Abstract

Title of Dissertation:	JOINT MODELING OF RESPONSES, RESPONSE TIME, AND ANSWER CHANGES IN TESTLET-BASED ASSESSMENT FOR COGNITIVE DIAGNOSIS
	Chengbin Yin, Doctor of Philosophy, 2022
Dissertation directed by:	Dr. Hong Jiao, Professor, Measurement, Statistics and Evaluation, Department of Human Development and Quantitative Methodology

To address the scenario of testlet-based assessment, this dissertation proposes a joint model of responses, response time, and answer change patterns for testlet-based cognitive diagnostic assessments. A simulation study was conducted to assess the impact of accounting for dual item and item time dependency and of incorporating answer change patterns as an additional data source on model fit, classification accuracy at the attribute and attribute profile level, and parameter estimation. Through manipulating three factors, the simulation study examined the extent to which the manipulated factors impacted the performance of the proposed model and two comparison models in recovering model parameters. Application of the proposed model was demonstrated with an empirical dataset.

JOINT MODELING OF RESPONSES, RESPONSE TIME, AND ANSWER CHANGES IN TESTLET-BASED ASSESSMENT FOR COGNITIVE DIAGNOSIS

by

Chengbin Yin

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
2022

Advisory Committee:
Professor Hong Jiao, Chair
Professor Jeffrey R. Harring
Professor Robert Lissitz
Professor Yang Liu
Professor Robert J. Mislevy
Professor Yan Li, Dean's Representative

Dedication

To my family.

Acknowledgements

It is with great admiration that I write this to acknowledge the contributions of the members of my advisory committee to my dissertation study, members who hail from a community where I once belonged in the land that has my roots and where I still wish I could belong.

In this journey toward reaching new heights and achieving new dreams, I have had the fortune to study under the guidance of Dr. Hong Jiao, nourished by her exceptional mentorship and meticulous guidance, without which this dissertation would have been an impossible mission. Even before that it was Dr. Robert Mislevy, who opened my eyes to a field full of wonders and opportunities to learn, and who, notwithstanding his departure from the campus of the Terps, guides me with his wisdom and lends me a helping hand whenever he can. My heartfelt gratitude goes to Dr. Robert Lissitz for the role model he has been for all aspiring psychometricians. Just as inspirational are Dr. Jeffrey Harring's talks about the Newton—Raphson method and Maximum Likelihood Estimation. It seems only yesterday that I sat in his class and endeavored to grasp the algorithms and MCMC. To him I owe my knowledge of the mathematical foundations for simulation techniques. To Dr. Yang Liu, whose exemplary research is for me to follow, and to Dr. Yan Li, an accomplished scholar and researcher, I am forever grateful.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACCR: Attribute Correct Classification Rate

ACs: Answer Changes

CDM: Cognitive Diagnostic Model

JAD: JRT-DINA-R/RT/AC Model

JAD-TT: JRT-AC-DINA for Testlet Model

JD-TT: Joint Testlet-DINA Model

PCCR: Pattern Correct Classification Rate

RT: Response Time

Chapter 1: Introduction

The past decade has seen growing attention in designing and developing cognitive assessments that tap into and are informed by cognitive processes that students engage in during learning or test-taking (Ercikan & Pellegrino, 2017). The advent of computer technology and availability of rich data have concurrently brought forth opportunities for a range of approaches to modeling response and response process data that hold promise for building a holistic understanding of the processes that test-takers engage in and strategies they use as they interact with the assessment tasks and test-taking environment. Cognitive diagnostic models (CDMs) configure assessment tasks as functions of component skills and relate their features to skills required for performing them through a Q-matrix (Leighton & Gierl, 2007; Mislevy, 2018; Nichols, Chipman, & Brennan, 1995; Rupp, Templin, & Henson, 2010). Latest advances in cognitive diagnostic modeling have seen the integration of the response process data in the modeling of responses for cognitive diagnosis (Jiao, Liao, & Zhan, 2019; Zhan, Jiao, Liao, 2018a). This joint modeling approach, however, has largely been limited to response time (RT) data only, although other relevant data sources can potentially inform ability estimation and diagnosis of students' mastery status on the skills and attributes of interest. One recent example of integrating response process data other than RT is the integration of answer changes (ACs) data in cognitive diagnosis modeling, an approach validated by an empirical data analysis (Jiao, Ding, & Yin, 2020). The model, however, does not address local dependencies incurred by testlets. Testlets are widely used units of test construction in educational assessments for assessing skills and competencies across domains.

Motivated by these developments, this dissertation research proposes to expand on current models of responses and response time (RT) by incorporating answer changes (ACs) as an additional supplemental data for testlet-based cognitive diagnostic assessments.

1.1 Statement of the Problem

CDMs are process models for assessment tasks that closely resemble the knowledge, patterns, and rules people use when responding to the tasks (Kane & Mislevy, 2017). CDMs structure assessment tasks around selected attributes or skill sets sampled from a narrowly-defined domain of interest and model item responses as functions of an assembly of required skills, based upon which process-model interpretations and inferences are drawn regarding what examinees can or cannot do and are often at a finer grain size compared to summary statements. As psychometric models, specific CDMs are multivariate discrete latent variable models which assume a theory of cognitive processes for the responses to the items. Examples include the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977), the deterministic noisy "or" gate (DINO; Templin & Henson, 2006; Templin, 2016) and the linear logistic model (LLM; Hagenaars, 1993). CDMs are widely used in learning and assessment applications, particularly in formative assessments to understand students' mastery or nonmastery of core skills and strategies of interest (e.g., Deonovic, Chopade, Yudelson, de la Torre, & van Davier, 2019; Leighton & Gierl, 2007; Rupp et al., 2010).

Advances in cognitive diagnostic modeling have seen a joint modeling approach integrating response and response process data from computer-based

assessments of student learning (Jiao et al., 2019; Zhan et al., 2018a). This approach extends van der Linden's (2007) hierarchical framework for modeling speed and accuracy to incorporate RTs in the modeling of response data for cognitive diagnosis. Compared to stand-alone CDMs such as the DINA model, the joint model of item responses and RTs improves attribute level and attribute profile level classification accuracy and yields more accurate and precise estimates of model parameters (Zhan et al., 2018a).

Response processes are the thought process, strategies, and behaviors of the examinees as they interact with assessment tasks (Ercikan & Pellegrino, 2017). In computer-based assessments, data on examinee response processes can be collected through response logs which document examinee interaction with stimulus materials and indicate which task elements are used and manipulated. Examples of response process data include RT (van der Linden, 2006, 2009), answer changes (ACs; Jeon, De Boeck, & van der Linden, 2017; Liu, Bridgeman, Gu, Xu, & Kong, 2015; Sinharay & Johnson, 2016; van der Linden & Jeon, 2012), eye-tracking (Cho, Brown-Schmidt, Naveiras, & De Boeck, 2020; Oranje, Gorin, Jia, & Kerr, 2017; van Gog & Sheiter, 2010), and hint requests (Bolsinova, Deonovic, Attali, & Maris, 2020). RT as a widely used response process data records the lengths of time an examinee spends on an item or particular aspects of the item. Eye-tracking data, as another example, include traces of performance and thinking and points of gaze on the computer screen (Man & Harring, 2019). ACs are another type of response process data, the patterns, and outcomes of which are closely associated with test-takers ability level (Liu, Bridgeman, Gu, Xu, & Kong, 2015). As such, response process data can provide

information about examinees' engagement with assessment tasks and the extent to which examinees utilize resources and information related to test items.

With the use of computer technology and digital devices becoming ubiquitous, rich response process data become readily available. The modeling of response process data is emerging and can be integrated with standard item response models. To date, however, response modeling incorporating process data other than RT is limited. The joint modeling approach proposed by Zhan et al. (2018a) and Jiao et al. (2019) illustrates the use of RT to improve parameter estimation and classification accuracy in CDMs. As discussed by Jiao et al. (2019), the effects of integrating response process data in the modeling of responses for cognitive diagnosis can be evident when a test is not ideally designed for cognitive diagnosis and can potentially improve model parameter estimation. A recent development in the joint modeling of response and response process data is the integration of ACs as an additional process data in the modeling of response and RT for cognitive diagnosis, an approach which is validated by analyses of empirical data (Jiao, Ding, & Yin, 2020). The joint model of response, RT, and ACs proposed in this study, however, does not address local dependencies incurred by testlets. Testlet-based assessments are widely used in educational assessments for assessing skills and competencies at various levels and across domains. Studies have shown that models incorrectly assuming local item independence can result in biased parameter estimates (Yen, 1993). Considering this, this dissertation research proposes to extend the joint modeling of response and response process data for cognitive diagnosis by integrating

both RT and ACs as process data to specifically account for the testlet design and examine their effects on model fit and parameter estimation.

