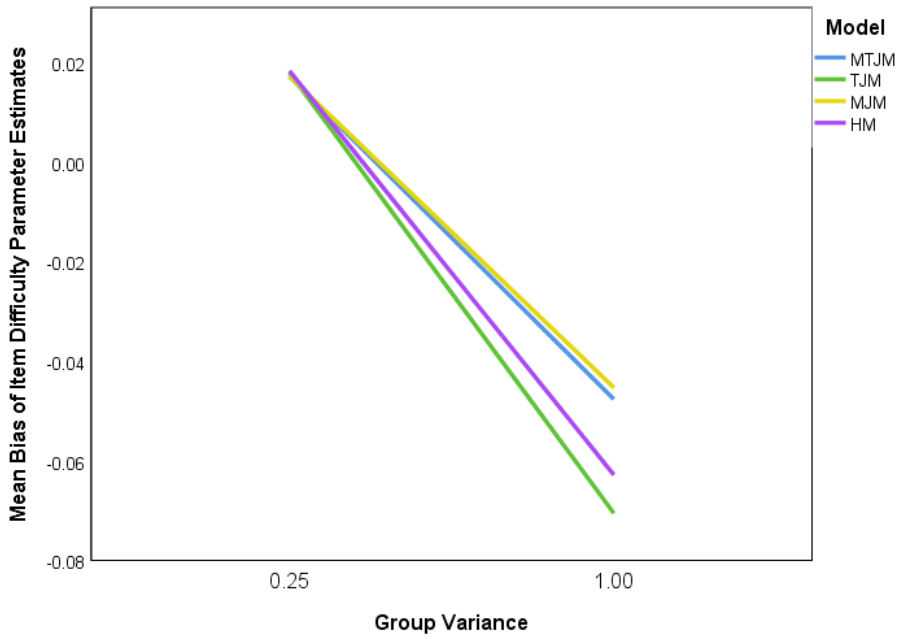

Figure 8. Significant two-way interaction of Model*group_var on the Bias of the item difficulty parameter estimates, b_i when I=24.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2).

For the study conditions with the shorter test length (I=24), the two-way interaction between the model and the testlet variance exerted a significant small effect on the RMSE of b_i (F=103.28, p<.001, partial η^2=.051). So does the two-way interaction between the model and group variance (F=101.98, p<.001, partial η^2=.051), as seen in Table 12. The factor, Model, has a medium effect (F=240.91, p<.001, partial η^2=.112) on the RMSE of b_i. The group variance factor has a large effect size (F=218.34, p=<.001, partial η^2=.102).

Table 12

The ANOVA Results of the Bias of the Item Difficulty Estimates (I=24)

Source	RMSE of b_i		
	F Statistics	p-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	262.67	<0.001	.121
Model*testlet_var	103.28	<0.001	.051

Model*group_var	101.98	<0.001	.051
Between-Subject Effects			
group_var	218.34	<0.001	.102

Note. Number of test items (I); Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

To better understand the two-way interactions, two figures are plotted and presented in Figures 9 and 10. The two-way interaction between the model and the testlet variance is presented in Figure 9. The mean RMSE of b_i is smaller for models incorporating a testlet parameter (MTJM and TJM) for the study condition with a large testlet variance, σ_γ^2; and is larger for the models that do not incorporate a testlet parameter (MJM and HM) at the larger level of σ_γ^2.

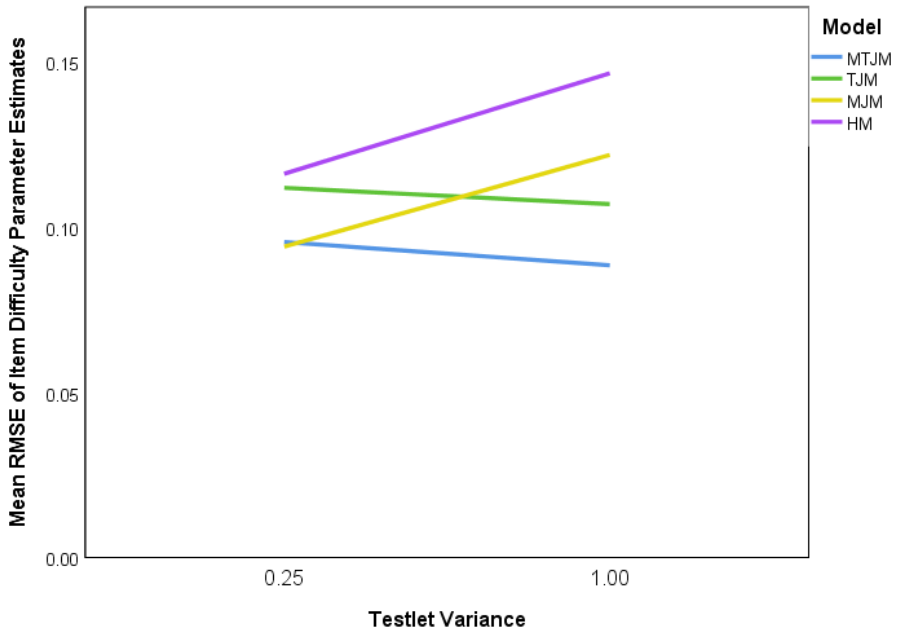

Figure 9. Significant two-way interaction of Model*testlet_var on the RMSE of the item difficulty parameter estimates, b_i when I=24.
Note. Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2).

The two-way interaction between the Model and group variance, σ_g^2 is represented in Figure 10 for the mean RMSE of b_i. In general, RMSE is smaller when the group variance, σ_g^2, is small for all models. For the study conditions with larger group variance, σ_g^2, the mean RMSE of b_i is smaller for the models that incorporate a group parameter (MTJM and MJM) than for the models that do not (TJM and HM).

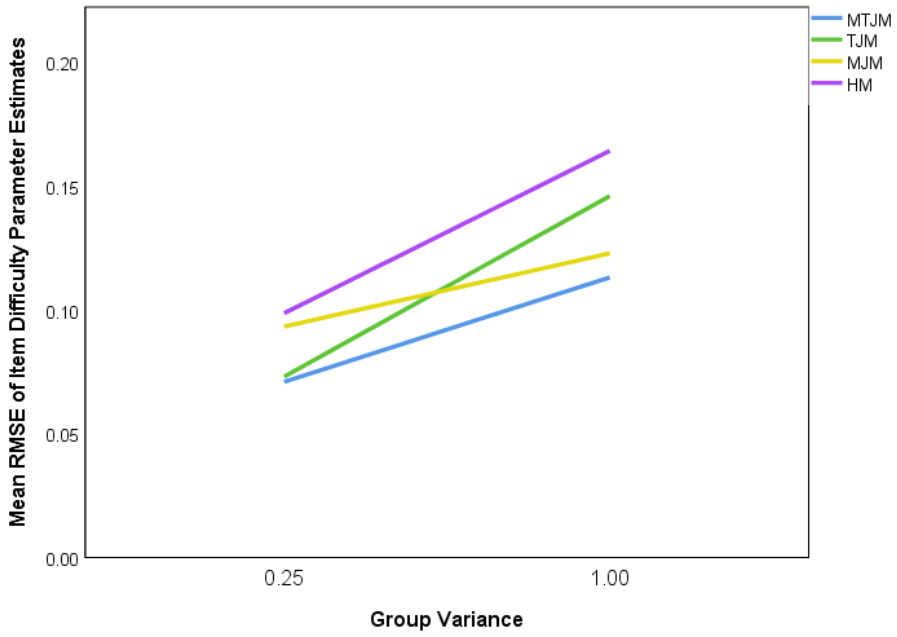

Figure 10. Significant two-way interaction of Model*group_var on the RMSE of the item difficulty parameter estimates, b_i when I=24.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2).

For the study conditions with a longer test length, the summary results of the mixed-effect ANOVA analyses are presented in Table 13. There were no significant effects for bias of b_i. However, several factors significantly affected the RSME of item difficulty parameter estimation with non-negligible effect sizes as presented in Table 13. Two significant three-way interactions: model, testlet variance, and correlation between ability and speed (F=70.31, p<0.001, partial η^2=.018) and model, group variance, and correlation between ability and speed, (F=37.71,

p<0.001, partial η^2=.010) were significant, each had a small effect. Both of these interactions are related to the correlation between person ability and speed ($\rho_{\theta\tau}$). The two-way interactions between the Model and the factors included in the three-way interactions (testlet variance , group variance, and correlation between theta and tau) were each found to be statistically significant: model and testlet variance (F=153.20, p<0.001, partial η^2=.038), model and group variance (F=164.50, p<0.001, partial η^2=.041), and model and correlation between theta and tau (F=74.72, p<0.001, partial η^2=.019). There was a medium effect for the Model factor (F=602.27, p<0.001, partial η^2=.136). The group variance and the correlation between theta and tau had a significant two-way interaction with a significant small effect (F=151.13, p=<0.001, partial η^2=.038). In addition, two factors had significant effects with small effect sizes: testlet variance (F=90.61, p<0.001, partial η^2=.023 and group variance (F=173.92, p<0.001, partial η^2=.043).

Table 13

The ANOVA Results of the RMSE of the Item Difficulty Estimates (I=48)

Source	RMSE of b_i		
	F Statistics	p-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	602.27	<0.001	.136
Model*testlet_var	153.20	<0.001	.038
Model*group_var	164.50	<0.001	.041
Model*theta_tau_corr	74.72	<0.001	.019
Model*testlet_var*theta_tau_corr	70.31	<0.001	.018
Model*group_var*theta_tau_corr	37.71	<0.001	.010
Between-Subject Effects			
testlet_var	90.61	<0.001	.023
group_var	173.92	<0.001	.043
group_var*theta_tau_corr	151.13	<0.001	.038

Note. Number of test items (I); Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

To better understand the multiple multi-way interactions, the mean plots for the three-way interaction among the model, testlet variance and the correlation between theta and tau are presented in Figure 11. For the study conditions with a smaller testlet variance, σ_γ^2, the MTJM and MJM with the multilevel structure, have smaller mean RMSE of b_i than the models that do not model the multilevel structure. In the left panel for the study conditions with smaller correlations between the ability and speed parameters, the larger testlet variance leads to a greater mean RMSE of b_i, an ordinal pattern is found among the model performance. In the right panel for the study conditions with larger correlations between the ability and speed parameters, the RMSE for the item difficulty parameters for the two models with a testlet parameter (MTJM and TJM) were not affected by the magnitude of the testlet variance. This is not the case observed in the models that do not address local item dependence (MJM and HM).That is, the mean RMSE of b_i, appears much greater for the study conditions with larger testlet variance (σ_γ^2=1.0).

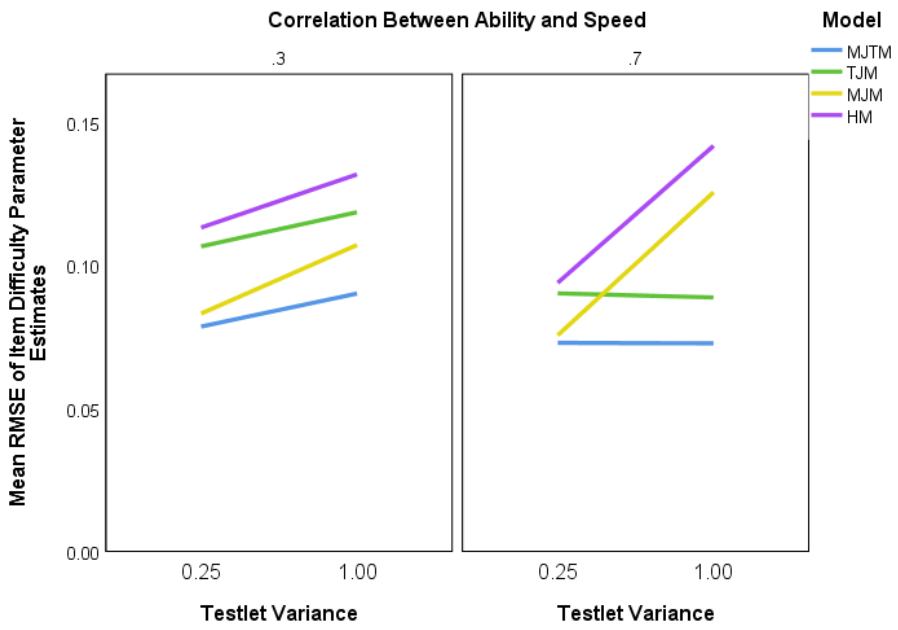

Figure 11. Significant three-way interaction of Model*testlet_var*theta_tau_corr on the RMSE of the item difficulty parameter estimates, b_i, when I=48.
Note. Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

The mean RMSE of b_i is represented in Figure 12 for the three-way interaction between model, group variance, and correlation between theta and tau. When the correlation between ability and speed was smaller ($\rho_{\theta\tau}$=.3) as presented in the left panel, the mean RMSE of b_i appeared similar for the smaller level of group variance. When the group variance is large, the two models with grouping parameter (MTJM and MJM) produced smaller mean RMSE compared to the two models that do not include a grouping parameter (TJM and HM). In the right panel that presents the results for the conditions with larger correlation, the MTJM and MJM show a slight decrease in the mean RMSE of b_i while the TJM and HM show slight increase with the true model lead to the smallest RMSE and the HM which ignores both item and person clustering effects the largest.

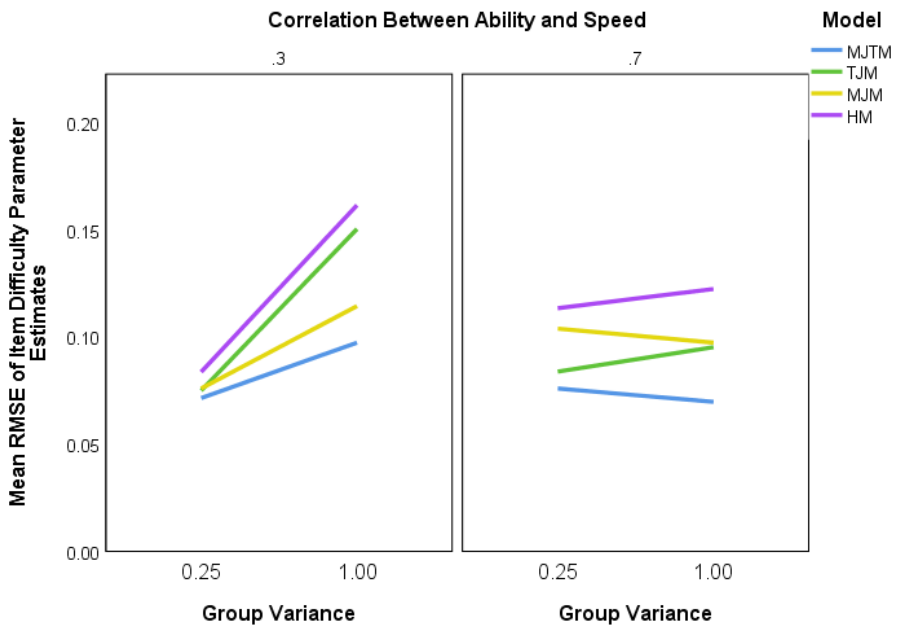

Figure 12. Significant three-way interaction of Model*group_var*theta_tau_corr on the RMSE of the item difficulty parameter estimates, b_i, when I=48.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

Item time intensity. The same analyses were conducted for the recovery of the item time intensity parameter, β_i. The results are summarized for study conditions with two different test lengths, respectively.

As presented in Table 15 for the study conditions with shorter test length, the two-way interaction between the model and the group variance had a small significant effect on the bias of the item time intensity parameter recovery (F=42.15, p<0.001, partial η^2=.022). The factor Model had a significant small effect (F=66.66, p=<0.001, partial η^2=.034). The testlet variance had a significant small effect (F=30.48, p=<0.001, partial η^2=.016).

Table 14

The ANOVA Results of the Bias of the Item Time Intensity Estimates (I=24)

Source	Bias of β_i

	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	66.66	<.001	.034
Model*group_var	42.15	<.001	.022
Between-Subject Effects			
testlet_var	30.48	<.001	.016

Note. Number of test items (I); Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$); group_var=Group variance magnitude (σ_g^2).

The significant two-way interaction is presented in Figure 13. The two-way interaction Model and group variance shows at the smaller level of σ_g^2 the mean bias of time intensity, β_i, the models have error values that appear to be very similar (note that the metric of the y-axis is in 0.005 increments). For the larger level of the σ_g^2, the mean bias of β_i has decreased substantially for the MJM, but in absolute terms of bias is nearly the same as in the smaller level of σ_g^2. The other models appear to have relatively similar mean bias of β_i.

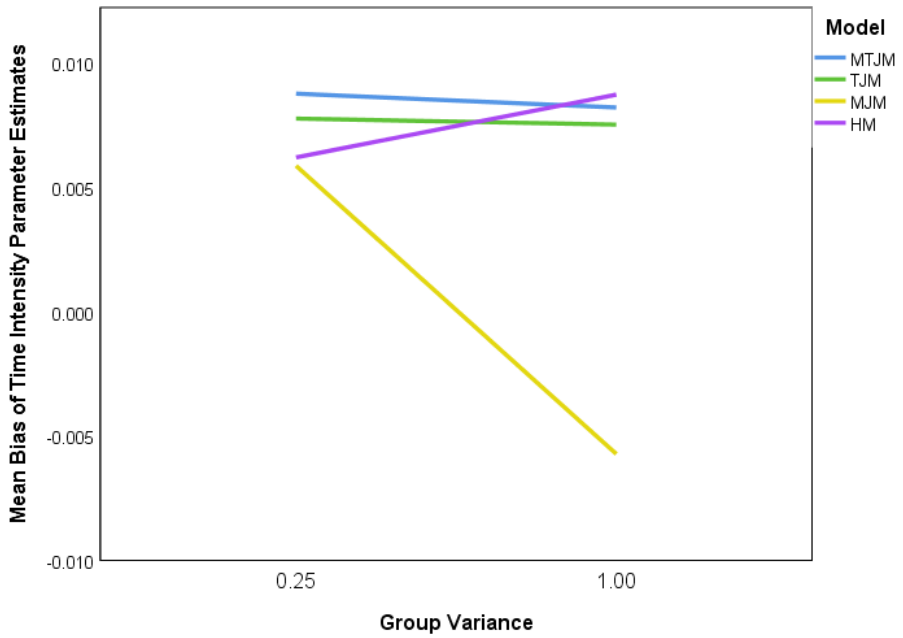

Figure 13. Significant two-way interaction of Model*group_var on the bias of the item time intensity parameter estimates, β_i, when I=24.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2).

The mixed-effect ANOVA for the RMSE of β_i when I=24 has a small three-way interaction effect among model, testlet variance, and group variance (F=38.69, p<0.001, partial η^2=.020) according to Table 15. The two-way interaction model and group variance had a significant medium effect (F=290.10, p<0.001, partial η^2=.132). The main effect of Model (F=536.78, p<0.001, partial η^2=.219) was large.

Figure 14 shows the three-way interaction among model, testlet variance, and group variance. For the small group variance (σ_g^2=0.25), the mean RMSE of β_i are very similar. For the large σ_g^2 , the models without a group parameter lead to larger mean RMSE of β_i than the models accounting for the group clustering effect. The difference was even larger for the level with a greater σ_γ^2. The main effect of group variance (F=207.19, p=<.001, partial η^2=.098) was medium. The two-way interaction of group variance and the correlation between the theta and tau parameters was small (F=24.02, p=<.001, partial η^2=.012). The three-way interaction among testlet variance, group variance, and the correlation between the theta and tau parameters had a significant small effect (F=19.67, p=<0.001, partial η^2=.010).

Table 15

The ANOVA Results of the RMSE of the Item Time Intensity Estimates (I=24)

Source	RMSE of β_i		
	F Statistics	p-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	536.78	<0.001	.219

Model*group_var	290.10	<0.001	.132
Model*testlet_var*group_var	38.69	<0.001	.020
Between-Subject Effects			
group_var	207.19	<0.001	.098
group_var*theta_tau_corr	24.02	<0.001	.012
testlet_var*group_var*theta_tau_corr	19.67	<0.001	.010

Note. Number of test items (I); Model=Estimation model type; group_var=Group variance magnitude (σ_g^2); testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

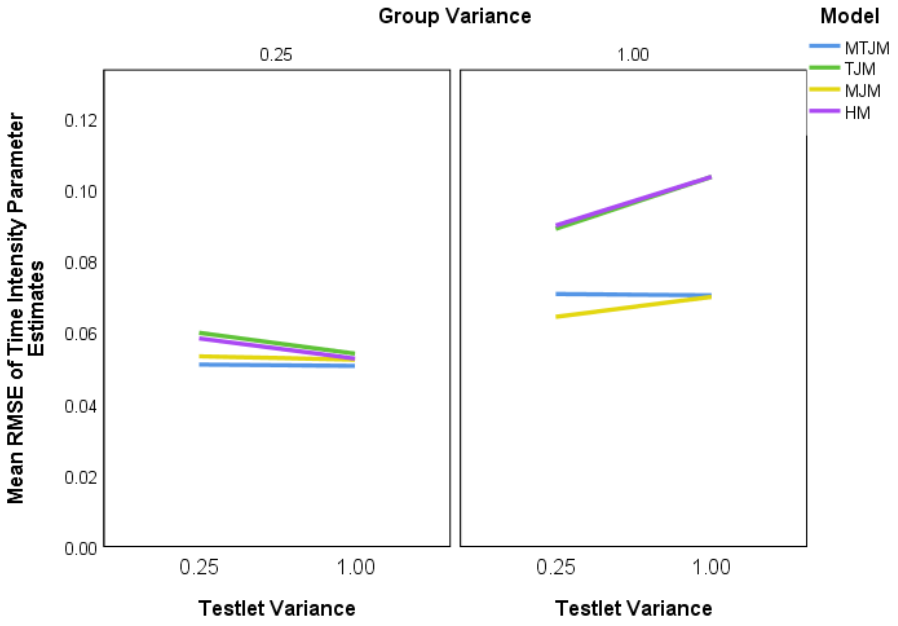

Figure 14. Significant three-way interaction of Model*testlet_var*group_var on the RMSE of the item time intensity parameter estimates, β_i, when I=24.
Note. Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2).

The estimates of the β_i for the study conditions with longer test length (I=48) were evaluated with the mixed-effect ANOVA. All the possible three-way interactions that could include the Model factor were significant. The three-way interaction among the model, testlet variance, and group variance (F=42.91, p<0.001, partial η^2=.011) is small, as are the other two three-way interactions among model,

testlet variance, and the correlation between theta and tau (F=63.88, p<0.001, partial η^2=.016) and that among model, group variance, and the correlation between theta and tau (F=51.35, p<0.001, partial η^2=.013), presented in Table 16. The disordinal three-way interaction among model, testlet variance, and group variance is depicted in Figure 15. In general, the mean bias of β_i generally is positive for the smaller σ_g^2 level with the smaller level of σ_γ^2, and negative for the larger level of σ_γ^2. In the panel that shows the larger level σ_g^2, the models show a trend of decreased mean bias of β_i, in absolute value for the larger testlet variance level, σ_γ^2. The change in bias is less pronounced for the MTJM.

Table 16

The ANOVA Results of the Bias of the Item Time Intensity Estimates (I=48)

Source	Bias of β_i		
	F Statistics	p-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	36.77	<0.001	.010
Model*theta_tau_corr	46.88	<0.001	.012
Model*testlet_var*group_var	42.91	<0.001	.011
Model*testlet_var*theta_tau_corr	63.88	<0.001	.016
Model*group_var*theta_tau_corr	51.35	<0.001	.013
Between-Subject Effects			
testlet_var*group_var	51.41	<0.001	.013
testlet_var*theta_tau_corr	276.16	<0.001	.067
group_var*theta_tau_corr	42.33	<0.001	.011

Note. Number of test items (I); Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

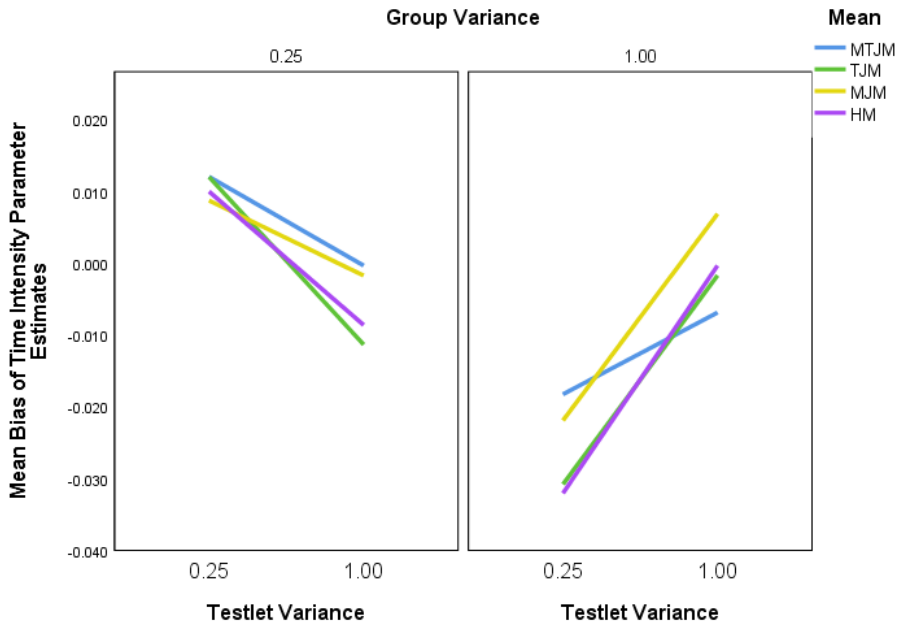

Figure 15. Significant three-way interaction of Model*testlet_var*group_var on the bias of the item time intensity parameter estimates, β_i, when I=48.
Note. Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2).

The three-way interaction among model, testlet variance, and theta-tau correlation is displayed visually in Figure 16. For the smaller level of correlation between speed and ability, $\rho_{\theta\tau}$, the absolute value of the bias for each model appears approximately the same. A larger absolute bias is observed for the larger level of $\rho_{\theta\tau}$ compared with the smaller testlet variance, σ_γ^2 . As seen in Figure 17, the interaction of σ_g^2 with $\rho_{\theta\tau}$ is disordinal. For the larger level of $\rho_{\theta\tau}$, the mean bias of β_i is larger compared to that for study conditions with smaller correlation. In addition to the three-way interactions, the two-way interaction between model and the correlation between theta and tau had a small effect on the RMSE of item intensity (F=46.88, p=<0.001, partial η^2=.012) while Model had a small effect (F=36.77, p=<0.001, partial η^2=.010) as well.

There are three significant two-way interactions between the study factors. Two of these interactions were small, specifically, that between testlet variance and group variance (F=51.41, p=<0.001, partial η^2=.013) and that between group variance and the correlation between the theta and tau parameters (F=42.33, p=<0.001, partial η^2=.011). The two-way interaction of testlet variance and the correlation between the theta and tau parameters (F=276.16, p=<0.001, partial η^2=.067) was of medium effect size.

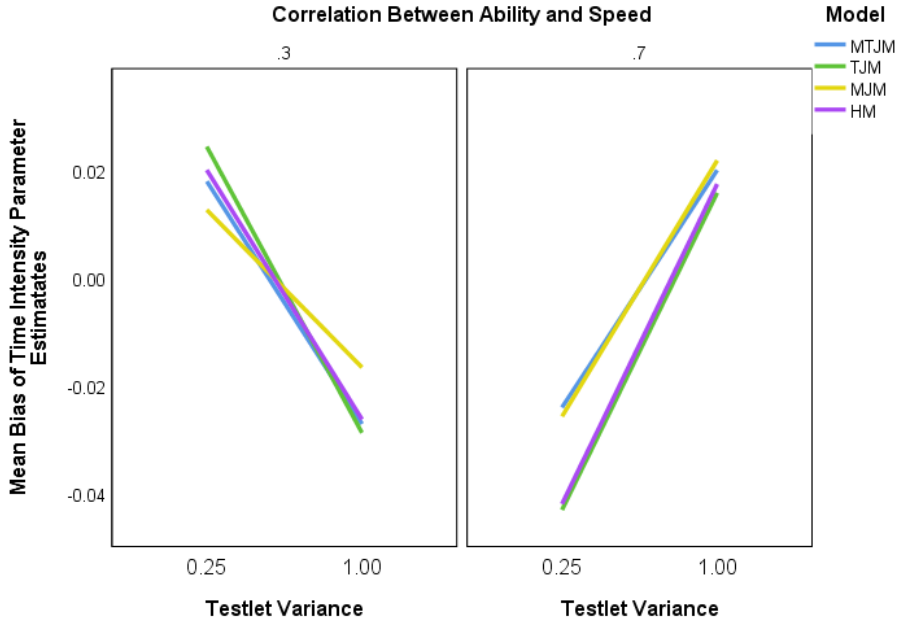

Figure 16. Significant three-way interaction of Model*testlet_var*theta_tau_corr on the bias of the item time intensity parameter estimates, β_i, when I=48.
Note. Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

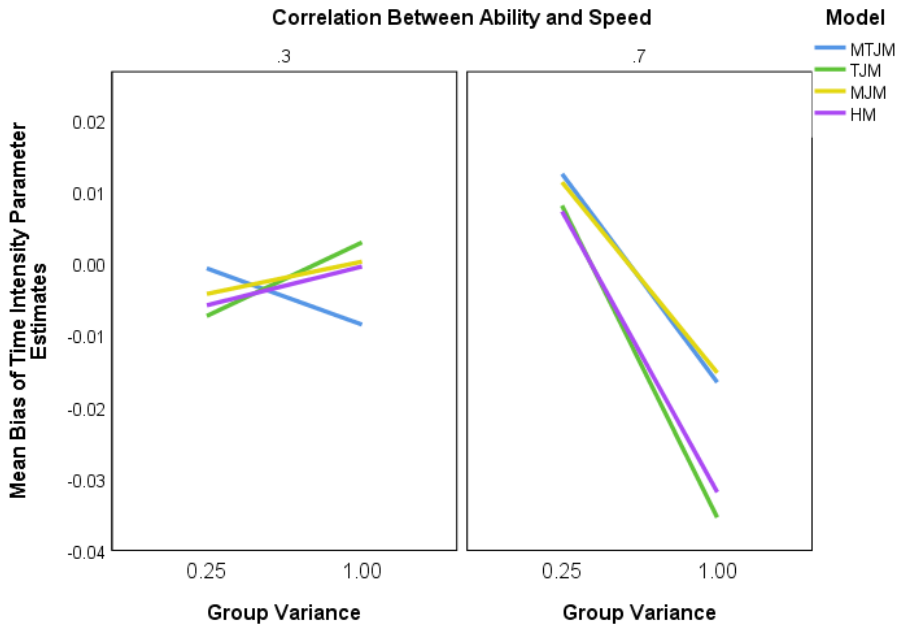

Figure 17. Significant three-way interaction of Model*group_var*theta_tau_corr on the bias of the item time intensity parameter estimates, β_i, when I=24.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

The ANOVA results for the RMSE of β_i are presented in Table 17. Two higher-order interactions for the longer test (I=48) were significant. These interaction effects had small effect sizes. The three-way interactions are model, testlet variance, and group variance (F=64.68, p<0.001, partial η^2=.017), and model, group variance, and correlation between theta and tau (F=42.96, p<0.001, partial η^2=.011). Two two-way interactions were also evaluated with a small effect for the interaction between model and testlet variance (F=66.42, p<0.001, partial η^2=.017) and with a large effect for the interaction between model and testlet variance (F=890.38, p<0.001, partial η^2=.189). The factor Model effect was large (F=2088.47, p<0.001, partial η^2=.353). The two-way interaction between group variance and the correlation between theta and tau (F=85.81, p<0.001, partial η^2=.022) was small. Two main effects were significant: testlet variance (F=45.44, p<0.001, partial η^2=.012) had a small effect

size and group variance (F=396.85, p<0.001, partial η^2=.094) had a medium effect size.