1.2 Purpose of the Study

This dissertation study proposes an extension of the joint model of response and RT for cognitive diagnosis (Jiao et al., 2019; Zhan et al. 2018a) by incorporating another type of response process data, ACs, as an additional data source in testlet-based cognitive diagnostic assessments. Its primary purpose is to identify whether the inclusion of ACs can improve classification accuracy at the attribute and attribute profile level, model fit, and parameter estimation compared to the joint model of response and RT only.

ACs, also called erasures or response revisions, refer to the fact that examinees, after making an initial decision, subsequently revisit the decision and revert to an alternative option as their best choice (Jeon, Deboeck, & van der Linden, 2017). Patterns and outcomes of ACs are associated with examinee ability, with high-performing examinees making more wrong to right changes and attaining greater score gains from making them compared to low-performing examinees (e.g., Jeon et al., 2017, Liu et al., 2015; Milia, 2007). ACs directly contribute to changes in response patterns. The inclusion of ACs as an additional response process data for cognitive diagnostic assessments thus may result in improved estimation of examinee ability and diagnosis of skills and strategies specified in the assessment blueprint.

This dissertation study additionally examines the effects of ignoring dependency of response and RT on model performance and parameter estimation in testlet-based cognitive diagnostic assessments. Testlets are units of test construction

that aggregate a group of test items around a common stimulus (Wainer & Kiely, 1987). The use of testlets in educational assessments is a common practice and is likely to induce local response dependence, with responses to groups of items embedded within the same context. In joint models of response and RT, testlet design can additionally induce dependency among RT, resulting in local RT dependence. As fitting standard IRT models to testlet responses can produce overestimation of measurement precision, and biased estimation of item parameters, this dissertation research proposes to investigate how the exclusion of testlet effects in joint models of responses, RTs, and ACs in testlet-based cognitive diagnostic assessments affects the performance of the proposed joint models in comparison with alternative comparison models and the estimation of model parameters.

This research conducts a simulation study to investigate the effects of including ACs in addition to RTs and of accounting for local response and RT dependence on model performance and parameter estimates in the joint model of responses, RTs, and ACs in testlet-based cognitive diagnostic assessments. It manipulates three factors identified as likely contributing to systematic variation in parameter estimates: sample size, correlation between speed and ability, and testlet effect size. These factors simulate the conditions in the real world educational tests that likely vary by the magnitude of the testlet effects and by the time constraints identified as likely contributing to different levels of correlation between speed and ability (van der Linden, 2009).

The simulation study in this dissertation research is guided by the following research questions:

1. How does the proposed joint model of response, RT, and ACs for testlet-based cognitive diagnostic assessment perform compared to testlet-based joint model of response and RT in terms of model fit, attribute and attribute profile classification accuracy, and parameter estimates?
2. How do the factors manipulated in this study, i.e., correlation between speed and ability, testlet effects size, and sample size, affect comparisons of the joint model of response, RT, and ACs for testlet-based cognitive diagnostic assessment and testlet-based joint model of response and RT?
3. How does the proposed joint model of response, RT, and ACs accounting for testlet effects perform compared to the alternative model ignoring these effects in terms of model fit, attribute and attribute profile classification accuracy, and parameter estimation?
4. How do the factors manipulated in this study, i.e., the correlation between speed and ability, the magnitude of the testlet effects, and the sample size, affect comparisons of the joint model of response, RT, and ACs accounting for testlet effects and the joint model of response, RT, and ACs ignoring testlet effects?

To evaluate and validate the proposed joint model, this research conducts an analysis of an empirical dataset consisting of the response, RT, and ACs of 71 examinees for a standardized large-scale mathematics assessment. The empirical data analyses are guided by research questions 3 to identify the best-fitting model, based upon which parameter estimates, mixing proportions, and attribute patterns are summarized.

1.3 Significance of the Study

As noted earlier in this chapter, capacities for response and response process data collection have significantly expanded as computer technology becomes increasingly integrated into learning and assessment practices, and computer-based assessments have become the norm rather than the exception. Through the use of response logs and other data collection techniques, rich process data such as RT, eye-tracking data, ACs, and examinees' use of help features become readily available, promoting considerations of new approaches to modeling and integrating different types of process data. Response modeling incorporating process data other than RT is emerging and yet is very limited. Methodologically, a modeling approach integrating response data and different types of process data for cognitive diagnosis can generate improved parameter estimates and diagnosis of examinees' mastery status on the skills and strategies assessed by the educational tests compared to existing models that only consider response and RT. To this end, this dissertation study as an extension of the joint model of responses and RT by Jiao et al. (2019) and Zhan et al. (2018a) to incorporate a second type of process data, ACs, for cognitive diagnosis will likely serve as a modeling approach to incorporating multiple response process data in the modeling of responses for cognitive diagnosis, with possible extensions to accommodate other types of response process data and psychometric models other than CDMs.

Moreover, the proposed joint model distinguishes from existing models by explicitly accounting for testlet effects likely resulting from groups of items nested within the same content area and sharing the same stimulus and examines the effects

of fitting standard joint models of response and response process data on model performance and parameter estimation in testlet-based cognitive diagnostic assessments. van der Linden's (2007) hierarchical modeling framework assumes independence of response given ability and of RT given speed, and independence between response and RTs given examinees' speed and ability. In reality, the assumption of local independence of response and RT are unlikely to be tenable given the prevalent use of testlets as units of test construction in educational assessments. As such, by explicitly accounting for testlets effects in testlet-based cognitive diagnostic assessments and examining their effects under simulated conditions that resemble real world scenarios, the proposed model will likely demonstrate how the inclusion of testlet parameters will yield improved model fit, parameter estimates, and classification accuracy at the attribute and attribute profile level.

Finally, as stated in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), "validity refers to the degree to which evidence and theory support the interpretations of test scores for the proposed use of tests. The process of validation involves accumulating relevant evidence to propose a sound scientific basis for proposed score interpretations" (p.11, as cited in Kane & Mislevy, 2017). Advances in the use of computer technology for educational assessments expand evidence accumulation as going beyond traditional summary scores indicating overall achievement in a specific domain to encompass evidence garnered from supplemental data such as RT and log files. Fine-grained inferences regarding examinees' use of component skills and strategies as enabled by

CDMs modeling tasks as functions of component skills and abilities can be supplemented with additional response process data to support both trait interpretations and process-model interpretations of the meaning of test scores (Kane & Mislevy, 2017). In this regard, integrating additional process data in the modeling of responses in testlet-based diagnostic assessment is a step forward in extending validation of score meaning with supplemental response process data.

1.4 Overview of the Chapters

This chapter describes recent advances in educational assessments and psychometric modeling that motivate this research study and states the purposes of this research and its significance. The following chapters are organized as follows. Chapter 2 is a comprehensive review of current approaches to cognitive diagnostic modeling and the modeling of testlet effects, RT, and ACs, and distinct frameworks for modeling speed and accuracy. This chapter additionally introduces the model estimation method to be used in this research: fundamentals of Bayesian inference, the Monta Carlo simulation method, and diagnostics for assessing model convergence. Chapter 3 proposes the joint model, describing its formulation and parameterization, and the design of a simulation study to investigate the impact of the manipulated factors on model performance and parameter estimates, and concludes with a description of an empirical study designed to evaluate and validate the proposed model. Results of the simulation study and empirical data analyses are presented in Chapters 4. This dissertation concludes with Chapter 5 discussing the results of the simulation study and empirical data analyses and addressing limitations of this research.

Chapter 2: Literature Review

This chapter reviews approaches and methods for cognitive diagnostic modeling, and for modeling of RT, testlets, and AC patterns, which lays the foundation for the proposed modeling approach in Chapter 3. The first three sections review cognitive diagnostic models, testlet models, and RT models, with emphasis on the DINA model, the G-DINA model, the testlet response model (Wainer, Bradlow, & Wang, 2007), and van der Linden's (2006) lognormal RT model. The fourth section reviews the methods for jointly modeling responses and RTs. The last section presents Bayesian MCMC method for parameter estimation of the proposed model and the rationale for employing this method.

2.1 Cognitive Diagnostic Modeling

Cognitive diagnostic models, also known as diagnostic classification models, restricted latent class models, structured located latent class models, or multiple classification latent class models, were originally a collection of models of within-person production system to between-person measurement models characterized by coarser-grained attributes (Mislevy, 2018; Nichol, Chipman, & Brennan, 1995). Persons and tasks are described in terms of attributes, i.e., clusters or properties of knowledge and rules comprising a production system. Link functions such as identity, logit, or logarithmic determine the probability distributions for a person's performance on a task given the person and task values with respect to the attributes (Mislevy, 2018).