Table 17

The ANOVA Results of the RMSE of the Item Time Intensity Estimates (I=48)

Source	Bias of β_i		
	F Statistics	p-value	Partial η^2
Within-Subject Effects (with Huynh-Feldt Adjustment)			
Model	2088.47	<0.001	.353
Model*testlet_var	66.42	<0.001	.017
Model*group_var	890.38	<0.001	.189
Model*testlet_var*group_var	64.68	<0.001	.017
Model*group_var*theta_tau_corr	42.96	<0.001	.011
Between-Subject Effects			
testlet_var	45.44	<0.001	.012
group_var	396.85	<0.001	.094
group_var*theta_tau_corr	85.81	<0.001	.022

Note. Number of test items (I); Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

Figures 18 and 19 present the two significant three-way interactions. Both levels of σ_g^2 in Figure 18 show disordinal interactions where the models behave similarly for the level with smaller group variance. The models with a group parameter (MTJM and MJM) had lower mean RMSE of β_i than the models that do not have this parameter (TJM and HM). There is a slight decrease in the measurement error comparing across levels of σ_γ^2. That is, the mean RMSE of β_i is lower for the conditions with larger σ_γ^2 compared to those with smaller σ_γ^2.

In Figure 19, both levels of $\rho_{\theta\tau}$ display disordinal interactions. Within each level of $\rho_{\theta\tau}$, a larger mean RMSE of β_i is found for the larger level of group variance, σ_g^2, compared to the smaller level, for the models that do not include a

group parameter (TJM and HM) compared with the models that incorporate a group parameter (MTJM and MJM).

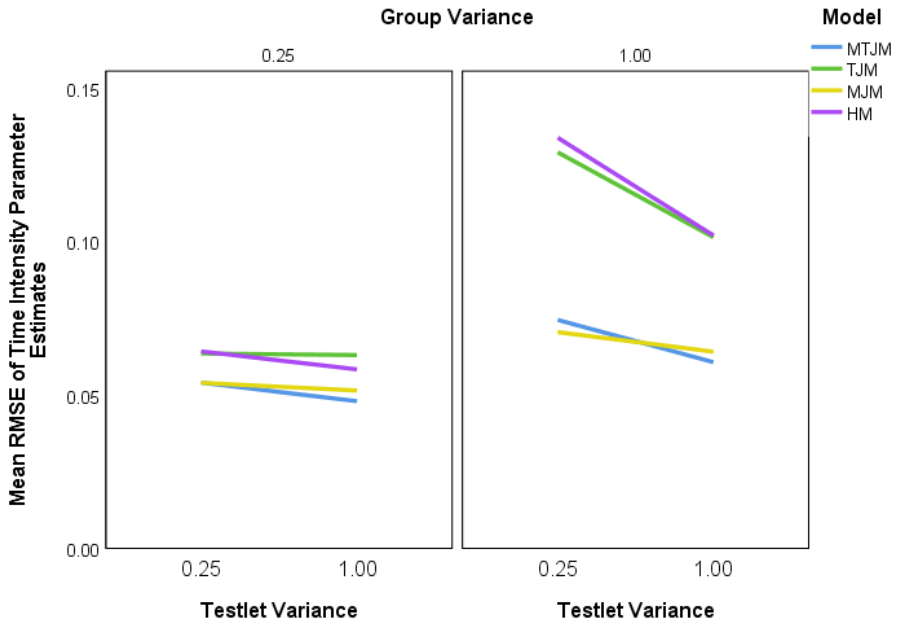

Figure 18. Significant three-way interaction of Model*testlet_var*group_var on the RMSE of the item time intensity parameter estimates, β_i, when I=48.
Note. Model=Estimation model type; testlet_var=Testlet variance magnitude (σ_γ^2); group_var=Group variance magnitude (σ_g^2).

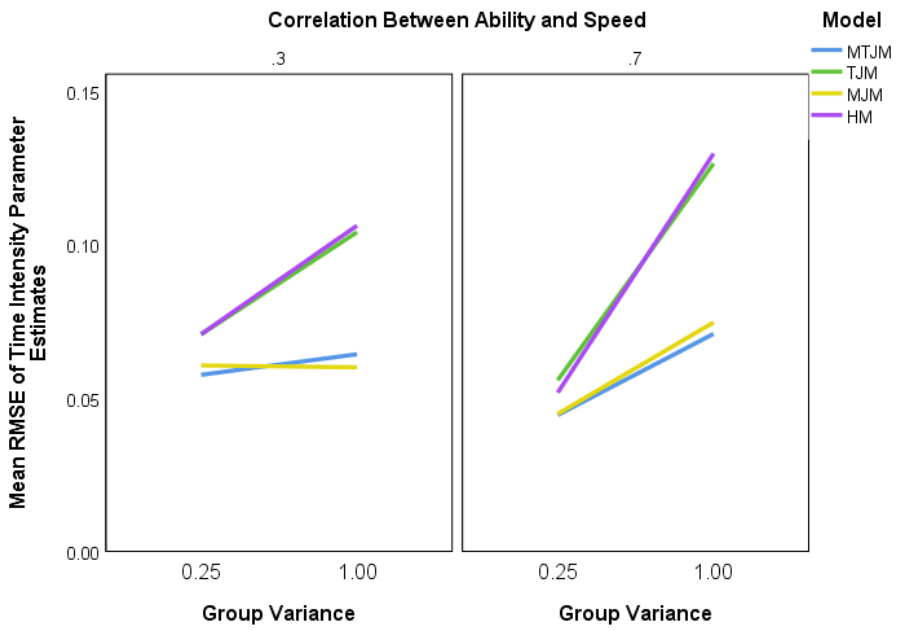

Figure 19. Significant three-way interaction of Model*group_var*theta_tau_corr on the RMSE of the item time intensity parameter estimates, β_i, when I=48.
Note. Model=Estimation model type; group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

4.1.3 Variance and Correlation Parameters

In model parameter estimation, the variance and covariance parameters were estimated. As the correlation was generated as the true values of the strength of relationship between the accuracy and speed parameters in the simulation, the estimated correlations were derived from the associated variances and covariances. This section is organized in the following presentation order (1) the estimates of variances and correlation for the person parameters, (2) the estimates of groups and testlet variances, and (3) the estimates of variances and correlation for the item parameters. The variance of the individual-specific ability was fixed at 1 in estimation for the scale identification. Summaries of the parameter estimates by condition are compared to the known "true" values. Appendix B includes the descriptive statistics for each of these parameters.

Person Speed Parameter Variance and Correlation Estimates. The variance of the person speed parameter, σ_τ^2, was estimated for the RT model. The parameter true value, 1.0, was very well recovered by the MTJM and MJM models, regardless of the manipulated level for a condition. The models that ignore the person clustering structure (TJM and HM) did not recover σ_τ^2 well. The variance of person speed was double the true value when the manipulated group variance, σ_g^2, factor was at the larger level (σ_g^2=1.0). The estimated mean σ_τ^2 summarized by condition is provided in Figure 20.

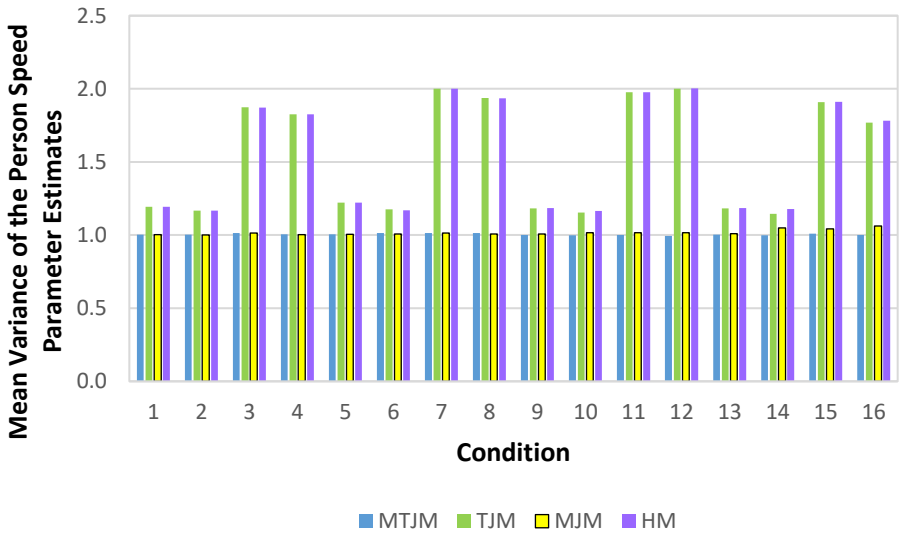

Figure 20. Mean of variance of person speed parameter, σ_τ^2 estimates by simulation condition.

The correlation between the person ability and speed parameters, $\rho_{\theta\tau}$, was estimated to determine the strength of association between the IRT and RT models for the person parameters. For the even-numbered conditions, the true value of $\rho_{\theta\tau}$ was .7, for the odd-numbered conditions the true value was .3. As seen in Figure 21, the parameter was very well recovered by the MTJM and MJM for all conditions.

The $\rho_{\theta\tau}$ parameter estimates reflect the true value for the TJM and HM in many conditions. In conditions where the manipulated group variance was larger, σ_g^2, the recovery of the $\rho_{\theta\tau}$ true values was poor; less than half of the strength of the true correlation was estimated by the models that do not account for group modeling.

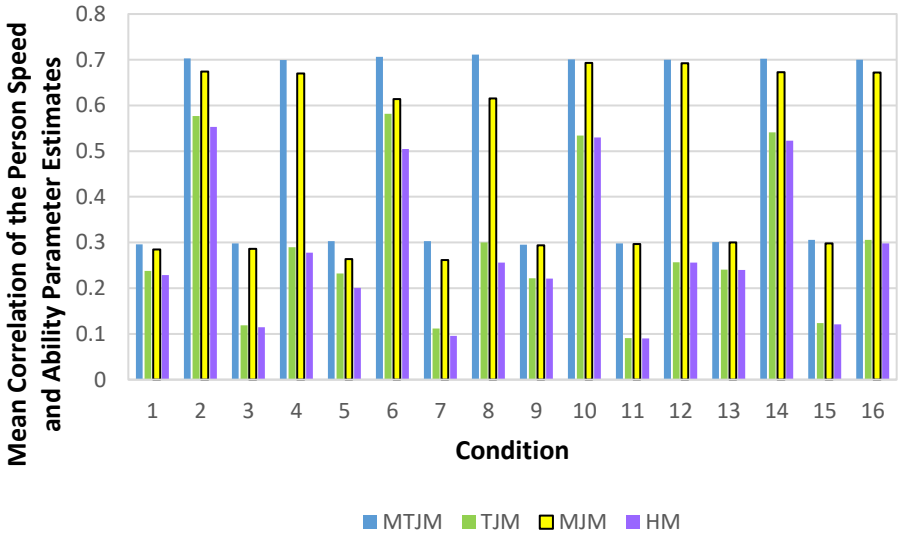

Figure 21. Mean correlation between person ability and speed parameter, $\rho_{\theta\tau}$, estimates by simulation condition.

Group Parameter Variance Estimates. Group parameters were estimated for both the IRT and RT models. In the IRT model, the group-specific ability parameter variance, $\sigma_{\theta_g}^2$, is estimated only for the models that account for multilevel structure, the MTJM and MJM. Figure 22 visually depicts the mean of the estimates of the $\sigma_{\theta_g}^2$ for the simulation study conditions. Parameter recovery was generally good. For the conditions with the smaller level of group variance ($\sigma_{\theta_g}^2$=0.25), the two models performed very similarly. For the conditions with larger group variance ($\sigma_{\theta_g}^2$=1.0), the MTJM better recovered the true value. Within this group variance level, the

recovery differences were most pronounced in conditions 7, 8, 15, and 16, conditions with larger testlet variance, σ_{γ}^{2}.

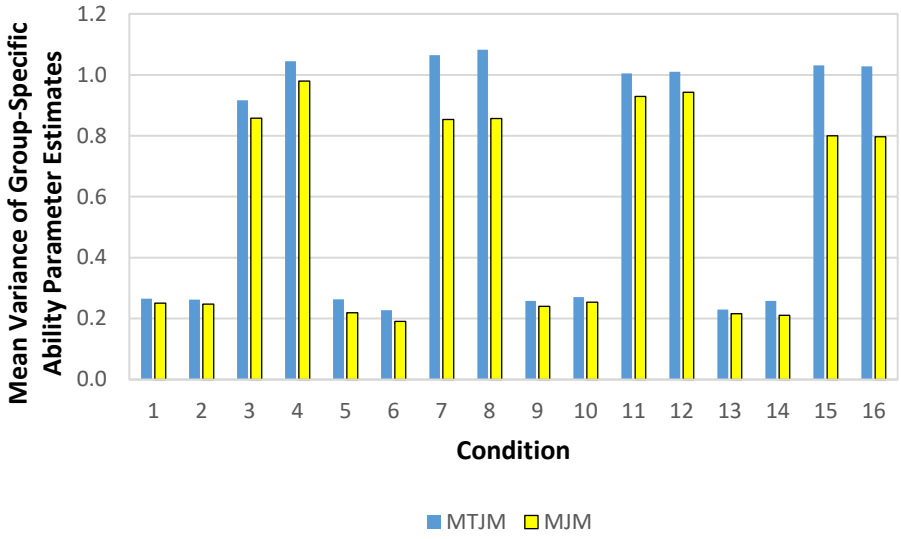

Figure 22. Mean of group-specific ability variance parameter, $\sigma_{\theta_g}^{2}$, estimates by simulation condition.

The group speed parameter variance, $\sigma_{\tau_g}^{2}$, represents the effect of group clustering in the multilevel structure in the RT model. The recovery of this parameter for the related models (MTJM and MJM) was overall quite good. In the conditions with larger group variance ($\sigma_{\theta_g}^{2}$=1.0), as seen in Figure 23, the variances may be overestimated for conditions 7, 8, 11, and 12 and underestimated for conditions 3, 4, 15 and 16. There is no clear pattern associated with other manipulated study factors.

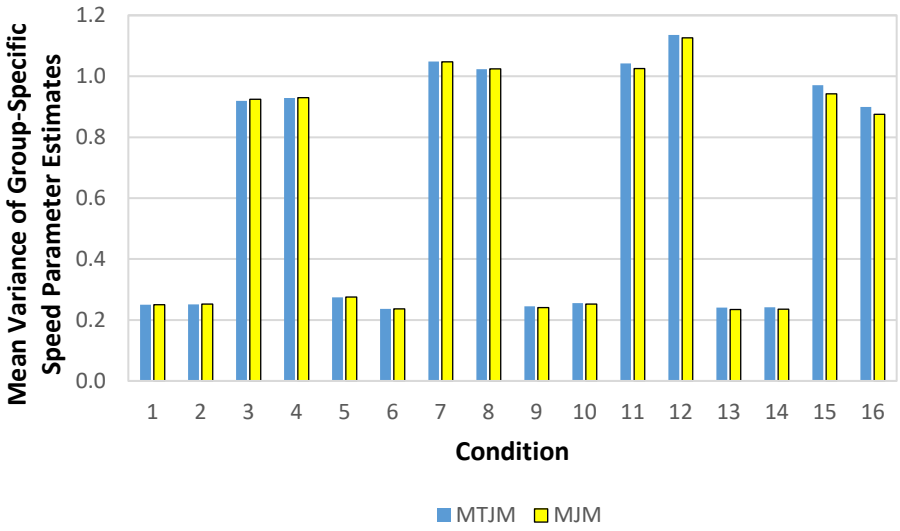

Figure 23. Mean of group-specific speed variance parameter, $\sigma^2_{\tau_g}$, estimates for estimation by simulation condition.

Testlet Variance Estimates. Two models in the simulation included testlet parameters, MTJM and TJM. All testlet variance parameters were well recovered. For ease of readability, Figure 24 provides the recovery of the variance of testlet 3. The descriptive statistics for estimated testlet variance for each testlet are provided in Appendix B.

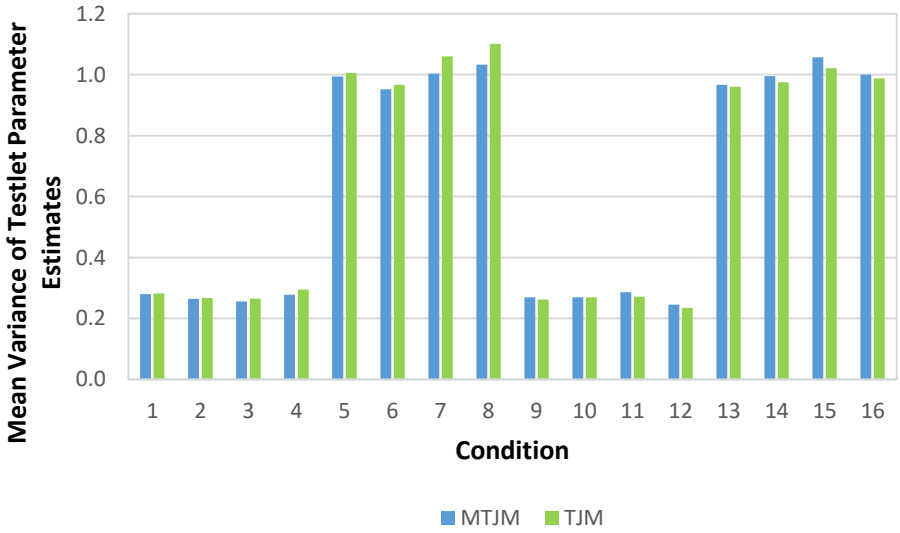

Figure 24. Mean of testlet variance parameter, σ_γ^2, estimates by simulation condition. Note. Testlet 3 values provided as an example.

Item Parameter Variance and Correlation Estimates. The variances of item difficulty parameter, σ_b^2, were estimated in the IRT model. As Figure 25 reflects, the recovery of σ_b^2 for a model that incorporates a testlet parameter, TJM, was most consistent. This model outperformed the models that ignore the local dependence of items within testlets (MJM and HM). In general, the TJM also outperformed the more complex MTJM. The differences between TJM and the models which ignore the testlet is the greatest in the conditions where the σ_γ^2 is larger (σ_γ^2=1.0 compared to σ_γ^2=0.25). The mean σ_b^2 estimated by the MTJM resulted in consistent overestimation of this parameter.

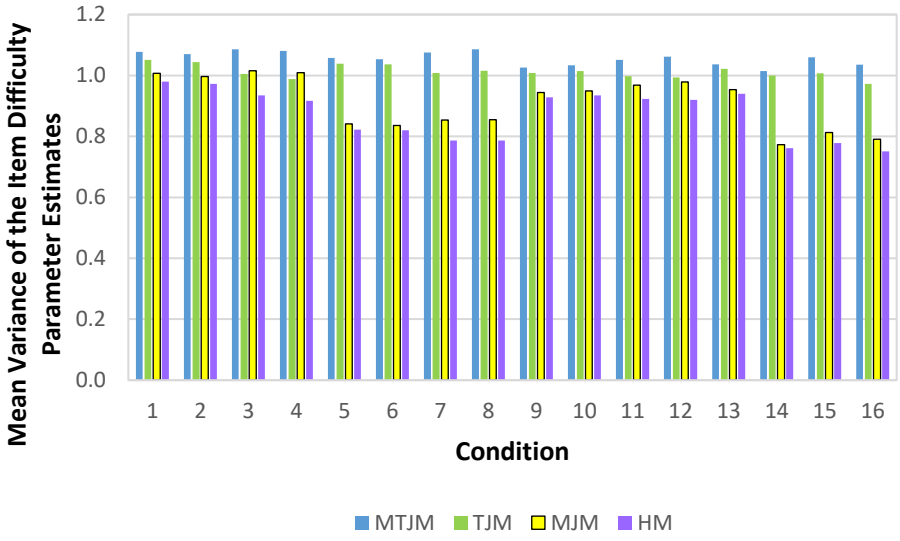

Figure 25. Mean of item difficulty variance parameter, σ_b^2, estimates by simulation condition.

The estimates of variance of item time intensity parameter, σ_β^2, were recovered extremely consistently for all models across all conditions. As shown in

Figure 26, the mean σ_{β}^{2} was slightly higher than the true value, 1.0 indicating overestimation of this parameter.

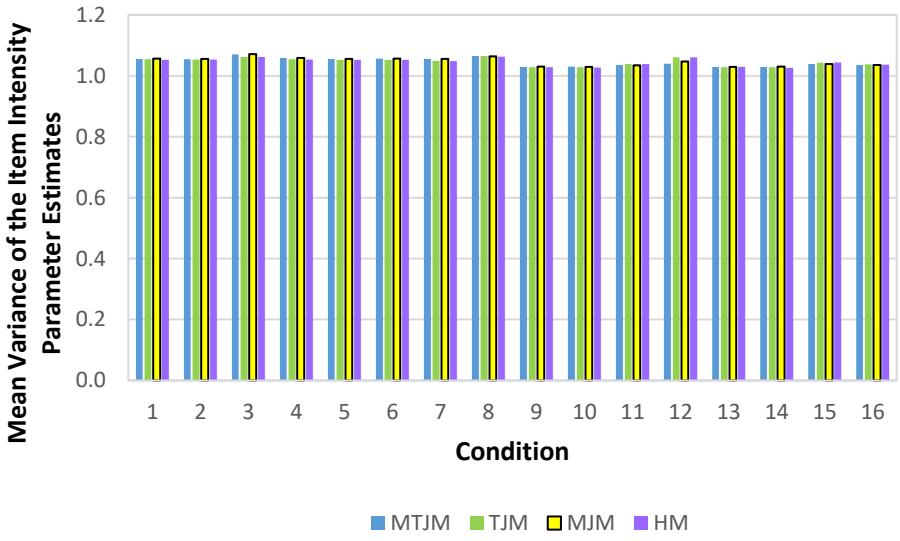

Figure 26. Mean of item intensity variance parameter, σ_{β}^{2}, estimates by simulation condition.

The strength of the relationship between item parameters in the IRT and RT models was characterized by the correlation of the item difficulty and item time intensity, $\rho_{b\beta}$. This value was fixed for the true parameter at .3 for all conditions. As seen in Figure 27, the recovery of this parameter was good. The mean estimates consistently underestimated the $\rho_{b\beta}$. There are no patterns that suggest the manipulated factors result in differences in parameter recovery performance.

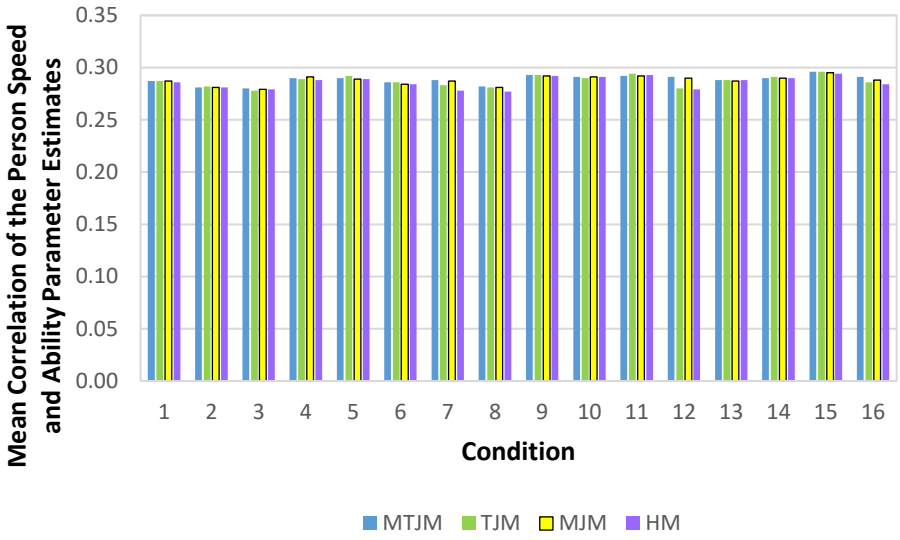

Figure 27. Mean of correlation of item difficulty and item intensity parameter, $\rho_{b\beta}$, estimates by simulation condition.

To summarize, generally, the MTJM which accounts for both LPD and LID recovered these parameters well. The person parameter estimates were also well recovered by the MJM which addresses the simulated grouping of the examinees. Person parameters were poorly recovered by the models that do not account for LDP (TJM and HM) when the effect of the person grouping was at the larger level (σ_g^2=1.0). Of the multilevel models accounting for person clustering, MTJM and MJM, the IRT group ability variance, $\sigma_{\theta_g}^2$, was recovered less well by the model which ignores the testlet effect, MJM, compared to the MTJM when the testlet variance was at the larger level (σ_γ^2=1.0). The IRT group speed variance, $\sigma_{\tau_g}^2$, was well recovered by both of these models. Overall, the models that incorporated a testlet parameter (MTJM and TJM) performed well recovering the testlet variances, σ_γ^2. For the item difficulty variance, σ_b^2, parameter recovery was mixed. The performance was poorest for the models that ignore item clustering (TJM and HM) when such effect

was larger (σ_{γ}^2=1.0). All models performed well in the recovery of the item intensity variance, σ_{β}^2, and the true item correlation parameter values, $\rho_{b\beta}$.

4.2 Model Fit

The relative model fit indices deviance and DIC was applied to model selection. Table 18 presents the frequency of selection of the true model as the best fitting model based on the deviance and DIC. Deviance may be viewed as a component of DIC. The deviance is an estimation of the -2log likelihood of the data given the parameters. DIC includes a penalty for parsimony based on the effective number of parameters.

For all conditions, deviance identified the data generating model as the best fitting model for the majority of the time. In 14 out of 16 conditions, the deviance selected the MTJM as the best fitting model 100% of the time. The DIC did not identify the data generating model as the best fitting model as successfully as the deviance. The MJM was selected by DIC as the best fitting model with the highest frequency. For 9 out of 16 conditions, DIC selected the MJM most frequently, including all conditions where the testlet variance, σ_{γ}^2, was of the smaller level. When the σ_{γ}^2 was at the larger level, the DIC selected the data generating MTJM in 75% of the conditions. In the remaining 25% of these larger σ_{γ}^2 conditions, the TJM was selected by DIC as the best fitting model. When the TJM was selected by DIC, the group variance, σ_{g}^2, was at the smaller level. In summary, deviance 87.5% of the time selected the data-generating model as the best fitting model. The results for DIC were more mixed. In more than half of the conditions, DIC identified the MJM as the best-fitting model for the majority of simulated data runs within a condition.

Table 18

Frequency of Identifying Each Model as the Best-Fitting Model in the Simulation Study

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	Deviance				DIC			
				MTJM	TJM	MJM	HM	MTJM	TJM	MJM	HM
24	.25	.25	.3	**10**	0	0	0	0	0	**10**	0
			.7	**10**	0	0	0	0	0	**10**	0
		1.0	.3	**10**	0	0	0	0	0	**10**	0
			.7	**10**	0	0	0	0	0	**10**	0
	1.0	.25	.3	**10**	0	0	0	4	**5**	0	1
			.7	**10**	0	0	0	**8**	2	0	0
		1.0	.3	**10**	0	0	0	**7**	0	3	0
			.7	**10**	0	0	0	**9**	0	1	0
48	.25	.25	.3	**10**	0	0	0	0	0	**9**	1
			.7	**10**	0	0	0	0	0	**7**	3
		1.0	.3	**10**	0	0	0	0	0	**10**	0
			.7	**8**	2	0	0	0	0	**10**	0
	1.0	.25	.3	**10**	0	0	0	0	0	**9**	1
			.7	**10**	0	0	0	4	**6**	0	0
		1.0	.3	**9**	1	0	0	**7**	0	3	0
			.7	**10**	0	0	0	**8**	0	2	0

Note. The largest numbers of replications among the three model under each condition are bolded. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

4.3 Empirical Study

The proposed model was applied for the estimation of model parameters on an empirical dataset. The dataset was excerpted from the PISA 2015 (OECD, 2017) mathematics assessment. For data cleaning the procedure was followed as described in Section 3.4. The dataset consists of 1,478 participants and 17 items, 8 of which were bundled as testlets of 2 items apiece. The items were scored dichotomously. In addition to the proposed MTJM, the three alternative models (MJM, TJM, and HM) were also used for parameter estimation. The model fit index DIC, which was described previously, was applied for the evaluation of the best fitting model to the

data. The convergence criteria that were used in the simulation study were applied in this empirical data study. In the Bayesian MCMC estimation, the $\hat{R}$ criterion for all parameter estimates less than 1.2 was achieved; the result was obtained that all parameter estimates had $\hat{R}$ values less than 1.05. Inspection of diagnostic plots supported the decision that convergence was met.