1.1.1 Latent Class Analysis

Specifically, latent class analysis (LCA) conditions response probabilities upon an unobserved latent categorical variable known as class membership (Macready & Dayton, 1977). This model is based upon three assumptions: a) response probabilities are class-specific and class-dependent; b) local independence, i.e., observed responses are independent given their class membership; and c) classes are mutually exclusive and exhaustive (von Davier & Lee, 2019). Mathematically, probabilities of response patterns in LCA are given as:

$$P(X_r = x_r) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1-x_{ir}} \quad (2.1)$$

where X_r and x_r are the observed responses, x_{ir} is the response to item i, π_{ic} is the probability of a correct response to item i for examinees in class c, and v_c is the mixing proportion, i.e., the proportion of examinees in class c. The joint probability of a particular response pattern is thus given as the product of the probability of a correct response and of an incorrect response across all items, assuming independence of item responses given class membership. This is weighted by the mixing proportion v_c of a given latent class and summed across all latent classes, giving the marginal likelihood of observed response patterns. An unrestricted LCA model of responses to i items generates 2^i response patterns or latent classes, with 2^i-1 mixing proportions and $2^i i$ item parameters to be estimated.

1.1.2 The DINA Model

The deterministic inputs, noisy "and" gate (DINA) model is essentially a restricted latent class model (Macready & Dayton, 1977), and is considered the

foundation of several cognitive diagnostic models (Junker & Sijtsma, 2001). This model identifies latent response variables as

$$\xi_{ij} = \prod_{k:Q_{jk}=1} \alpha_{ik} = \prod_{k=1}^{K} \alpha_{ik}^{Q_{jk}} \tag{2.2}$$

where α_{ik} indicates whether examinee *i* possesses attribute *k* and Q_{jk} indicates whether attribute *k* is required for task or item *j*. Specific attributes required for correctly responding to an item is fixed a priori by a $J \times K$ Q-matrix where the *J* rows are items and *K* columns are attributes that the items are designed to measure (Embreston, 1984; Tatsuoka, 1995). In this matrix, the q_{jk}th element yields a value of 1 when a correct response to the *j*th item requires mastery of the *k*th attribute; otherwise it takes on a value of zero. The notation ξ_{ij} indicates whether examine *i* has all the attributes required for item *j*. Tatsuoka (1995) defines the latent vectors $\boldsymbol{\alpha}_{i.}$ = ($\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}$) as *knowledge states*, and the vectors $\boldsymbol{\xi}_{i.} = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{ij})$ as representing a deterministic prediction of task performance based on each examinee's knowledge state.

The Item Response Function for a given item is presented as follows:

$$P(X_{ij} = 1 | \alpha, s, g) = (1 - s_j)^{\xi^{ij}} g_j^{1-\xi_{ij}} \tag{2.3}$$

where $s_j = \mathrm{P}(X_{ij} = 0 | \xi_{ij} = 1)$ and $g_j = \mathrm{P}(X_{ij} = 1 | \xi_{ij} = 0)$. Item parameters s_j and g_j, mnemonically termed slipping and guessing probabilities, are false negative and false positive rates for detecting ξ_{ij} from noisy observations X_{ij} (Junker & Sijtsma, 2001). ξ_{ij} as a binary function of binary inputs yields a value of 1 if and only if all the inputs are 1s and functions as the "and" gate component combining

deterministic input $\alpha_{ik}^{Q_{jk}}$. Assuming local independence, the joint probability of all responses is:

$$P(X_{ij} = x_{ij}, i, j|\alpha, s, g) = \prod_{i=1}^{N}\prod_{j=1}^{J}[(1-s_j)^{x^{ij}} {s_j}^{1-x^{ij}}]^{\xi^{ij}} [{g_j}^{x^{ij}}(1-g_j)^{1-x^{ij}}]^{1-\xi^{ij}} \quad (2.4)$$

1.1.3 The G-DINA Model

Based upon the DINA model, the G-DINA (*generalized deterministic inputs, noisy "and" gate*) model is a saturated model with more relaxed assumptions postulated to relate several CDMs with different formulations (de la Torre, 2011; de la Torre & Minchen, 2019). Assuming K_j^* attributes are required to correctly respond to item *j*, the G-DINA model classifies examinees into $2^{K_j^*}$ latent groups, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ denotes the number of attributes required for item *j*, and α_{lj}^* denotes the reduced attribute vector the elements of which are required for item *j*. For two given attribute vectors $\boldsymbol{\alpha}_{lj}^*$ and $\boldsymbol{\alpha}_{l'j}^*$, $\boldsymbol{\alpha}_{\boldsymbol{lj}}^* \leq \boldsymbol{\alpha}_{\boldsymbol{l'j}}^*$ if and only if $\boldsymbol{\alpha}_{\boldsymbol{lk}}^* \leq \boldsymbol{\alpha}_{\boldsymbol{l'k}}^*$, i.e., a reduced vector subsuming another reduced vector will have more ones. The G-DINA model estimates $2^{K_j^*}$ number of parameters for item *j* and is thus a generalization of the more restricted DINA model.

Three link functions used in specifying models for cognitive diagnosis are *identity*, *logit*, and *logarithmic* (de la Torre, 2011). Regardless of the link functions, all models in their saturated forms result in an estimation of $2^{K_j^*}$ number of parameters for item *j* and provide identical model-data fit. The G-DINA model using the identity link expresses the response probability given α_{lj}^* as:

$$P(\alpha_{lj}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^{*}} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk} \quad (2.5)$$

In this expression, δ_{j0} is the intercept for item *j*, representing the probability of a correct response in the absence of mastery of none of the required attribute. δ_{jk} is the main effect of α_k and indicates a change in the probability attributable to mastery of α_k. $\delta_{jkk'}$ is a first-order interaction effect of α_k and $\alpha_{k'}$, denoting the change in probability due to the mastery of both α_k and $\alpha_{k'}$. This effect is above and beyond the main effects of α_k and $\alpha_{k'}$. $\delta_{j12\ldots K_j^{*}}$ is the interaction effect of α1, ..., and $\alpha_{K_j^{*}}$, representing the change in probability as a result of mastery of all required attributes, This effect is above and beyond the main effects of the required attributes and the lower-order interaction effects.

The general CDM model using the logit link is equivalent to the log-linear CDM (Hagenaars, 1993; Henson & Templin, 2019) and is expressed as follows:

$$\text{logit}[P(\alpha_{lj}^{*})] = \lambda_{j0} + \sum_{k=1}^{K_j^{*}} \lambda_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \lambda_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \lambda_{j12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk} \quad (2.6)$$

Similarly, using the log link, the response probability given α_{lj}^{*} is expressed as:

$$\log P(\alpha_{lj}^{*}) = \upsilon_{j0} + \sum_{k=1}^{K_j^{*}} \upsilon_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \upsilon_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \upsilon_{j12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk} \quad (2.7)$$

Notwithstanding similar formulations in the three general CDMs, the nature of the effects described by these models is different (de la Torre, 2011). G-DINA and logit CDM describes the additive effect of attribute mastery on the probability and logit of the probability of success, whereas in the log CDM the effect is multiplicative. De la Torre (2011) notes the importance of noting this distinction as different link functions result in different reduced models even as the same set of constraints are imposed.

1.1.4 Summary and Discussion

LCA was developed as probabilistic models for classifying examinees with respect to mastery of specific concepts or skills (Macready & Dayton, 1977). The within-class model of LCA assumes independence of responses, which can be violated in assessment situations where test structure can induce highly related responses, as in testlet-based assessments (see section 2.2 for a detailed discussion of testlet modeling). The other drawback of LCA is its flexibility in accounting for dependence between observed variables by increasing the number of latent classes, which can result in the model overfitting the observed dependencies and a substantial increase in the number of parameters to be estimated (von Davier & Lee, 2019).

The DINA model is a commonly used CDM based upon which fine-grained inferences about cognitive information of the items and attribute profiles can be drawn to inform classroom instruction and learning and is appropriate to use when assessment tasks require the conjunctive use of several equally important attributes, and when lacking one required attribute is the same as lacking all required attributes (de la Torre, 2008; de le Torre & Douglas, 2004; Rupp et al 2010). This model is restrictive and parsimonious as only two parameters, a slipping parameter and a

guessing parameter, are required for estimating each item. As described earlier in this section, an integral component of the DINA model is the Q-matrix, which describes how items are related to attributes. Fit analysis in the DINA model assumes that the Q-matrix is correctly specified. Model fit analysis without verification of the completeness and accuracy of the Q-matrix can only be incomplete and partial (de la Torre, 2008). Further, as a non-compensatory model which specifies that the lack of one or more of the attributes required for correctly responding to an item cannot be compensated for by the presence of another, the DINA model assumes the attribute vectors in the same group to have the same probability of correctly responding to the items, which may not always be true. Attribute vectors in the same group may have varying level of deficiency with regard to the required attributes, hence their probabilities of producing a correct response may not be identical (de la Torre, 2011). Additionally, the DINA model was found to be affected by identifiability issues, with one empirical study showing it not being able to identify all attribute patterns (DeCarlo, 2011), and other studies showing that model identifiability requires the use of single-loaded items that only measures one attribute (DeCarlo, 2011; Fang, Liu, & Ying, 2017).