The results of the item fit index selection mirrored those of the simulation study. That is, the deviance value favored the MTJM, but when adjusted with a complexity penalty in favor of parsimony, DIC selection resulted in a model that did not address both anticipated sources of local dependence. Table 19 provides the values obtained from the DIC evaluation. DIC did not choose a more complex model, thus ignoring or minimizing the known structural dependencies of person clustering and testlet-designed items present in the assessment implementation.

Table 19

Model Fit Indices for PISA Mathematics Dataset

Model	Deviance	DIC
MTJM	**84447.30**	94481.08
TJM	84595.12	94023.53
MJM	85528.25	88529.37
HM	85616.01	**88473.75**

Note. DIC=deviance information criterion. The lowest Deviance and DIC values among the competing models are bolded. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

The parameter estimates of the MTJM and the three alternative models (TJM, MJM, and HM) are summarized in Table 20. The empirical dataset shows the presence of the clustering effects for both items (the testlet variance, σ_γ^2) and persons (the group variance, σ_g^2). Specifically, the testlet variance for the four testlets in the

mathematics subtest had means ranging from 0.265 to 1.344, in the MTJM, these values represent small to large testlet effects. In the TJM, these four testlets had means that ranged from 0.436 to 1.296. When comparing individual testlets across models with testlets, the mean magnitudes were similar. For the group clustering effects the proposed model estimated a negligible effect for the speed parameter, $\sigma^2_{\tau_g}$, and small to medium effect for the ability parameter ($\sigma^2_{\theta_g}$=0.376). The group clustering effects estimated using the MJM, yielded magnitudes much like the MTJM, negligible for the speed parameter, $\sigma^2_{\tau_g}$, and small to medium effect for the ability parameter ($\sigma^2_{\theta_g}$=0.351). The mean and standard deviation values for the individual ability, $\theta_{j(g)}$, individual speed, $\tau_{j(g)}$, item difficulty, b_i, and time intensity, β_i, differed by no more than 0.09 when the proposed model was compared to the alternative models. Appendix C provides the descriptive statistics for the person and item parameter estimates for the proposed and the alternative competing models. Overall, the parameter estimation was appeared similar for the four models. The negligible effect of the group speed parameter, $\sigma^2_{\tau_g}$, coupled with the small to medium effect of the ability parameter $\sigma^2_{\theta_g}$ provides the reasonable explanation of why neither of the multilevel models in this study were selected as best-fitting by the relative fit indices. That is, although there is complex sampling in the administration of the assessment, the data do not exhibit person clustering effects in both the response and RT models. In addition, the speed parameter variance was near zero, so the benefit in estimation to come from joint modeling of response and RT may have been muted.

Table 20

Parameter Estimates for the Data-Fitting Models

Parameters	MTJM		TJM		MJM		HM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\theta_{j(g)}$	0.002	0.802	0.002	0.889	0.000	0.800	0.000	0.887
$\tau_{j(g)}$	0.000	0.100	0.000	0.126	0.000	0.100	0.000	0.126
b_i	-0.058	0.832	-0.077	0.812	-0.055	0.811	-0.068	0.794
β_i	4.184	0.397	4.178	0.398	4.183	0.397	4.179	0.398
σ_τ^2	0.026	0.003	0.039	0.004	0.026	0.003	0.039	0.004
$\sigma_{\theta\tau}$	-0.092	0.009	-0.046	0.009	-0.093	0.009	-0.047	0.009
$\sigma_{\tau_g}^2$	0.067	0.016	-	-	0.067	0.015	-	-
$\sigma_{\theta_g}^2$	0.376	0.091	-	-	0.351	0.087	-	-
$\rho_{\theta\tau}$	-0.575	NA	-0.234	NA	-0.575	NA	-0.237	NA
$\sigma_{\gamma 1}^2$	0.441	0.137	0.436	0.147	-	-	-	-
$\sigma_{\gamma 2}^2$	0.265	0.081	0.253	0.083	-	-	-	-
$\sigma_{\gamma 3}^2$	1.344	0.283	1.296	0.279	-	--	-	--
$\sigma_{\gamma 4}^2$	0.555	0.158	0.441	0.129	-	-	-	-
σ_b^2	0.808	0.312	0.771	0.306	0.772	0.304	0.736	0.288
$\sigma_{b\beta}$	0.254	0.137	0.245	0.139	0.251	0.135	0.242	0.131
σ_β^2	0.236	0.092	0.235	0.095	0.236	0.091	0.234	0.092
$\rho_{b\beta}$	0.581	NA	0.575	NA	0.589	NA	0.583	NA

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model. Correlation parameters are derived from the respective variance-covariance matrices. NA= not available. Testlets are identified in the subscript of the testlet variances. Parameters that are not present in a model are represented with "-".

Chapter 5: Discussion

The joint modeling of response and response time using IRT and RT has received increasing attention as the computational capabilities and access to process data expands (e.g., Im, 2015; Liao, 2018; Man et al., 2019). The present study investigates local dependence due to item clustering and person clustering in the joint modeling of responses and response time. The proposed multilevel testlet joint model (MTJM) simultaneously addressed both types of local dependence. The performance of the proposed model was evaluated in an experiment that manipulated variables frequently employed in this line of research and compared the proposed model to models that addressed neither or only one type of local dependence. The chapter is organized as follows: findings from the simulation and empirical study are summarized first, and then limitations and future directions are elaborated.

5.1 Summary of the Study Results

In this section, findings from the simulation study regarding the impact of the manipulated factors in terms of the recovery of model parameters as well as model selection in the simulation and empirical analyses were discussed. Generally, parameter recovery was good for the joint response and RT models accounting for dual local dependence for the individual person and item parameters as measured in terms of bias and RMSE; the recovery of the variance and correlation parameters was not performed as successfully by all models. The mixed-effect ANOVAs were conducted for the estimated parameters of person ability, person speed, item difficulty, and item time intensity. Only the effects which are statistically significant and practically significant effects with at least small effect sizes were reported and

discussed. The discussions related to interaction effects were up to three-way for easiness of interpretation. Further, the high-level interactions of the studied factors were also visually represented and discussed. The parameters that were not evaluated in ANOVAs were summarized by comparing the estimate to the known true value per condition. The findings are summarized in responding to each of the research questions proposed for this dissertation research below.

How do the manipulated factors affect the proposed model parameter estimates?

The results from the ANOVAs of the parameters estimated for the proposed MTJM indicate that outcome measures, i.e., the measurement errors of the model parameter estimates, were impacted by at least one of the manipulated factors, namely the number of items (test length), the testlet effects, the person clustering effects, and the correlation of ability and speed. For the test length variable there were two levels, a shorter test with 24 items and a longer test with 48 items. The testlet effect factor was operationalized in this study by manipulating two levels of testlet variance (σ_γ^2=0.25 and σ_γ^2=1.0). The group variances for the two levels were σ_g^2=0.25 and σ_g^2 =1.0. The correlation between person ability and speed ($\rho_{\theta\tau}$) was manipulated to provide another factor in the investigation. The two levels of this factor were $\rho_{\theta\tau}$=.3 and $\rho_{\theta\tau}$=.7.

Test length had a small effect (F=6008.03, p<0.001, partial η^2=.029) on the RMSE of the ability parameter, θ_j, as did testlet variance (F= 2496.93, p<0.001, partial η^2=.012). Further, test length also had a small effect (F= 8678.32, p<0.001, partial η^2=.042) on the RMSE of the speed parameter, τ_j.

The estimation errors for item parameters were analyzed separately by test length (I=24, I=48). For the shorter test, the testlet variance factor has a significant small effect on the bias of item difficulty, b_i (F= 25.33, p<0.001, partial η^2=.013). The group variance factor had medium effects on the bias (F= 162.20, p<0.001, partial η^2=.078) and on the RMSE (F= 166.22, p<0.001, partial η^2=.080) of item difficulty. The univariate ANOVA of the results from the longer tests indicate a significant small two-way interaction among group variance and the correlation between theta and tau (F= 71.11, p<0.001, partial η^2=.018) on the RMSE of the item difficulty, b_i. The theta-tau correlation also had a small effect (F= 37.08, p<0.001, partial η^2=.010) on the RMSE of item difficulty, b_i. For the longer test, no study factors produced any statistically significant effects on the bias for b.

For the shorter test, the testlet variance affected the bias of time intensity, β_i with a small effect (F= 26.00, p<0.001, partial η^2=.013). The correlation between theta and tau had a small effect (F= 25.75, p<0.001, partial η^2=.013) on the bias of time intensity, β_i. Further, group variance had a small effect (F= 97.43, p<0.001, partial η^2=.048) on RMSE.

For the longer tests, most factors had significant interactions effects on the item time intensity parameter,β_i. The two-way interaction between the testlet variance and the correlation between theta and tau had medium effect (F= 386.34, p<0.001, partial η^2=.092) on the bias of β_i. So did the group variance with a small effect (F= 65.83, p<0.001, partial η^2=.017). In addition, group variance had a small effect on the RMSE of β_i (F= 132.05, p<0.001, partial η^2=.033). There was a significant two-way interaction between group variance and theta-tau correlation on

RMSE of β_i with small effect size (F= 46.72, p<0.001, partial η^2=.012). The factor testlet variance had a small effect (F= 47.05, p<0.001, partial η^2=.012) on the RMSE of β_i. The summary tables for all the significant univariate ANOVAs are provided in Appendix D.

How do violations of local person independence and local item independence affect parameter recovery when fitting the data with standard joint models of response and RT?

The simulation study included the proposed model which accounts for local item and local person dependency (MTJM) and three models which do not include a parameter to account for the local person dependency, the local person dependency, or both of these dependencies (the TJM, MJM, and HM, respectively). The violations of local independence assumptions were found to affect the person parameters, item parameters, and the variance and correlation parameters. The results of the mixed-effect ANOVAs and descriptive analyses provide the answers to this research question.

For the person ability parameter, two significant interactions: model and test length, and model and group variance were found on the RMSE. Test length had a smaller effect. The mean RMSE were smaller for the models that incorporate a group parameter to address the LPD.

For the RT speed parameter, the interaction between the model and group variance had a significant effect on RMSE. As was observed for the ability parameter recovery, the models that account for LPD performed better in the speed parameter recovery.

For the item difficulty parameters, the interactions between models and group variances and the interaction between models and testlet variances had significant effects. Specifically, in the short test length conditions, the interaction between models and group variances had significant effects on both bias and RMSE. For this RMSE, the interaction of model and testlet was also significant. For the longer test lengths, the three-way interactions among the group variance, testlet variance, and the correlation of the ability and speed had an impact on the RMSE.

The bias of the item time intensity in the short test conditions were affected by the model and group variance interaction while the RMSE were affected by the three-way interaction among the model, testlet variance, and group variance. On the other hand, the two-way interactions between group and testlet variance factors had an effect on the bias and the RMSE.

It was found that the group variance and the correlation parameters for the person related model parameters were better recovered by the models that incorporate the group parameter (MTJM and MJM). The group variance for the person ability parameters estimated by the MJM was adversely affected by ignoring the LID. For the item difficulty parameter variance, the models that did not incorporate the testlet parameter performed more poorly in estimating the true testlet variances.

These results indicate that that ignoring LID, LPD or both have negative impact on the recovery of true model parameters. Overall, the proposed model performed better in terms of parameter recovery than the models that did not incorporate a parameter for local dependence when local dependence was present.

How does model selection perform for the proposed model compared to alternative competing models when LID and/or LPD is ignored in simulated and empirical data analysis?

A Bayesian model fit index was used to model selection. The deviance and deviance information criterion (DIC) was calculated for each model. The DIC selected the multilevel joint model (MJM) more frequently than the other models including the proposed multilevel testlet joint model (MTJM). The deviance, without the penalty imposed in the DIC, was best performing in terms of identifying the true model for data generation. The better performance for model selection by deviance was also reported in Liao (2018). In this study, deviance selected the proposed model as the best fitting model in 98% of the replications.

An empirical dataset from the PISA 2015 mathematics was analyzed to provide a comparison of model selection for the proposed model and three alternative competing models. The competing models were the same as those in the simulation study. For this dataset, the MTJM was again not selected by DIC as the best-fitting model. Instead the HM, which does not include parameters to accommodate LID and LPD, was selected by DIC as the best fitting model. Like in the simulation study, the deviance selected the proposed model as the best-fitting model. The parameters estimated by the proposed model for the testlets and for the group variances were very similar to those estimated by the models that included a parameter for these effect (the TJM and MJM, respectively). This indicates that there is presence of LID and of marginal LPD in the empirical dataset, although the DIC fit index did not identify the proposed model as the best fitting model.

5.2 Limitations and Future Directions

This study proposed a joint response and RT model that addresses complex local dependence issues often present in assessment contexts. As such, different study conditions were simulated to explore model performance under different assessment contexts that are consistent with prior research and is appropriately scoped for this current independent investigation. The following presents the limitations of this study. Five areas were identified for future exploration.

Model Extensions

The present study proposed one type of measurement model of RA and one type of RT to develop the joint model. The models used for comparison were parameterized to provide a means of evaluation of similar models that differed only with respect to ignoring the LID, LPD, or both. There are several alternatives for modeling that could be used such as the IRT model for polytomous items in a testlet (e.g., Huang & Wang, 2014; Jiao & Zhang, 2015), RT where the speed is not fixed (e.g., Bolsinova et al., 2017, Fox & Marianti, 2016; Meng, Tao, & Chang, 2015), RT that includes a testlet (e.g., Im, 2015), and joint models of response and RT that includes covariates (e.g., Klein Entink, Fox, and van der Linden, 2009). Future study may explore possible overfitting of the data further through simulation where multiple generating models are used.

RT Model Extensions

Assumptions of how the data are modeled were considered in the design of this study. For the RT model the lognormal distribution was used for data generation and data modeling. This is based on the relative simplicity of applying this distribution and its wide application (e.g., van der Linden, 2006, 2007). Other

possible distributions may be explored instead such as the Box-Cox normal model (e.g., Klein Entink, van der Linden, and Fox, 2009), or the Semiparametric Cox proportional hazards (PH) model (e.g., Ranger & Ortner, 2012; Wang, Fan, Chang, & Douglas, 2013).

Prior Distributions

For the simulation study, the means of the item parameters were fixed. These values were based on the study by Liao (2018). An analysis of a sample run of the estimation for each of the four models found no consequential differences in item parameter estimates compared to when the means had normal hyperprior distributions. Klein Entink, Fox, and van der Linden (2009) performed an analysis of simulated datasets that investigated the sensitivity of different priors for the person correlation parameter. Similar sensitivity analyses on the effect of varying priors and hyperpriors in joint response and RT modeling would be of benefit to the research community.

The RT model had a fixed item discrimination parameter to provide an analogous model to the IRT Rasch model for RA. An explicit time discrimination parameter was specified in the joint multilevel model of Klein Entink, Fox, and van der Linden (2009). Some other newly proposed joint models have also included this parameter (e.g., Liao, 2018, Man et al., 2019). Future study extending the proposed model may include varying the discrimination parameter in the RT model.

Model Fit

A relative model fit index, DIC, was used in this study for model comparison in both simulation and empirical data analyses. Recent joint response and RT models

(e.g., Im, 2015; Man et al., 2019) applied only the DIC as a relative model fit index. Other relative model fit indices are available to the practitioner for model comparison such as Akaike's information criterion (AIC; Akaike, 1987), AIC adjusting for small sample sizes (AICc; Sugiura, 1978), and Bayesian information criterion (BIC; Schwarz, 1978). In addition to relative model fit indices, other methods of fit analysis such as the posterior predictive model check (PPMC, Guttman, 1967; Rubin, 1981, 1984) could be used in future research on other extended joint response and RT modeling.

Other Factors to Consider

In multilevel studies, having a larger number of groups is usually favored over more participants within a group (e.g., Gelman & Hill, 2007; Raudenbush & Bryk, 2002). However, future studies extending this current joint multilevel response and RT modeling may investigate the impact of the cluster number as well as the cluster sizes. In addition, the number of testlets and the number of items per testlet can be other manipulated factors for future exploration.

To sum up, this study proposed a joint response and RT model that account for local item dependence and local person dependence in analyzing response and RT data. The model parameter s could be well recovered based on the Bayesian approach explored in this research study. Further, the comparison of the proposed model and three competing models that ignore one or both types of local dependence found in complex sampling of persons and items bundled in testlets revealed the impact of ignoring local dependence on the ability, person speed parameters, item difficulty and item time intensity parameters.

The collection of process and product data in the computer-based assessment is expanding dramatically. Test stakeholders are gaining easier access to a wealth of data in this information age. As the modeling of data from multiple sources may also increase in terms of access and complexity, this study intends to contribute in this area.

Appendix A

Table A. 1

Bias and RMSE of the Estimates of the Person Ability Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	Bias of θ_j				RMSE of θ_j			
				MTJM	TJM	MJM	HM	MTJM	TJM	MJM	HM
24	.25	.25	.3	-	-.003	-	-	.392	.487	.394	.496
			.7	-	-.001	-	-.002	.360	.455	.363	.471
		1.0	.3	.001	-.001	.001	-	.398	.666	.399	.685
			.7	.001	-	.002	-	.367	.676	.370	.702
	1.0	.25	.3	-	-	-	.001	.476	.537	.486	.565
			.7	.001	-	-	.001	.423	.476	.451	.524
		1.0	.3	.002	-	.002	.004	.485	.686	.496	.750
			.7	.002	.002	.001	.002	.430	.657	.458	.737
48	.25	.25	.3	-	.001	.001	-.002	.298	.443	.299	.445
			.7	.001	.001	.002	-	.283	.431	.285	.437
		1.0	.3	-.001	-	.001	-.001	.302	.713	.303	.715
			.7	.001	-.003	.003	-.002	.288	.713	.293	.718
	1.0	.25	.3	-	-	.001	-.002	.296	.428	.297	.429
			.7	-	-.003	.001	-.003	.351	.446	.364	.468
		1.0	.3	-	.001	.002	.002	.384	.719	.388	.730
			.7	-	-.002	.001	.001	.347	.692	.365	.712

Note. Bias values that approach 0 (i.e., -.001<Bias<.001) are represented with "-".
MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

Table A. 2

Bias and RMSE of the Estimates of the Person Speed Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	Bias of τ_j				RMSE of τ_j			
				MTJM	TJM	MJM	HM	MTJM	TJM	MJM	HM
24	.25	.25	.3	.001	-.001	-	-.003	.171	.404	.171	.405
			.7	-	-.002	-	-.001	.171	.405	.171	.405
		1.0	.3	-	-.001	-	.003	.173	.765	.173	.765
			.7	.001	.001	.001	-.004	.169	.761	.169	.761
	1.0	.25	.3	-.001	-.001	-.001	-.005	.173	.427	.173	.427
			.7	.001	-.001	-	-.001	.172	.391	.171	.391
		1.0	.3	-	-.007	-.001	.002	.172	.806	.171	.806
			.7	.001	.002	-	-.002	.174	.814	.174	.814
48	.25	.25	.3	.001	-	-.001	-.002	.127	.386	.134	.386
			.7	.001	.003	.001	-	.128	.390	.132	.390
		1.0	.3	-	.001	-.001	-.007	.116	.816	.130	.816
			.7	.002	-.001	.003	.004	.117	.853	.128	.853
	1.0	.25	.3	.001	-	.001	.004	.125	.389	.133	.389
			.7	.001	-.006	.001	-.004	.126	.377	.132	.377
		1.0	.3	-	-.002	.001	-.001	.117	.791	.136	.791
			.7	-.005	-.008	-.004	-.006	.122	.761	.142	.761

Note. Bias values that approach 0 (i.e., -.001<Bias<.001) are represented with "-".
MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

Table A. 3

Bias and RMSE of the Estimates of the Item Difficulty Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	Bias of b_i				RMSE of b_i			
				MTJM	TJM	MJM	HM	MTJM	TJM	MJM	HM
24	.25	.25	.3	.033	.032	.033	.033	.077	.079	.079	.083
			.7	.018	.023	.021	.021	.070	.072	.071	.076
		1.0	.3	-.005	-.010	-.008	-.009	.133	.164	.129	.165
			.7	-.055	-.080	-.051	-.075	.103	.133	.098	.141
	1.0	.25	.3	.017	.017	.013	.017	.068	.068	.107	.114
			.7	.001	.001	.001	.002	.069	.073	.116	.122
		1.0	.3	-.078	-.113	-.067	-.097	.110	.155	.134	.180
			.7	-.052	-.079	-.056	-.071	.107	.132	.131	.170
48	.25	.25	.3	.006	.004	.004	.001	.064	.068	.072	.078
			.7	.001	-	.003	-.001	.077	.085	.079	.090
		1.0	.3	.011	.019	.020	.017	.093	.145	.095	.148
			.7	.008	.016	.008	.017	.069	.095	.072	.097
	1.0	.25	.3	.017	.020	.016	.017	.079	.082	.080	.089
			.7	.025	.023	.018	.021	.075	.082	.128	.136
		1.0	.3	-.017	-.017	-.013	-.013	.101	.155	.134	.174
			.7	-	-.007	-.001	-.002	.070	.095	.122	.147

Note. Bias values that approach 0 (i.e., -.001<Bias<.001) are represented with "-".
MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

Table A. 4

Bias and RMSE of the Estimates of the Item Time Intensity Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	Bias of β_i				RMSE of β_i			
				MTJM	TJM	MJM	HM	MTJM	TJM	MJM	HM
24	.25	.25	.3	-.013	-.015	-.014	-.017	.049	.061	.053	.058
			.7	.018	.015	.012	.016	.052	.059	.053	.059
		1.0	.3	.003	-.003	-.016	.001	.070	.089	.063	.088
			.7	-.009	-.012	-.020	-.017	.072	.088	.066	.092
	1.0	.25	.3	.005	.009	.003	.004	.053	.059	.057	.059
			.7	.025	.022	.022	.022	.049	.049	.047	.046
		1.0	.3	.004	.013	-.001	.022	.057	.089	.057	.088
			.7	.034	.032	.015	.029	.083	.118	.082	.119
48	.25	.25	.3	.022	.022	.020	.021	.061	.070	.062	.074
			.7	.002	.002	-.003	-.001	.047	.057	.046	.055
		1.0	.3	.014	.027	.005	.019	.068	.107	.054	.115
			.7	-.051	-.088	-.049	-.083	.081	.151	.087	.153
	1.0	.25	.3	-.023	-.037	-.029	-.032	.054	.071	.059	.068
			.7	.023	.014	.025	.015	.042	.055	.044	.049
		1.0	.3	-.031	-.021	-.005	-.020	.060	.101	.066	.098
			.7	.017	.017	.018	.020	.061	.102	.062	.106

Note. Bias values that approach 0 (i.e., -.001<Bias<.001) are represented with "-".
MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

Appendix B

Table B. 1

MTJM and TJM Estimates of Variance of the Individual-Specific Speed Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\tau_{jg}}^2$				TJM $\sigma_{\tau_{jg}}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.992	1.017	1.003	0.007	1.148	1.290	1.194	0.043
			.7	0.981	1.020	1.002	0.013	1.110	1.213	1.167	0.036
		1.0	.3	0.998	1.042	1.015	0.013	1.577	2.064	1.873	0.158
			.7	0.979	1.035	1.005	0.015	1.572	2.000	1.826	0.156
	1.0	.25	.3	0.979	1.021	1.005	0.012	1.158	1.302	1.222	0.053
			.7	0.992	1.024	1.015	0.009	1.091	1.234	1.175	0.040
		1.0	.3	0.991	1.042	1.015	0.015	1.707	2.299	2.000	0.191
			.7	0.970	1.043	1.015	0.024	1.623	2.340	1.937	0.241
48	.25	.25	.3	0.988	1.013	1.000	0.007	1.116	1.238	1.183	0.047
			.7	0.990	1.015	0.999	0.008	1.088	1.201	1.153	0.037
		1.0	.3	0.991	1.010	1.001	0.006	1.740	2.158	1.976	0.152
			.7	0.977	1.024	0.994	0.015	1.589	2.375	2.000	0.258
	1.0	.25	.3	0.985	1.015	1.003	0.009	1.117	1.263	1.182	0.044
			.7	0.981	1.017	0.999	0.015	1.064	1.224	1.146	0.053
		1.0	.3	0.997	1.023	1.010	0.008	1.539	2.168	1.908	0.188
			.7	0.980	1.020	1.000	0.017	1.574	2.090	1.769	0.174

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 2

MJM and HM Estimates of Variance of the Individual-Specific Speed Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MJM $\sigma_{\tau_{jg}}^2$				HM $\sigma_{\tau_{jg}}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.991	1.017	1.003	0.007	1.145	1.290	1.194	0.043
			.7	0.979	1.015	1.000	0.012	1.109	1.211	1.166	0.036
		1.0	.3	0.998	1.043	1.014	0.013	1.577	2.065	1.872	0.158
			.7	0.977	1.031	1.002	0.015	1.573	1.998	1.825	0.154
	1.0	.25	.3	0.978	1.022	1.005	0.012	1.158	1.302	1.221	0.054
			.7	0.988	1.017	1.008	0.009	1.084	1.230	1.170	0.040
		1.0	.3	0.989	1.041	1.013	0.015	1.709	2.300	2.000	0.191
			.7	0.961	1.035	1.007	0.024	1.619	2.337	1.934	0.241
48	.25	.25	.3	1.000	1.028	1.007	0.009	1.117	1.241	1.185	0.047
			.7	1.008	1.031	1.017	0.007	1.096	1.214	1.164	0.038
		1.0	.3	0.992	1.055	1.016	0.019	1.741	2.160	1.976	0.151
			.7	0.992	1.053	1.016	0.017	1.592	2.378	2.004	0.258
	1.0	.25	.3	0.997	1.020	1.010	0.008	1.119	1.264	1.184	0.044
			.7	1.035	1.068	1.049	0.014	1.092	1.261	1.179	0.053
		1.0	.3	1.005	1.160	1.042	0.045	1.542	2.170	1.912	0.189
			.7	1.025	1.136	1.063	0.033	1.588	2.105	1.783	0.173

Note. MJM= Multilevel Joint Model, HM= Hierarchical Model.

Table B. 3

MTJM and TJM Derived Estimates of Correlation of the Individual-Specific Ability and Speed Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\rho_{\theta\tau}$				TJM $\rho_{\theta\tau}$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.286	0.323	0.296	0.012	0.193	0.270	0.238	0.026
			.7	0.692	0.724	0.703	0.010	0.533	0.614	0.577	0.025
		1.0	.3	0.283	0.322	0.298	0.013	-0.062	0.238	0.119	0.086
			.7	0.684	0.716	0.699	0.010	0.209	0.397	0.290	0.068
	1.0	.25	.3	0.265	0.335	0.303	0.025	0.138	0.272	0.232	0.037
			.7	0.659	0.734	0.706	0.022	0.517	0.642	0.582	0.041
		1.0	.3	0.278	0.374	0.303	0.031	0.026	0.290	0.112	0.078
			.7	0.685	0.740	0.711	0.018	0.211	0.426	0.300	0.070
48	.25	.25	.3	0.274	0.322	0.295	0.018	0.194	0.258	0.222	0.022
			.7	0.683	0.713	0.701	0.010	0.472	0.579	0.534	0.034
		1.0	.3	0.284	0.327	0.298	0.012	0.023	0.230	0.091	0.057
			.7	0.683	0.713	0.700	0.010	0.139	0.379	0.257	0.076
	1.0	.25	.3	0.281	0.317	0.301	0.010	0.202	0.289	0.241	0.033
			.7	0.679	0.731	0.702	0.018	0.468	0.613	0.541	0.041
		1.0	.3	0.271	0.339	0.306	0.018	0.028	0.224	0.124	0.069
			.7	0.686	0.722	0.700	0.011	0.229	0.346	0.306	0.040

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 4

MJM and HM Derived Estimates of Correlation of the Individual-Specific Ability and Speed Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MJM $\rho_{\theta\tau}$				HM $\rho_{\theta\tau}$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.275	0.312	0.285	0.012	0.184	0.261	0.229	0.025
			.7	0.663	0.692	0.674	0.010	0.509	0.589	0.553	0.024
		1.0	.3	0.270	0.308	0.286	0.013	-0.059	0.228	0.115	0.082
			.7	0.654	0.687	0.670	0.011	0.199	0.383	0.278	0.066
	1.0	.25	.3	0.234	0.289	0.264	0.020	0.122	0.234	0.201	0.031
			.7	0.575	0.643	0.614	0.019	0.451	0.553	0.505	0.034
		1.0	.3	0.242	0.325	0.262	0.028	0.020	0.252	0.096	0.068
			.7	0.593	0.644	0.615	0.016	0.179	0.363	0.256	0.061
48	.25	.25	.3	0.274	0.320	0.294	0.017	0.192	0.256	0.221	0.022
			.7	0.676	0.705	0.693	0.010	0.468	0.575	0.530	0.034
		1.0	.3	0.282	0.325	0.297	0.012	0.023	0.229	0.090	0.057
			.7	0.675	0.708	0.692	0.010	0.138	0.377	0.256	0.075
	1.0	.25	.3	0.279	0.316	0.300	0.010	0.200	0.288	0.240	0.033
			.7	0.653	0.698	0.673	0.016	0.453	0.589	0.523	0.038
		1.0	.3	0.260	0.328	0.298	0.019	0.030	0.219	0.121	0.068
			.7	0.658	0.696	0.672	0.012	0.224	0.338	0.298	0.038

Note. MJM= Multilevel Joint Model, HM= Hierarchical Model.