As a generalization of the DINA model, the G-DINA model extends the flexibility and usefulness of cognitive diagnostic modeling by allowing the fitting of different reduced models without necessitating estimation of the parameters from the original response data (de la Torre, 2011). Parameter estimation for several saturated models and a special class of reduced models can use the more efficient maximum likelihood estimation (MMLE), which is faster than the Markov chain Monte Carlo

(MCMC) algorithm. The model's flexibility becomes useful when researchers cannot specify the reduced CDMs a priori as they can still obtain parameter estimates based on the final model specification. Additional strengths include the possibility of using several CDMs simultaneously to construct a test, and of empirically verifying researchers' hypothesis regarding the underlying process for a subset or all of the items. This framework, however, cannot be used when constraints need to be set across items, as it allows for reduced models to be estimated one item at a time. The other drawback is the model's maximum likelihood estimation method which is not applicable to parameter estimates for reduced models that do not belong to the special class of reduced models (de la Torre, 2011).

In the joint model of response, RT, and ACs in testlet-based assessments for cognitive diagnosis proposed in this research study, the DINA model is the CDM chosen to fit the response data. As described above, the DINA model is a parsimonious CDM specifying two parameters for every item and is thus more interpretable and tractable compared to other CDMs. Examples of the application of the DINA model can be found in Macready and Dayton (1977), Junker and Sijstma (2001) and de la Torre and Douglas (2004). Given the prevalent use of testlets in educational assessments and dependence of responses likely resulting from this, the DINA model can be extended to specifically account for local response dependence. The next section reviews measurement models specifically addressing local item dependence induced by testlets.

2.2 Testlet Modeling

A testlet, also known as an item cluster or item bundle, is a unit of test construction that aggregates a group of items around a single content area (Wainer & Kiely, 1987). Testlets are typically characterized by a predetermined and fixed number of paths that an examinee can take, hierarchical, linear, or a combination of both. Compared to stand-alone items, testlets leverage tests' capabilities for efficiently addressing problems such as contextual effects, item ordering, and content balancing (Wainer & Kiely, 1987; Wainer & Lewis, 1990). The use of testlets as testing units is likely to induce local item dependence, as responses to a group of items embedded within the same context are likely to be more highly related and are thus a violation of the assumption of conditional independence (CI) for standard item response theory (IRT) models (Yen, 1993). Fitting standard IRT models to testlet responses induces overestimation of measurement precision and biased estimation of item difficulty and discrimination parameters, resulting in inaccurate inferences about the parameters (Wainer & Wang, 2000; Yen, 1993).

2.2.1 The Bayesian Random Effects Model for Testlets

Testlet models are developed primarily to account for the nesting of items within the same testlets and dependence of item responses as incurred by a common stimulus. Jiao, Wang, and He (2013) summarizes multiple perspectives on testlet modeling as indicative of how testlet effects are conceptualized. Essentially viewing the lack of CI as a form of unidimensional proficiency model misfit, Bradlow, Wainer, and Wang (1999) proposed the Bayesian random-effects model for testlets to explicitly account for the dependence structure of testlet items. Their Bayesian

hierarchical model modifies standard two-parameter IRT models to include an additional interaction term to model person-specific testlet effects for item test scores composed of a mixture of binary independent and testlet items. This model was extended into a three-parameter testlet model (Wainer, Bradlow, & Du, 2000), and to tests composed of a mixture of binary and polytomous items, independent and nested within testlets (Wang, Bradlow, & Wainer, 2002).

The general Bayesian model for testlets as proposed by Wang et al (2002) builds upon two basic IRT models: the three-parameter logistic model for binary items (Birnbaum, 1968) and Samejima's (1969) polytomous IRT model. These models are given as:

$$p_{ij}(1) = P\left(y_{ij} = 1 \middle| \theta, \omega_j\right) = c_j + \left(1 - c_j\right) logit^{-1}\left(t_{ij}\right), \text{and}$$

$$p_{ij}(r) = P\left(y_{ij} = r \middle| \theta, \omega_j, d\right) = \Phi\left(g_r - t_{ij}\right) - \Phi\left(g_{r-1} - t_{ij}\right), \qquad (2.8)$$

where $p_{ij}(r)$ is the probability of examine $i = 1, \dots, I$ receiving score $r = 1, \dots, R_j$ on item $j = 1, \dots, J$, c_j is the lower asymptote for binary item j, ω_j is the set of item j's parameters, g_r is the latent cutoff for the polytomous items, Φ is the normal cumulative density function, and t_{ij} is the latent linear score predictor. Standard form for the linear predictor t_{ij} is given as:

$$t_{ij} = a_j\left(\theta_i - b_j\right) \qquad (2.9)$$

where a_j, b_j, and θ_i are standard item slope, item difficulty, and latent trait parameters. Bradlow, Wainer, and Wang (1999) extends it to

$$t_{ij} = a_j\left(\theta_i - b_j - \gamma_{id(j)}\right), \qquad (2.10)$$

where $\gamma_{id(j)}$ denotes the testlet effect of item j to person i nested within testlet $d(j)$.

Prior distributions for the set of parameters in this model as embedded in the Bayesian hierarchical framework are specified as follows:

$$a_j \sim N(\mu_a, \sigma_a^2)$$

$$b_j \sim N(\mu_b, \sigma_b^2)$$

$$q_j \sim N(\mu_q, \sigma_q^2)$$

$$\theta_i \sim N(0,1)$$

$$\gamma_{id(j)} \sim N\left(0, \sigma_{d(j)}^2\right) \quad (2.11)$$

where $N(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2 and $q_j = logit\ (c_j)$. The mean and variance of the ability distribution are fixed at 0 and 1 to identify the model. $\sigma_{d(j)}^2$ representing testlet-specific excess dependence is allowed to vary across testlets. Hyperpriors for the means of the set of parameters in the prior distributions are specified as $\mu_a \sim (0, V_a)$, $\mu_b \sim (0, V_b)$, and $\mu_q \sim (0, V_q)$, where $V_a^{-1} = V_b^{-1} = V_q^{-1}$ were set to 0. Slightly informative hyperpriors specified for all prior variances are given by $\sigma_z^2 \sim \sigma_{g_z}^2$ which is an inverse chi-square random variable with g_z degrees of freedom where $g_z = 0.5$ for all distributions.

2.2.2 The Rasch Testlet Model

Similarly, Wang and Wilson (2005) proposed the Rasch testlet model for both dichotomous and polytomous items in testlet-based tests, demonstrating it to be a special case of the multidimensional random coefficients multinomial logit model (MRCMLM).

The one-parameter Rasch testlet model is given as:

$$p_{ni1} = \frac{\exp\left(\theta_n - b_i + \gamma_{nd(i)}\right)}{1 + \exp\left(\theta_n - b_i + \gamma_{nd(i)}\right)} \tag{2.12}$$

This equation can be expressed as:

$$\log\left(\frac{p_{ni1}}{p_{ni0}}\right) = \theta_n - b_i + \gamma_{nd(i)} \tag{2.13}$$

where p_{ni1} and p_{ni0} are the probabilities for scoring 1 and 0 on item i for person n. For polytomous items, this can be extended to:

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - b_{ij} + \gamma_{nd(i)}, \tag{2.14}$$

where p_{nij} and $p_{ni(j-1)}$ are the probabilities for scoring j and $j-1$ on item i for person n and b_{ij} is the jth step difficulty for item i. Let

$$b_{ij} = b_i + \pi_{ij} \tag{2.15}$$

where b_i is the difficulty of item i, and π_{ij} is the jth threshold parameter of item i. Constraining the threshold parameters to be the same across items, such as $\pi_{ij} = \pi_j$, the Rasch testlet model for polytomous items is reduced to

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - (b_i + \pi_i) + \gamma_{nd(i)} \tag{2.16}$$

θ and γ are assumed to be independently and normally distributed. Thus $\theta' = [\theta, \gamma_1, \ldots, \gamma_d, \ldots, \gamma_D]$ has a multivariate normal distribution $N(\mu, \Sigma)$, where for model identification, μ is set to zero, and Σ is constrained to be a diagonal matrix.

2.2.3 Higher Order Testlet Response Models

To simultaneously account for both testlet design and hierarchical structure in latent traits, Huang and Wang (2012) developed higher order testlet response models on the basis of higher order IRT models for hierarchical latent traits (e.g. de la Torre

& Douglas, 2004; de la Torre & Hong, 2010; de la Torre & Song, 2009) and testlet response models (Bradlow et al, 1999; Wainer et al, 2000; Wainer et al, 2007; Wang & Wilson, 2005). In these models, the relationship between first and second order latent traits is specified as:

$$\theta_{nv}^{(1)} = \beta_v \theta_n^{(2)} + \varepsilon_{nv}^{(1)} \tag{2.17}$$

where the vth first-order latent traits denoted by $\theta_{nv}^{(1)}$ for person n is a linear function of the same second-order latent trait denoted by $\theta_n^{(2)}$. β_v is the regression weight (factor loading) indicating the magnitude of the effect of $\theta_n^{(2)}$ on $\theta_{nv}^{(1)}$. The error term $\varepsilon_{nv}^{(1)}$ is assumed to be normally distributed with a mean of zero and independent of other ε and θ.

The 3 PL higher order testlet response model (3P-HTM) for dichotomous items accounting for hierarchical latent traits is given as:

$$P_{ni1v} = \pi_{iv} + (1 - \pi_{iv}) \times \frac{\exp\left[\alpha_{iv}\left(\beta_v \theta_n^2 - \delta_{iv} + \varepsilon_{nv}^{(1)} + \gamma_{nd(i)v}\right)\right]}{1 + \exp\left[\alpha_{iv}\left(\beta_v \theta_n^2 - \delta_{iv} + \varepsilon_{nv}^{(1)} + \gamma_{nd(i)v}\right)\right]} \tag{2.18}$$

where P_{ni1v} is the probability of responding correctly to item i in test v for person n, π_{iv} is the asymptotic parameter for item i in test v, α_{iv} is the discrimination parameter, and δ_{iv} is the item difficulty parameter. $\gamma_{nd(i)v}$ is the additional latent trait accounting for local dependence for items nested within testlet d of test v and is assumed to be normally and independently distributed (Wainer et al, 2007; Wang & Wilson, 2005).

The 3 PL higher order testlet response model (3P-HTM) for polytomous items, referred to as the generalized partial credit higher order testlet model (GP-HTM) is given as:

$$\log\left(\frac{P_{nijv}}{P_{ni(j-1)v}}\right) = \alpha_{iv}\left(\beta_v\theta_n^2 - \delta_{iv} - \tau_{ijv} + \varepsilon_{nv}^{(1)} + \gamma_{nd(i)v}\right) \tag{2.19}$$

where P_{nijv} and $P_{ni(j-1)v}$ are the probabilities of scoring j and $j-1$ on item i in test v for person n, and τ_{ijv} is the jth threshold parameter for item i in test v.

2.2.4 The Three-Level One-Parameter Testlet Model and Its Extensions

Jiao, Wang, and Kamata (2005) proposed the three-level one-parameter model for dichotomous items where the contextual effect of testlets on items is accounted for from the hierarchical generalized linear modeling perspective, configuring it to be a three-level hierarchical generalized linear model for item analysis (Kamata, 2001). At level 1, the log-odds of person j responding to item i nested within testlet d is expressed as:

$$\log\left(\frac{p_{idj}}{1-p_{idj}}\right) = \eta_{idj} = \beta_{odj} + \sum_{q=1}^{k-1}\beta_{qdj}X_{qidj} \tag{2.20}$$

where p_{idj} is the probability of person j correctly responding to item i nested within testlet d, X_{qidj} is the qth dummy variable for person j, β_{odj} is an intercept term representing the reference item effect, and β_{qdj}, the coefficient for X_{qidj} where $q = 1, \ldots, k-1$, and k is the total number of items on the test, represents the unique effect for item q relative to β_{odj}. Level 2 models the testlet effects and is given as:

$$\beta_{odj} = \gamma_{00j} + \mu_{0dj} \text{ and}$$

$$\beta_{qdj} = \gamma_{q0j} \tag{2.21}$$

where level 1 reference item effect is decomposed into a fixed effect γ_{00j} and a random testlet effect μ_{0dj} comparable to the testlet effect parameter $\gamma_{jd(i)}$ in the

general Bayesian random effects testlet model (Bradlow et al, 1999). Level 3 models person effects by further partitioning the level 2 fixed effect γ_{00j} into a fixed component and a random component. Level 3 model is given as

$$\gamma_{00j} = \pi_{000} + \omega_{00j} \text{ and } \gamma_{q0j} = \pi_{q00} \tag{2.22}$$

where $\omega_{00j} \sim N(0, \sigma_{\omega}^2)$ is the random person effect and the effects for the items remain fixed. Jiao, Wang, and He (2013) demonstrate the equivalence between this model and the Rasch testlet model, as the three equations can be combined into

$$P_{jdi} = \frac{1}{1 + \exp\left[-\left(\omega_{00j} - \left(-\pi_{000} - \pi_{q00}\right) + \mu_{0dj}\right)\right]} \tag{2.23}$$

where $\omega_{00j} = \boldsymbol{\theta_j}, -\pi_{000} - \pi_{q00} = \boldsymbol{b_j}, \ \mu_{0dj} = \gamma_{jd(i)}$.

In educational settings, the nesting of students within classes and of classes within schools is a norm rather than an exception (Bryk & Raudenbush, 1992). Applying standard IRT models to tests with person dependence structure may result in biased parameter estimates, compromising the validity of the inferences we can draw regarding item parameters and student proficiency. To simultaneously account for both item dependence and person dependence, Jiao, Kamata, Wang, and Jin (2012) further extended the three level one-parameter testlet model to a four-level IRT model for dual local dependence, where the levels in their ascending order respectively represent item effects, testlet effects, the effects of persons fully crossed with testlets and items, and examinee group effects as incurred by the nesting of examinees within classes, schools, or school districts.

The four-level IRT model for dual dependence for dichotomous items is given by

$$P_{jdig} = \frac{1}{1 + \exp\left[-(\theta_j + \theta_g - b_i + \gamma_{jd(i)}\right]} \quad (2.24)$$

where θ_j denotes the person-specific ability for person j, θ_g denotes the group-specific ability for group g, b_i is the item difficulty for item i, and $\gamma_{jd(i)}$ represents the testlet effect for person j on testlet d. In this model, $\sigma^2_{\theta_g}$ indicates the magnitude of the group effects. This model further assumes the effects of item-clustering and person-clustering, and person ability to be mutually exclusive and independent and the residuals variance to be uncorrelated after controlling for the three variances.

This model was generalized into polytomous multilevel testlet models for person and item clustering for dichotomous and polytomous items (Jiao & Zhang, 2015). Mathematically, this model is given as

$$P_{jtigk}\left(X_i = x \middle| a_i, d_{ik}, \theta_{jg}, \delta_g, \gamma_{jt(i)}\right) = \frac{exp\left[\sum_{s=0}^{x} a_i\left(\theta_{jg} + \delta_g + \gamma_{jt(i)} - d_{ik}\right)\right]}{\sum_{s=0}^{K} exp\left[\sum_{s=0}^{k} a_i(\theta_{jg} + \delta_g + \gamma_{jt(i)} - d_{ik})\right]} \quad (2.25)$$

where $\sum_{s=0}^{0} a_i(\theta_{jg} + \delta_g + \gamma_{jt(i)} - d_{ik}) = 0$, a_i is the item discrimination parameter, d_{ik} is the item step difficulty, δ_g is the group effect, θ_{jg} is the person-specific ability indicating the deviation of person j's ability from the group ability, and $\gamma_{jt(i)}$ is the testlet effect for person j interacting with testlet t. This equation thus represents the probability of person j with person-specific ability θ_{jg} in group g receiving a score of X on item i with a step difficulty of d_{ik} in testlet t. Person clustering effect δ_g, if present, will vary across groups. Its variance indicates the magnitude of the person clustering effect. Likewise, item clustering effect $\gamma_{jt(i)}$ varies across persons and across testlets, but remains constant for the same person on the items nested within the same testlet. The variance of $\gamma_{jt(i)}$ indicates the magnitude of the item clustering

effect. Item clustering effects, person clustering effects, and person ability are additive, and assumed to be mutually exclusive and independent. Further, residual variances area assumed to be uncorrelated after controlling for the three random effects.