Table B. 5

MTJM and MJM Estimates of Variance of the Group-Specific Ability Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\theta_g}^2$				MJM $\sigma_{\theta_g}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.196	0.345	0.265	0.045	0.185	0.323	0.251	0.042
			.7	0.210	0.324	0.262	0.045	0.199	0.310	0.247	0.043
		1.0	.3	0.569	1.145	0.917	0.168	0.528	1.080	0.858	0.158
			.7	0.713	1.458	1.045	0.243	0.672	1.367	0.980	0.223
	1.0	.25	.3	0.207	0.310	0.263	0.035	0.174	0.259	0.219	0.028
			.7	0.192	0.268	0.227	0.024	0.167	0.223	0.191	0.019
		1.0	.3	0.853	1.305	1.065	0.171	0.687	1.034	0.854	0.134
			.7	0.758	1.767	1.082	0.287	0.600	1.394	0.857	0.225
48	.25	.25	.3	0.189	0.339	0.258	0.054	0.180	0.312	0.240	0.048
			.7	0.206	0.407	0.271	0.061	0.197	0.381	0.254	0.057
		1.0	.3	0.596	1.391	1.005	0.202	0.551	1.296	0.929	0.188
			.7	0.567	1.473	1.010	0.245	0.520	1.373	0.943	0.230
	1.0	.25	.3	0.196	0.314	0.230	0.039	0.180	0.300	0.216	0.038
			.7	0.174	0.350	0.258	0.056	0.151	0.301	0.211	0.046
		1.0	.3	0.765	1.290	1.031	0.197	0.596	0.995	0.800	0.148
			.7	0.906	1.283	1.028	0.123	0.674	0.967	0.797	0.099

Note. MTJM= Multilevel Testlet Joint Model, MJM= Multilevel Joint Model.

Table B. 6

MTJM and MJM Estimates of Variance of the Group-Specific Speed Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\tau_g}^2$				MJM $\sigma_{\tau_g}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.209	0.343	0.250	0.042	0.205	0.342	0.250	0.042
			.7	0.180	0.293	0.251	0.037	0.180	0.293	0.252	0.038
		1.0	.3	0.624	1.118	0.919	0.163	0.636	1.127	0.924	0.159
			.7	0.664	1.096	0.929	0.161	0.668	1.101	0.930	0.163
	1.0	.25	.3	0.220	0.353	0.274	0.047	0.217	0.361	0.275	0.050
			.7	0.162	0.287	0.237	0.035	0.163	0.291	0.237	0.036
		1.0	.3	0.742	1.358	1.048	0.196	0.728	1.362	1.047	0.201
			.7	0.716	1.420	1.023	0.234	0.729	1.434	1.024	0.233
48	.25	.25	.3	0.192	0.292	0.245	0.039	0.181	0.295	0.241	0.041
			.7	0.207	0.304	0.255	0.032	0.204	0.306	0.252	0.034
		1.0	.3	0.845	1.232	1.042	0.151	0.802	1.215	1.025	0.149
			.7	0.742	1.468	1.136	0.246	0.720	1.476	1.126	0.257
	1.0	.25	.3	0.186	0.310	0.241	0.041	0.169	0.305	0.234	0.041
			.7	0.183	0.297	0.242	0.043	0.186	0.288	0.235	0.038
		1.0	.3	0.586	1.214	0.971	0.197	0.577	1.205	0.942	0.215
			.7	0.727	1.232	0.899	0.172	0.712	1.154	0.875	0.159

Note. MTJM= Multilevel Testlet Joint Model, MJM= Multilevel Joint Model.

Table B. 7

MTJM and TJM Estimates of Variance of the Testlet 1 Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\gamma 1}^2$				TJM $\sigma_{\gamma 1}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.200	0.349	0.253	0.044	0.193	0.344	0.251	0.045
			.7	0.211	0.371	0.295	0.052	0.219	0.371	0.297	0.053
		1.0	.3	0.204	0.342	0.285	0.050	0.236	0.403	0.303	0.056
			.7	0.258	0.399	0.311	0.044	0.259	0.463	0.319	0.058
	1.0	.25	.3	0.798	1.153	0.952	0.115	0.790	1.208	0.964	0.123
			.7	0.860	1.143	0.997	0.099	0.838	1.154	0.994	0.110
		1.0	.3	0.870	1.146	0.981	0.093	0.928	1.323	1.086	0.122
			.7	0.891	1.185	1.049	0.083	0.922	1.399	1.138	0.139
48	.25	.25	.3	0.202	0.312	0.278	0.033	0.200	0.314	0.275	0.033
			.7	0.229	0.343	0.281	0.034	0.221	0.322	0.272	0.031
		1.0	.3	0.202	0.373	0.272	0.046	0.197	0.340	0.258	0.046
			.7	0.198	0.325	0.272	0.037	0.183	0.329	0.255	0.038
	1.0	.25	.3	0.807	1.147	0.976	0.108	0.801	1.141	0.969	0.108
			.7	0.853	1.178	0.991	0.101	0.835	1.185	0.979	0.110
		1.0	.3	0.899	1.123	1.014	0.086	0.869	1.098	0.985	0.076
			.7	0.879	1.120	0.982	0.066	0.849	1.102	0.944	0.077

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 8

MTJM and TJM Estimates of Variance of the Testlet 2 Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\gamma 2}^2$				TJM $\sigma_{\gamma 2}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.185	0.315	0.247	0.036	0.195	0.300	0.252	0.034
			.7	0.203	0.340	0.299	0.048	0.224	0.365	0.298	0.046
		1.0	.3	0.230	0.382	0.289	0.047	0.249	0.351	0.291	0.036
			.7	0.193	0.353	0.251	0.058	0.183	0.400	0.269	0.057
	1.0	.25	.3	0.919	1.140	1.049	0.076	0.920	1.168	1.058	0.079
			.7	0.851	1.302	1.073	0.135	0.853	1.337	1.085	0.138
		1.0	.3	0.829	1.160	1.032	0.099	0.919	1.255	1.098	0.106
			.7	0.883	1.163	1.040	0.100	0.963	1.308	1.127	0.099
48	.25	.25	.3	0.253	0.371	0.295	0.040	0.248	0.358	0.289	0.037
			.7	0.229	0.326	0.276	0.026	0.227	0.319	0.274	0.026
		1.0	.3	0.222	0.375	0.286	0.048	0.203	0.362	0.261	0.049
			.7	0.223	0.342	0.285	0.035	0.216	0.325	0.265	0.032
	1.0	.25	.3	0.808	1.146	0.970	0.110	0.809	1.144	0.967	0.111
			.7	0.883	1.181	1.059	0.102	0.917	1.195	1.057	0.100
		1.0	.3	0.880	1.181	0.993	0.099	0.841	1.119	0.965	0.085
			.7	0.726	1.132	0.954	0.111	0.739	1.043	0.917	0.085

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 9

MTJM and TJM Estimates of Variance of the Testlet 3 Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\gamma 3}^2$				TJM $\sigma_{\gamma 3}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.208	0.347	0.280	0.040	0.201	0.347	0.282	0.042
			.7	0.204	0.297	0.264	0.027	0.199	0.300	0.267	0.032
		1.0	.3	0.203	0.310	0.256	0.037	0.203	0.336	0.265	0.045
			.7	0.226	0.311	0.278	0.031	0.214	0.343	0.295	0.042
	1.0	.25	.3	0.833	1.110	0.994	0.100	0.854	1.159	1.006	0.103
			.7	0.823	1.108	0.952	0.085	0.840	1.132	0.967	0.081
		1.0	.3	0.880	1.103	1.004	0.078	0.988	1.147	1.060	0.055
			.7	0.919	1.218	1.033	0.085	0.941	1.316	1.101	0.108
48	.25	.25	.3	0.188	0.345	0.270	0.064	0.182	0.338	0.262	0.061
			.7	0.211	0.342	0.270	0.040	0.200	0.327	0.269	0.040
		1.0	.3	0.212	0.408	0.286	0.071	0.188	0.396	0.272	0.073
			.7	0.194	0.315	0.245	0.040	0.181	0.313	0.235	0.038
	1.0	.25	.3	0.813	1.130	0.967	0.103	0.807	1.123	0.961	0.104
			.7	0.846	1.185	0.995	0.102	0.834	1.157	0.975	0.092
		1.0	.3	0.867	1.344	1.057	0.137	0.810	1.348	1.021	0.148
			.7	0.812	1.164	1.001	0.103	0.778	1.133	0.988	0.111

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 10

MTJM and TJM Estimates of Variance of the Testlet 4 Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\gamma 4}^2$				TJM $\sigma_{\gamma 4}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
		1.0	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
	1.0	.25	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
		1.0	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
48	.25	.25	.3	0.206	0.341	0.269	0.049	0.204	0.341	0.266	0.048
			.7	0.234	0.384	0.295	0.052	0.232	0.378	0.287	0.049
		1.0	.3	0.232	0.339	0.280	0.035	0.207	0.331	0.263	0.040
			.7	0.219	0.325	0.282	0.032	0.197	0.302	0.260	0.031
	1.0	.25	.3	0.828	1.178	0.978	0.116	0.834	1.179	0.975	0.116
			.7	0.821	1.108	0.977	0.074	0.813	1.109	0.981	0.078
		1.0	.3	0.768	1.148	0.986	0.104	0.745	1.146	0.962	0.105
			.7	0.874	1.124	1.028	0.080	0.827	1.083	0.981	0.097

Note. Testlets not present for the shorter test length condition are represented with "-".
MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 11

MTJM and TJM Estimates of Variance of the Testlet 5 Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\gamma 5}^2$				TJM $\sigma_{\gamma 5}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
		1.0	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
	1.0	.25	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
		1.0	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
48	.25	.25	.3	0.221	0.343	0.281	0.037	0.212	0.348	0.277	0.038
			.7	0.204	0.345	0.274	0.044	0.206	0.342	0.268	0.048
		1.0	.3	0.239	0.388	0.298	0.046	0.221	0.364	0.278	0.048
			.7	0.235	0.319	0.279	0.029	0.227	0.298	0.264	0.025
	1.0	.25	.3	0.836	1.135	0.987	0.104	0.834	1.132	0.981	0.101
			.7	0.806	1.192	0.981	0.103	0.769	1.143	0.965	0.104
		1.0	.3	0.710	1.191	0.990	0.151	0.725	1.147	0.955	0.133
			.7	0.817	1.132	1.000	0.113	0.799	1.203	0.981	0.128

Note. Testlets not present for the shorter test length condition are represented with "-".
MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 12

MTJM and TJM Estimates of Variance of the Testlet 6 Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\sigma_{\gamma 6}^2$				TJM $\sigma_{\gamma 6}^2$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
		1.0	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
	1.0	.25	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
		1.0	.3	-	-	-	-	-	-	-	-
			.7	-	-	-	-	-	-	-	-
48	.25	.25	.3	0.205	0.315	0.280	0.033	0.199	0.309	0.275	0.032
			.7	0.226	0.310	0.278	0.027	0.214	0.312	0.272	0.028
		1.0	.3	0.213	0.309	0.254	0.034	0.198	0.303	0.237	0.034
			.7	0.226	0.377	0.291	0.050	0.192	0.368	0.266	0.055
	1.0	.25	.3	0.823	1.130	0.981	0.106	0.820	1.133	0.979	0.105
			.7	0.839	1.223	1.006	0.109	0.818	1.178	1.001	0.099
		1.0	.3	0.868	1.111	0.990	0.090	0.778	1.058	0.950	0.095
			.7	0.929	1.156	1.065	0.076	0.888	1.167	1.038	0.082

Note. Testlets not present for the shorter test length condition are represented with "-".
MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 13

MTJM and TJM Estimates of Variance of the Item Difficulty Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM σ_b^2				TJM σ_b^2			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	1.017	1.122	1.077	0.035	0.998	1.098	1.051	0.032
			.7	1.038	1.106	1.070	0.022	1.012	1.082	1.044	0.025
		1.0	.3	1.008	1.169	1.086	0.043	0.927	1.076	1.005	0.040
			.7	1.031	1.170	1.081	0.038	0.938	1.137	0.988	0.061
	1.0	.25	.3	1.015	1.099	1.057	0.027	1.000	1.082	1.039	0.027
			.7	1.025	1.079	1.053	0.019	1.014	1.060	1.036	0.017
		1.0	.3	0.968	1.192	1.075	0.070	0.879	1.152	1.008	0.087
			.7	0.985	1.178	1.086	0.051	0.910	1.106	1.015	0.057
48	.25	.25	.3	0.968	1.057	1.026	0.031	0.959	1.047	1.008	0.029
			.7	0.976	1.074	1.033	0.031	0.959	1.047	1.014	0.028
		1.0	.3	0.987	1.086	1.051	0.028	0.950	1.058	0.998	0.035
			.7	1.017	1.114	1.062	0.032	0.955	1.054	0.993	0.031
	1.0	.25	.3	0.991	1.071	1.036	0.024	0.971	1.052	1.022	0.024
			.7	0.986	1.057	1.014	0.024	0.966	1.040	1.000	0.023
		1.0	.3	0.999	1.118	1.060	0.039	0.926	1.078	1.007	0.048
			.7	1.000	1.067	1.035	0.024	0.928	1.007	0.972	0.029

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 14

MTJM and TJM Estimates of Variance of the Item Difficulty Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MJM σ_b^2				HM σ_b^2			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.956	1.055	1.007	0.034	0.934	1.034	0.980	0.030
			.7	0.967	1.025	0.997	0.018	0.948	1.000	0.972	0.016
		1.0	.3	0.936	1.100	1.015	0.042	0.859	0.997	0.935	0.037
			.7	0.977	1.075	1.009	0.028	0.878	1.028	0.917	0.048
	1.0	.25	.3	0.802	0.881	0.841	0.025	0.790	0.861	0.822	0.022
			.7	0.806	0.871	0.836	0.020	0.796	0.848	0.820	0.019
		1.0	.3	0.774	0.928	0.854	0.047	0.689	0.896	0.787	0.063
			.7	0.781	0.932	0.855	0.040	0.710	0.874	0.787	0.048
48	.25	.25	.3	0.900	0.975	0.944	0.026	0.881	0.959	0.928	0.026
			.7	0.903	0.976	0.949	0.024	0.895	0.962	0.935	0.022
		1.0	.3	0.916	1.008	0.968	0.026	0.879	0.973	0.923	0.033
			.7	0.937	1.030	0.979	0.028	0.892	0.968	0.920	0.024
	1.0	.25	.3	0.906	0.987	0.954	0.024	0.892	0.970	0.940	0.024
			.7	0.745	0.803	0.773	0.019	0.739	0.790	0.761	0.018
		1.0	.3	0.775	0.846	0.813	0.025	0.725	0.820	0.778	0.033
			.7	0.772	0.807	0.791	0.012	0.725	0.771	0.751	0.018

Note. MJM= Multilevel Joint Model , TJM= Testlet Joint Model.