2.2.5 The Bayesian Multilevel Multidimensional IRT (BMMIRT) Model for Locally Dependent Data

A Bayesian approach differs from frequentist approaches in treating the model parameters not as fixed but as random and uses distributions that reflect *a priori* knowledge to model beliefs about them (Levy & Mislevy, 2016). The BMMIRT model for locally dependent data was proposed to investigate whether the assumption about the orthogonality of the dimensional structure in previously reviewed models is justifiable and the extent to which it is violated (Fujimoto, 2018). Based upon the generalized partial credit model (Muraki, 1992), the probability of a response of c to item k by person ij (person i in group j) given all model parameters (Ψ) is expressed as

$$P(Y_{ijk} = c|\Psi) = \frac{\prod_{x=0}^{c} H(\eta_{kx}) \prod_{l=c+1}^{m_k} [1 - H(\eta_{kl})]}{\sum_{w=0}^{m_k} (\prod_{x=0}^{w} H(\eta_{kx}) \prod_{l=w+1}^{m_k} [1 - H(\eta_{kl})]} \tag{2.26}$$

where $H(\eta_{kx})$ is the cumulative density function for the conditional probability of a response c to item k and is given as

$$P[Y_{ijk} = c|Y_{ijk} = c \vee Y_{ijk} = c - 1, \Psi] = H(\eta_{kc}) = \frac{\exp(\eta_{kc})}{1 + \exp(\eta_{kc})} \tag{2.27}$$

The systematic component (η_{kc}) for the model is

$$\eta_{kc} = \xi_k \theta_{ij}^T + \tilde{\xi}_k \theta_j^T - (\beta_k + \tau_{kc}), \tag{2.28}$$

where

$$\xi_k = \chi \circ \alpha_k \text{ and}$$

$$\tilde{\xi}_k = \tilde{x} \circ \tilde{\alpha}_k \tag{2.29}$$

In this model, discrimination parameters are α_k, $\tilde{\alpha}_k$, χ, and $\tilde{x}$, where α_k and $\tilde{\alpha}_k$ are the 1 x *D* vector of item-specific discriminations at Level 2 and Level 3, and χ and $\tilde{x}$ are 1 x $\widetilde{D}$ vector of overall discriminations at Level 2 and Level 3. Latent trait parameters are θ_{ij} and θ_j, with θ_{ij} being the 1 x *D* vector of latent trait dimensional positions for person i in cluster j at Level 2, and θ_j being the 1 x $\widetilde{D}$ vector of latent trait dimensional positions for cluster j at Level 3. β_k is the overall difficulty for item k and τ_{kc} is the cth element in the 1 x m_k vector of relative intercepts, denoting item k's relative intercept for category c.

2.2.6 The Bayesian Covariance Structure Model (BCSM) for Testlets

The Bayesian covariance structure model (BCSM) for testlets extends standard IRT models with a covariance structure to account for dependencies among testlet items (Fox, Wenzel, & Klotzke, 2020). Without using testlet effects, the model efficiently addresses problems such as sample size restrictions and computational burden as incurred by the inclusion of the testlet parameter to account for the dependence for each combination of testlet and test taker. By assuming a common covariance between responses to items in a test across test takers, BCSM estimates the same number of additional model parameters as the number of testlets, making it more suitable for small sample sizes, incomplete design, and for tests that contains just a few items for many testlets (Fox et al, 2020).

The BCSM for testlets expresses the latent responses to testlet d by individual i as

$$Z_i^d = a^d\theta_i - b^d + e_i^d$$

$$\theta_i \sim N(0, \sigma_\theta)$$

$$e_i^d \sim N(0, \Sigma_d) \quad (2.30)$$

where $\Sigma_d = \mathrm{I}_{nd} + \mathrm{J}_{nd}\sigma_{nd}$ and I_{nd} and J_{nd} are the identify matrix and a matrix of ones, respectively, both of dimensions n_d. This model assumes the latent responses to items in testlet d to be multivariate normally distributed. Further, σ_{γ_d} as a covariance parameter can be negative or zero, making it possible to represent a negative association among testlet items or no testlet effects.

2.2.7 Summary and Discussion

All but one testlet models reviewed in this section are modifications of standard IRT models and higher-order latent trait models to include an additional interaction parameter to specifically model the testlet effects of persons interacting with items nested within a given testlet. As is demonstrated by Jiao et al. (2013), the general Bayesian model for testlets (Wang et al., 2002), the Rasch testlet model (Wang & Wilson, 2005), and the three-level one-parameter testlet model (Jiao et al., 2005) are essentially equivalent models in terms of specifying the testlet effects as an additional parameter in the one-parameter Rasch model. Conceptually, however, in both the general Bayesian model for testlets and the Rasch testlet model, the testlet effects are represented by a person-specific random effects parameter, while the Rasch testlet model views them as additional dimensions, and the three-level one-parameter testlet model describes them as representing the contextual effects on items

nested within a context. Additionally, both the general Bayesian model for testlets and the Rasch testlet model apply to both binary and polytomous items nested within testlets, yet the estimation methods are different for the two models. The general Bayesian model for testlest embeds the modified 3PL IRT model (Birnbaum, 1968) and Semijima's (1969) polytomous IRT within the Bayesian framework and uses MCMC to obtain inferences regarding model parameters. The Rasch testlet model, demonstrated to be a special case of the MRCMLM, applies MMLE and Bock and Aitkin's (1981) formulation of the EM algorithm (Dempster, Laird, & Rubin, 1977). By way of contrast, the three-level one-parameter testlet model uses another estimation method, the sixth-order approximation Laplace (Laplace) method, for parameter estimation.

Compared with models assuming conditional independence, the general Bayesian model for testlets yields unbiased estimates of the tests' precision. An additional strength of the model is its use to score a test constructed of any combination of item formats, such as multiple-choice items, fill in the blank items, and items rated by expert raters, which allows test developers to choose item formats that best suit the constructs of interest. It further shows that as the sample size is increased, the root mean square error of the estimates decreases to an acceptable level. When fit to an empirical test data set, the Rasch testlet model fit the data statistically better than the standard IRT model, which overestimates the test reliability and yields difficulty estimates that shrink slightly toward the mean. This model additionally allows for simultaneous calibration of multiple tests (with or without testlets) and direct estimation of the correlations between latent traits and

more precise measures for individual persons than several calibrations of latent traits (Wang, Chen, & Cheng, 2004).

These models, however, only account for local item dependence, whereas in educational settings, the nesting of students within schools, and of schools within a larger geographical context can also incur person clustering and local person dependence. The four-level IRT model for dual dependence (Jiao et al., 2012) and its extension to polytomous items and complex sampling design, the polytomous multilevel testlet models (Jiao & Zhang, 2015), account for both testlet effects and person clustering effects. Compared to alternative models that ignore local item and/or person dependence, both models yield more accurate item and/or person ability parameter estimates and can be applied to K-12 state assessment programs and large-scale national and international assessment programs.

The other limitations with the general Bayesian model for testlets are the number of testlet parameters which can become very large if a test consists of many testlets and is administered to a large number of examinees and sample size restrictions which limits its applicability (Fox et al., 2020). An additional limitation is the use of the inverse-gamma prior for testlet variance, which is restricted to be positive and does not include the point of no testlet variance. As the testlet variance gets close to 0, the inverse-gamma prior can become biased by overstating the level of dependence of the testlet items. The BCSM for testlets (Fox et al., 2020) addresses these limitations by modeling the testlet dependences through an additional covariance structure, which significantly reduces the number of model parameters, and at the same time allowing testlet dependence to be set at 0 or negative when

estimating the model parameters. Compared to the general Bayesian model for testlets, the BCSM testlet model is more efficient and flexible and can be extended to model different types of clustered items.

The inclusion of a testlet parameter is the approach adopted in the proposed research study. One of the purposes of this research is to estimate the effects of accounting for dependence of responses and RTs on the precision with which model parameters, including person ability parameters and item parameters, are estimated. The proposed research additionally examines the effects of the inclusion of RT in the modeling of responses on model performance and parameter estimation. The next section reviews RT models, which is followed by a review of joint models of responses and RT with or without dependency in the Section 2.4.

2.3 Response Time (RT) Modeling

Response time (RT), also known as reaction time, is an important source of information for understanding aspects of cognitive processes underlying test performance (De Boeck & Jeon, 2019; van der Linden, 2007). Experimental psychologists in their long-standing tradition decomposed reaction time into stages of information processing to understand the structure of mental activity (Sternberg, 1969). From a test design perspective, RT as a type of process data translates directly into evidence accumulation and holds implications for the validity of the inferences regarding test-takers and test use.

Two traditions in RT modeling are distinct models for RT and models integrating response and RT (van der Linden, 2009). De Boeck and Jeon (2019) further classify approaches to RT modeling into four broad categories: a) RT models

with RT as the sole dependent variable; b) joint models in which RT and another kind of variable are both dependent variables; c) dependency models in which RT and other data are jointly modeled with the possibility of dependences; and d) RT as covariate models in which another variable varies as a function of RT. This section reviews distinct RT models and RT as covariate models, followed by a review of joint models for RT and response with or without dependences in Section 2.4.

2.3.1 Lognormal Response Time Model

RT distributions are positively skewed, with their means and variances positively correlated (Luce, 1986; Maris, 1993; Townsend & Ashby, 1983). One of the widely used distributions reported as having a good fit to RT data is lognormal distribution. van der Linden (2006) proposed a lognormal model for RT on test items in which RT distributions are determined by a distinct set of item and person parameters. Assuming response time t_i for a fixed person on item i is the realization of a random variable T_i, the normal density for the distribution of the log RT, $\ln T_i$, is written as:

$$f(t_i;\tau,\alpha_i,\beta_i)=\frac{\alpha_i}{t_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_i-(\beta_i-\tau)\right)\right]^2\right\} \qquad (2.31)$$

where τ is the person speed parameter, β_i denotes the time intensity of item i, and α_i is the discrimination parameter modifying the relationship between t_i and its mean $\beta_i-\tau$. When estimating the item parameters, the following constraint is imposed on the set of values τ_j to identify the model:

$$\sum_{j=1}^{N}\tau_j=0 \qquad (2.32)$$

This constraint implies that

$$n^{-1}\sum_{i=1}^{n}\beta_i = (nN)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N}\mu_{ij} \tag{2.33}$$

which equates the average item parameter β_i to the expected log time averaged over persons and items. The values of the person parameter τ_j thus represents deviations from this average. The proposed model is analogous to the 2PL IRT model in imposing a similar structure on the means of the variables and having the same discrimination parameter moderating the effects of an item and person parameter.

2.3.2 Alternative Distribution Models

Alternative distributions possessing the aforementioned properties are gamma, inverse Gaussian, ex-Gaussian, Weibull and Gumble, and shifted Wald (De Boek & Jeon, 2019; Maris 1993) and were used as descriptive distributions in the modeling and analysis of RT data. For example, Maris (1993) formulated additive, multiplicative, and combined additive-multiplicative models for gamma distributed random variables based upon which an analysis of response time data from a mental rotation experiment was conducted. Lo and Andrews (2015) in their reanalysis of three experiments investigating the effects of word frequency and stimulus quality applied generalized linear mixed-effect model (GLMM) to raw RT assuming Gamma or Inverse Gaussian distribution for the response time data. Relaxing the normality assumption for the dependent variable, GLMMs allows the assumptions regarding the relationship between the predictors and the dependent variable to be tested independently of the assumption for the dependent variable. Loeys, Rosseel, and Baten (2011) proposed a joint model for the reaction time and accuracy assuming a

log-normal distribution and a shifted three-parameter Weibull distribution for the response time. Using a Bayesian hierarchical framework, their joint model provides for estimation of the correlation between item intensity and difficulty at the item level, and between speed and ability at the subject level. A simulation study shows the reduction in bias gains compared to the separate modeling approach.

Ex-Gaussian and the shifted Wald distributions are used both as a measurement model and as an intra-individual cognitive process model for RT data (e.g., Anders, Alario, Van Maanen, 2016; Burbeck & Luce, 1982; Luce, 1986; Wald, 1947). Anders et al (2014) describe the shifted Wald distribution as an accumulation process model that provides a clear signal-to-response threshold interpretation of the RT data, with its parameters γ corresponding to the accumulation of the internal signal X, α to the threshold for initiating the physical process, and θ to the time lapsed external of signal accumulation. The versatility and usefulness of this model was demonstrated on three RT data sets representing different modes of responding. The ex-Gaussian distribution can be described by three parameters: the mean and standard deviation of the Gaussian component, μ and σ, and the mean of the exponential component, τ. This distribution was conjectured to represent the durations of various components of cognitive processing (e.g., Hohle, 1965). Matzke and Wagenmarkers (2009) on the other hand cautioned against interpreting the changes in the parameters of the ex-Gaussian and the shifted Wald distributions in terms of underlying cognitive processes. Their simulation study and empirical study demonstrated that the parameters of the two distributions do not correspond uniquely to the parameters of the diffusion model.

2.3.3 Response Time as a Function of Response Accuracy

RT as a function of response accuracy models are the models in which RT is the dependent variable (De Boeck & Jeon, 2019). van der Linden (2009) describes these models as "RT models that incorporate responses". These models usually condition the RT distributions upon responses, with distributions for correct responses being different from the distributions for incorrect responses. A well-known example of this type of models is Thissen's (1983) lognormal RT model for timed testing using 2PL IRT. This model assumes that the variations in item responses and response times were attributable to the same sources, and that responses and latencies could both be used in the estimation of examinee ability and item easiness. The model is given as

$$\log(t_{ij}) = v + s_i + u_j - b(a_j\theta_i + c_j) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2.34}$$

where v is the overall mean log response time, s_i and u_j are the person and item slowness parameters, b is a regression parameter reflecting how latency relates to effective ability and item easiness, and a_j, θ_i, and c_j are the discrimination, effective ability and item easiness parameters in the 2PL model. The person and item slowness parameters s_i and u_j are introduced to represent the extra effect of the examines and items on the RT that are not related to the trait that the items are designed to measure.

2.3.4 Response Time as a Covariate Models

RT as covariate models are models in which RT functions as a covariate, and response accuracy is the dependent variable (De Boeck & Jeon, 2019). van der Linden (2009) describes them as response models incorporating RT. Two modeling

approaches fall into this category: models assuming speed-accuracy trade-off (SAT) and generalized linear mixed model (De Boeck & Jeon, 2019).

Models Assuming SAT. SAT refers to the fact that there exists an inverse relationship between speed and accuracy. The relationship between mean RT and the probability of a correct response across conditions is called the Speed-Accuracy Trade-off Function (SATF) and is usually depicted as an increasing ogive-like function (Roskam, 1997). An example of this is Roskam's (1987, 1997) Rasch response time model for time-limit tests. In this model, the probability of a correct response conditional upon the response time for a given condition, the Conditional Accuracy Function (CAF; Luce, 1986), is given as

$$P\left(U_{ij}=1\middle|t_{ij},j,i\right)=\frac{\theta_j t_{ij}}{\theta_j t_{ij}+\varepsilon_i}=\frac{\exp\left(\xi_j+\tau_{ij}-\sigma_i\right)}{1+\exp\left(\xi_j+\tau_{ij}-\sigma_i\right)} \tag{2.35}$$

where θ is person ability, ε is item difficulty, and t is the RT. ξ, σ, and τ are the logarithms of θ, ε, and t. Roskam defines "the effective ability parameter for item i" as "mental speed times processing time". As the model implies, an increase in t results in an increase in CAF for item i, henceforth a trade-off between speed and accuracy. Roskam (1997) further assumes a Weibull distribution for RT where the hazard function defined as the probability density that the response is given at time t, conditional upon not given yet, is given as

$$h_{ij}(t)=\frac{\theta_j}{\varepsilon_i\delta_j}t \tag{2.36}$$

δ is the person persistence in continuing working on item i. This function implies a direct relationship between item difficulty and RT, and between person persistence

and RT, and an inverse relationship between mental speed and RT. The hazard function defines the probability density and probability distribution functions.

Another example is Wang and Hanson's (2005) four-parameter logistic response time (4PLRT) model that incorporates response times in the 3PL IRT model and can be applied to power tests. The probability of a correct response to item j by person i is given as

$$P\left(x_{ij}=1\middle|\theta_i,\rho_i,a_j,b_j,c_j,d_j,t_{ij}\right)=c_j+\frac{1-c_j}{1+e^{-1.7a_j\left[\theta_i-\left(\frac{\rho_i d_j}{t_{ij}}\right)-b_j\right]}} \qquad (2.37)$$

where a, b, and c are the discrimination, difficulty, and guessing parameters, and θ is the person ability parameter as they are interpreted in the 3PL IRT model. d is the item slowness parameter, ρ is the person slowness parameter, and t is the RT by this person to this item. The person slowness parameter indicates test-takers' work pace as they response to items. The item slowness parameter indicates how items react to RT. The product of these two parameters determines the rate of increase in the probability of a correct answer as a function of RT. An increase in RT results in an increase in the probability of a correct response. Wang and Hanson further note that the person slowness parameter does not indicate the amount of time a test-taker spends on a particular item and may or may not relate to RT. As RT for a particular item increases to infinity, the 4PLRT model reduces to the 3PL IRT model.