Table B. 15

MTJM and TJM Estimates of Variance of the Item Intensity Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM σ_β^2				TJM σ_β^2			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	1.039	1.081	1.055	0.015	1.034	1.078	1.054	0.014
			.7	1.032	1.077	1.054	0.015	1.032	1.078	1.053	0.016
		1.0	.3	1.055	1.086	1.070	0.010	1.045	1.086	1.062	0.012
			.7	1.032	1.088	1.059	0.015	1.025	1.086	1.054	0.017
	1.0	.25	.3	1.035	1.083	1.055	0.015	1.035	1.076	1.052	0.015
			.7	1.034	1.077	1.056	0.012	1.032	1.066	1.052	0.010
		1.0	.3	1.035	1.077	1.055	0.013	1.037	1.071	1.049	0.010
			.7	1.053	1.093	1.065	0.011	1.044	1.101	1.065	0.015
48	.25	.25	.3	1.018	1.049	1.029	0.008	1.018	1.051	1.028	0.010
			.7	1.010	1.048	1.030	0.011	1.002	1.039	1.028	0.012
		1.0	.3	1.017	1.064	1.036	0.013	1.021	1.080	1.039	0.019
			.7	1.031	1.072	1.040	0.012	1.023	1.206	1.061	0.054
	1.0	.25	.3	1.010	1.049	1.029	0.013	1.014	1.050	1.028	0.013
			.7	1.011	1.044	1.029	0.010	1.009	1.050	1.028	0.011
		1.0	.3	1.019	1.054	1.039	0.011	1.010	1.093	1.043	0.028
			.7	1.014	1.048	1.035	0.011	1.015	1.060	1.038	0.016

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 16

MJM and HM Estimates of Variance of the Item Intensity Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MJM σ_β^2				HM σ_β^2			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	1.035	1.076	1.056	0.015	1.034	1.072	1.052	0.014
			.7	1.028	1.076	1.055	0.017	1.029	1.077	1.053	0.016
		1.0	.3	1.049	1.093	1.071	0.013	1.043	1.090	1.062	0.015
			.7	1.031	1.088	1.059	0.016	1.020	1.091	1.053	0.021
	1.0	.25	.3	1.041	1.072	1.055	0.012	1.034	1.075	1.052	0.014
			.7	1.042	1.077	1.057	0.009	1.032	1.068	1.052	0.010
		1.0	.3	1.042	1.072	1.055	0.010	1.037	1.064	1.048	0.010
			.7	1.043	1.087	1.064	0.013	1.040	1.103	1.063	0.017
48	.25	.25	.3	1.017	1.055	1.030	0.010	1.016	1.054	1.028	0.011
			.7	1.011	1.040	1.029	0.009	1.005	1.039	1.027	0.010
		1.0	.3	1.023	1.049	1.034	0.009	1.014	1.085	1.039	0.021
			.7	1.031	1.115	1.047	0.024	1.019	1.195	1.061	0.051
	1.0	.25	.3	1.017	1.043	1.029	0.011	1.013	1.052	1.029	0.013
			.7	1.013	1.050	1.030	0.011	1.009	1.050	1.026	0.012
		1.0	.3	1.018	1.058	1.039	0.011	1.010	1.094	1.044	0.026
			.7	1.016	1.048	1.035	0.010	1.010	1.062	1.037	0.015

Note. MJM= Multilevel Joint Model, HM= Hierarchical Model.

Table B. 17

MTJM and TJM Derived Estimates of Correlation of the Item Difficulty and Item Intensity Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MTJM $\rho_{b\beta}$				TJM $\rho_{b\beta}$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.262	0.308	0.287	0.012	0.264	0.310	0.287	0.012
			.7	0.236	0.315	0.281	0.023	0.240	0.313	0.282	0.021
		1.0	.3	0.264	0.297	0.280	0.010	0.255	0.304	0.278	0.016
			.7	0.277	0.312	0.290	0.012	0.266	0.309	0.289	0.015
	1.0	.25	.3	0.271	0.312	0.290	0.013	0.274	0.312	0.292	0.013
			.7	0.257	0.305	0.286	0.014	0.256	0.309	0.286	0.016
		1.0	.3	0.263	0.310	0.288	0.016	0.244	0.328	0.283	0.025
			.7	0.243	0.302	0.282	0.018	0.201	0.310	0.281	0.032
48	.25	.25	.3	0.275	0.310	0.293	0.011	0.275	0.311	0.293	0.012
			.7	0.274	0.313	0.291	0.012	0.270	0.314	0.290	0.012
		1.0	.3	0.276	0.320	0.292	0.012	0.264	0.341	0.294	0.022
			.7	0.273	0.309	0.291	0.012	0.218	0.304	0.280	0.025
	1.0	.25	.3	0.269	0.308	0.288	0.012	0.270	0.307	0.288	0.012
			.7	0.268	0.314	0.290	0.014	0.270	0.308	0.291	0.014
		1.0	.3	0.269	0.321	0.296	0.015	0.244	0.346	0.296	0.027
			.7	0.278	0.307	0.291	0.011	0.255	0.308	0.286	0.016

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model.

Table B. 18

MJM and HM Derived Estimates of Correlation of the Item Difficulty and Item Intensity Parameter

I	σ_γ^2	σ_g^2	$\rho_{\theta\tau}$	MJM $\rho_{b\beta}$				HM $\rho_{b\beta}$			
				Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
24	.25	.25	.3	0.268	0.306	0.287	0.009	0.263	0.307	0.286	0.011
			.7	0.242	0.309	0.281	0.020	0.241	0.318	0.281	0.021
		1.0	.3	0.260	0.292	0.279	0.009	0.257	0.311	0.279	0.017
			.7	0.273	0.313	0.291	0.014	0.266	0.307	0.288	0.015
	1.0	.25	.3	0.273	0.306	0.289	0.012	0.274	0.307	0.289	0.011
			.7	0.259	0.307	0.284	0.015	0.254	0.306	0.284	0.016
		1.0	.3	0.264	0.311	0.287	0.016	0.247	0.319	0.278	0.023
			.7	0.248	0.296	0.281	0.017	0.201	0.305	0.277	0.030
48	.25	.25	.3	0.273	0.312	0.292	0.012	0.273	0.309	0.292	0.012
			.7	0.277	0.314	0.291	0.012	0.275	0.313	0.291	0.011
		1.0	.3	0.278	0.312	0.292	0.009	0.256	0.344	0.293	0.024
			.7	0.266	0.310	0.290	0.013	0.223	0.304	0.279	0.023
	1.0	.25	.3	0.266	0.302	0.287	0.011	0.267	0.308	0.288	0.013
			.7	0.268	0.311	0.290	0.014	0.268	0.308	0.290	0.013
		1.0	.3	0.277	0.317	0.295	0.012	0.246	0.342	0.294	0.025
			.7	0.275	0.307	0.288	0.011	0.258	0.309	0.284	0.016

Note. MJM= Multilevel Joint Model, HM= Hierarchical Model.

Appendix C

Table C. 1

Descriptive Statistics for Parameter Estimates by Model

Parameter	Model	Min.	Max.	Mean	SD
θ_i	MTJM	-2.365	2.212	0.002	0.802
	TJM	-2.162	2.230	0.002	0.889
	MJM	-2.352	2.201	0.000	0.800
	HM	-2.159	2.233	0.000	0.887
τ_i	MTJM	-0.242	0.480	0.000	0.100
	TJM	-0.285	0.703	0.000	0.126
	MJM	-0.242	0.481	0.000	0.100
	HM	-0.299	0.698	0.000	0.126
b	MTJM	-1.498	1.336	-0.058	0.832
	TJM	-1.482	1.286	-0.077	0.812
	MJM	-1.496	1.314	-0.055	0.811
	HM	-1.476	1.277	-0.068	0.794
β_i	MTJM	3.424	5.042	4.184	0.397
	TJM	3.416	5.038	4.178	0.398
	MJM	3.422	5.042	4.183	0.397
	HM	3.418	5.038	4.179	0.398

Note. MTJM= Multilevel Testlet Joint Model, TJM= Testlet Joint Model, MJM= Multilevel Joint Model, HM= Hierarchical Model.

Appendix D

Table D. 1

The Univariate ANOVA Results of the RMSE of the Person Ability Estimates

Source	RMSE of θ_j		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
test_length	6008.03	<0.001	.029
testlet_var	2496.93	<0.001	.012

Note. test_length=Number of test items (I); testlet_var=Testlet variance magnitude (σ_γ^2).

Table D. 2

The Univariate ANOVA Results of the RMSE of the Person Speed Estimates

Source	RMSE of τ_j		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
test_length	8678.32	<0.001	.042

Note. test_length=Number of test items (I).

Table D. 3

The Univariate ANOVA Results of the Bias of the Item Difficulty Estimates (I=24)

Source	Bias of b_i		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
testlet_var	25.33	<0.001	.013
group_var	162.20	<0.001	.078

Note. group_var=Group variance magnitude (σ_g^2); testlet_var=Testlet variance magnitude (σ_γ^2).

Table D. 4

The Univariate ANOVA Results of the RMSE of the Item Difficulty Estimates (I=24)

Source	Bias of b_i		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
group_var	166.22	<0.001	.080

Note. group_var=Group variance magnitude (σ_g^2).

Table D. 5

The Univariate ANOVA Results of the RMSE of the Item Difficulty Estimates (I=48)

Source	RMSE of b_i		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
theta_tau_corr	37.08	<0.001	.010
group_var*theta_tau_corr	71.11	<0.001	.018

Note. group_var=Group variance magnitude (σ_g^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

Table D. 6

The Univariate ANOVA Results of the Bias of the Item Time Intensity Estimates (I=24)

Source	Bias of β_i		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
testlet_var	26.00	<0.001	.013
theta_tau_corr	25.75	<0.001	.013

Note. testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

Table D. 7

The Univariate ANOVA Results of the RMSE of the Item Time Intensity Estimates (I=24)

Source	RMSE of β_i		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
group_var	97.43	<0.001	.048

Note. group_var=Group variance magnitude (σ_g^2).

Table D. 8

The Univariate ANOVA Results of the Bias of the Item Time Intensity Estimates (I=48)

Source	Bias of β_i		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
group_var	65.83	<0.001	.017
testlet_var*theta_tau_corr	386.34	<0.001	.092

Note. group_var=Group variance magnitude (σ_g^2); testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

Table D. 9

The Univariate ANOVA Results of the RMSE of the Item Time Intensity Estimates (I=48)

Source	RMSE of β_i		
	F Statistics	*p*-value	Partial η^2
Between-Subject Effects			
testlet_var	47.05	<0.001	.012
group_var	132.05	<0.001	.033
group_var*theta_tau_corr	46.72	<0.001	.012

Note. group_var=Group variance magnitude (σ_g^2); testlet_var=Testlet variance magnitude (σ_γ^2); theta_tau_corr=Correlation between person ability and speed ($\rho_{\theta\tau}$).

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, *21*(1), 1-23.

Aslett, L. J. M. (2019). *RStudio AMIs for Amazon EC2 cloud computing*. AMI ID ami-0226a8af83fcecb43, http://www.louisaslett.com/RStudio_AMI/.

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.

Binici, S. (2007). Random-effects differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods. (unpublished doctoral dissertation, Florida State University)

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126-1148.

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., & 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.

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.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

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.

Cohen, A. (1965). Estimates of linear combinations of the parameters in the mean vector of a multivariate distribution. *The Annals of Mathematical Statistics*, *36*(1), 78-87.

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

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of educational measurement*, *43*(2), 145-168.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104-121.

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.

Fox, J., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 269–286.

Fox, J. P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement, 15*(3-4), 261-28.

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

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.

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models*. New York, NY: Cambridge University Press.

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.

Glas, C. A., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In *Computerized adaptive testing: Theory and practice* (pp. 271-287). Springer, Dordrecht.

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.

Goldstein, H. (1987). *Multilevel models in education and social research.* Oxford University Press.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London, England: Arnold.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological), 29*(1), 83–100.Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Huang, H. Y., & Wang, W. C. (2014). Multilevel higher-order item response theory models. *Educational and Psychological Measurement, 74*(3), 495-515.

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.

Im, S. K. (2015). *The Hierarchical Testlet Response Time Model: Bayesian analysis of a testlet model for item responses and response times* (Doctoral dissertation, University of Kansas).

IBM Corp. (2019). *IBM SPSS Statistics for Windows, Version 26.0*. Armonk, NY: IBM Corp.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2010). *Simultaneous modeling of item and person dependence using a multilevel Rasch measurement model*. Paper presented at the meeting of the American Educational Research Association, Denver, CO.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). *A multilevel testlet model for dual local dependence. Journal of Educational Measurement, 49*(1), 82-1.

Jiao, H., Kamata, A., & Xie, C. (2015). A multilevel cross-classified testlet model for complex item and person clustering in item response modeling. In J. Harring, L. Stapleton, & S. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*. Charlotte, NC: Information Age Publishing.

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement*, *50*(2), 186-203.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of applied measurement*, *6*(3), 311.

Jiao, H., & Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology*, *68*(1), 65-83.

Kamata, A. (1999). *Some generalizations of the Rasch Model: An application of the hierarchical generalized linear model.* (unpublished doctoral dissertation, Michigan State University, East Lansing).

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

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.

Klein Entink, R., 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.

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), 621-64.

Klotzke, K., & Fox, J. P. (2019). Bayesian covariance structure modelling of responses and process data. *Frontiers in psychology*, *10*, 1675.

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.

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359-379.

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.

Liao, D. (2018). *Modeling the Speed-Accuracy-Difficulty Interaction in Joint Modeling of Responses and Response Time* (unpublished doctoral dissertation, University of Maryland).

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* (AddisonWesley

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.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26(*3), 307–33.

Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, 0146621618824853.

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

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Pacific Grove, CA: Brooks.

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., Tuerlinckx, F., & van der Maas, H. L. (2015). 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.

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.

Murphy, D. L., Dodd, B. G., & Vaughn, B. K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, *34*(6), 424-437.

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

OECD (2017), PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, PISA, OECD Publishing, Paris, https://doi.org/1.1787/9789264281820-en.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23-32.

Plummer, M (2017). *JAGS Version 4.3.0 User Manual.* Lyon, France. URL http://sourceforge.net/projects/mcmc-jags/.

R Core Team (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL https://www.Rproject.org/.

Ra, J. (2011). *Sensitivity of prior specification within testlet model* (unpublished doctoral dissertation, University of Georgia).

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Ranger, J., & Kuhn, J. T. (2014). 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., & Ortner, T. (2012). 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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, 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.

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.

Rubin, D. B. (1981). *Estimation in parallel randomized experiments. Journal of Educational Statistics, 6*(4), 377–401.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 1151–1172.

Rupp, A.A., Gushta, M., Mislevy, R.J., & Shaffer, D.W. (2010). Evidence-centered Design of Epistemic Games: Measurement Principles for Complex Learning Environments. Journal of Technology, Learning, and Assessment, 8(4). Retrieved [date] from http:// www.jtla.org.

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.

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 .5–7. *Available: CRAN. R-project. org/package= R2jags. (September 2015)*.

Suh, H. (2010). *A study of Bayesian estimation and comparison of response time models in item response theory* (unpublished doctoral dissertation, University of Kansas).

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.

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika*, *2*(4), 249-254.

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.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, (2013). *National Assessment of Educational Progress (NAEP), Assessments.*

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.

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 Maas, H. L. J., 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.

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*.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In *Computerized adaptive testing: Theory and practice* (pp. 245-269). Springer, Dordrecht.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185-201.

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

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.

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.

Zhan, P., Jiao, H., & Liao, D. (2017). A longitudinal diagnostic classification model. arXiv preprint arXiv:1709.03431.

Zhan, P., Liao, M. & Bian, Y. (2018) Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology, 9,* 607.