GLMM-based Models. Goldhammer, Steinwascher, Kroehne, & Naumann (2017) extends Roskam's (1987, 1997) Rasch RT model for time-limit tests for a single condition to estimate RT effects both within and across conditions. Speed differences can be a confounding factor when interpreting observed differences in

ability estimates when individuals completing a test have the option to choose their speed, as differences can be attributed to ability, speed chosen, or both (van der Linden, 2009). By using a GLMM approach, the proposed SATF and residual CAF model address the relation between speed and accuracy both intra-individually and inter-individually.

In this model, the observed RT of a person p $p = 1, \dots, P$) completing item $i(i = 1, \dots, I)$ in speed condition $c(c = 1, \dots, C)$ is decomposed into:

$$t_{pic} = \overline{t_{\dots}} + \left(\overline{t_{..c}} - \overline{t_{\dots}}\right) + \left(\overline{t_{p.c}} - \overline{t_{..c}}\right) + \left(\overline{t_{.ic}} - \overline{t_{..c}}\right) + t_{pic}^{(*)} \qquad (2.38)$$

where $\overline{t_{\dots}}$ is the grand mean of RT, $\overline{t_{..c}}$ is the condition average, $\overline{t_{p.c}}$ is the person average per condition, $\overline{t_{.ic}}$ is the item average per condition, and $t_{pic}^{(*)}$ is the residual. The residual $t_{pic}^{(*)}$ represents RT differences after removing the effects of item and person and is an indication of the extent to which the RT deviates from the expected RT in condition c given the person's speed and the item's time intensity. This is the random response time effect model for multiple speed conditions.

Goldhammer et al (2017) incorporate person and item differences in the SATF. The person SATF model describes the effect of within-person speed differences on the probability of a correct response and is given as

$$P\left(X_{pic} = 1\right) = \Psi^{-1}\left(\beta_0 + b_{0p} + b_{0i} + \left(\beta_1 + b_{1p}\right)\overline{t_{p.c}} + \left(\beta_2 + b_{2i}\right)\left(\overline{t_{.ic}} - \overline{t_{..c}}\right)\right) \qquad (2.39)$$

where Ψ^{-1} is the inverse logit function, β_0 is the general intercept, b_{0p} is the random person intercept for individual performance differences in the reference condition, and b_{0i} is the random intercept across items or item easiness. β_1 is the average person SATF slope representing the fixed effect of speed across conditions and b_{1p} is the

random person SAFT slope representing the individual deviation from it. β_2 is the fixed effect of relative time intensity across conditions, and b_{2i}, the random item slope, is the related item-specific deviation. The person SATF slope β_1 indicates the rate at which information is accumulated to give the response. The item SATF model describes the effect of intra-item differences in time intensity across speed conditions and is similarly given by

$$P(X_{pic} = 1) = \Psi^{-1}(\beta_0 + b_{0p} + b_{0i} + (\beta_1 + b_{1p})(\overline{t_{p.c}} - \overline{t_{..c}}) + (\beta_2 + b_{2i})\overline{t_{.ic}} \tag{2.40}$$

where β_1 denotes the fixed effect of relative speed across conditions, b_{1p} is the random person slope denoting related individual differences, β_2 is the average item SATF representing fixed effect of item time intensity, and b_{2i} is the random item SATF slope representing the related item-specific deviation. The item SATF slope β_2 indicates on average how fast information is gained and lost by item as speed changes.

Using the RT residual $t_{pic}^{(*)}$ as a predictor in the person SATF model, the overall SATF and residual CAF model is given by

$$P(X_{pic} = 1) = \Psi^{-1}\left(\beta_0 + b_{0p} + b_{0i} + (\beta_1 + b_{1p})\overline{t_{p.c}} + (\beta_2 + b_{2i})(\overline{t_{.ic}} - \overline{t_{..c}}) + \left(\sum_{k=1}^{C}(\beta_{3c} + b_{3pc} + b_{3ic})I_{c=k}t_{pic}^{(*)}\right)\right) \tag{2.41}$$

where the variable $I_{c=k}$ indicates whether the observation was made under condition c. The last term in this expression represent the effects of the residual RT, where, for

a given condition, β_{3c} is the residual CAF slope representing the fixed effect of residual RT, b_{3pc} is the random person component, and b_{3ic} is the random item component. The residual RT effects indicate how deviation from expected RT affect response accuracy. The CAF and residual CAF model for a single condition is given as

$$P(X_{pi} = 1) = \Psi^{-1}(\beta_0 + b_{0p} + b_{0i} + (\beta_1 + b_{1i})\overline{t_{p.}} + (\beta_2 + b_{2p})\overline{t_{.i}} + (\beta_3 b_{3p} + b_{3i})t_{pic}^{(*)}) \quad (2.42)$$

where β_1 is the average person CAF slope relating differences in speed to accuracy, b_{1i} is the variation of this relation across items, β_2 is the average item CAF slope relating differences in item intensity to accuracy, and b_{2p} is the variation of the relation across persons.

2.3.5 Summary and Discussion

This section reviews distribution models for RT, RT as a function of response accuracy models, and RT as a covariate model. Descriptive distributions used for the modeling and analysis of RT include lognormal, gamma, inverse Gaussian, ex-Gaussian, Weibull and Gumble, and shifted Wald distributions. The lognormal RT model (van der Linden, 2006) is analogous to the 2PL IRT model with its own set of RT parameters: person speed parameter τ, item time intensity parameter β_i and discrimination parameter α_i. Other distribution models reported as having a good fit for RT data are used for fitting RTs data from psychological experiments, such as mental rotation experiments (Maris, 1993), lexical decision tasks (Lo & Andrews,

2013), and word recognition tasks (Loeys et al, 2011). Two of the distributions, ex-Gaussian and the shifted Wald distributions, were used as both a measurement model and as an intra-individual cognitive process models for RT data, with parameters in the distribution model corresponding to the parameters in the cognitive process model (e.g., Anders et al, 2014; Burbeck & Luce, 1982; Luce, 1986; Wald, 1947). Luce (1986) concludes that it is unclear how cognitive processes relate to RT distributions. Matzke and Wagenmakers (2009) similarly caution again interpreting changes in the parameters as indicative of underlying cognitive processes. On the same note, van der Linden (2006) notes that the exponential and gamma distributions models follow strict assumptions on the problem-solving process underlying responses to an item, which is these models' weakness as the problem-solving process may not meet the assumptions of these models.

SAT models assume an inverse relationship between speed and accuracy and incorporate a RT parameter t_{ij} in IRT models for response accuracy such that an increase in t_{ij} results in an increase in the probability of a correct response (e.g., Raskam, 1997; Wang & Hanson, 2005). In reaction time research, SAT is typically represented by the positive relationship between the proportion of correct tasks and the average time on the tasks (e.g., Luce, 1986; van der Linden, 2009). As van der Linden (2009) notes, SAT in reaction time research is equivalent to speed-ability trade-off in testing, which is a within-person phenomenon and implies a monotonically decreasing relation between speed and ability. Speed-ability trade-off further implies that test scores do not reflect the level of test-takers' abilities unless constancy of speed is assumed. In this sense, speed and ability are related through the

function $\theta = \theta(\tau)$ and models of response and RT should treat the effective speed and ability of the test takers as fixed parameters (van der Linden, 2009). Thus, SAT models are for persons with fixed levels of ability and speed and seem to confound the within-person and fixed-person levels as speed-ability trade-off is only evidence when there is a change of strategy or condition (van der Linden, 2007).

The GLMM-based models provide for estimation of the effects of RT on response accuracy within and across experimental speed conditions (Goldhammer et al, 2017). The person and item SATF slope capture the between condition effects, indicating individuals' efficiency in accumulating information regarding the correctness of the response. The residual CAF slopes estimate the within condition effects and are an indicator for the mode of information processing. These models are proposed for measures including simple cognitive tasks with a strong speed component and can be extended with additional person and item indicators to provide better understanding of factors affecting the individual and item differences in the RT effects (Goldhammer et al, 2017). Analyses of the empirical datasets applying the models suggest that the association between speed and accuracy depends on the tasks and that the association is less negative (or more positive) for more difficult items.

The proposed research adopts van der Linden's (2006) lognormal RT model due to its fit to RT data for different item types in computer-based assessments and integrates it with the DINA model and partial credit AC model through a joint modeling approach. The next section reviews joint models of responses and RT with or without dependencies.

2.4 Joint Modeling of Response and Response Time

Joint modeling of response and RT entails distinct models for response and RT, each with its own set of parameters, with or without dependencies beyond those captured by latent variables and item parameters (De Boeck & Jeon, 2019; van der Linden, 2007, 2009). The prototypical model in this category is van der Linden's (2007) hierarchical framework for modeling speed and accuracy (De Boeck & Jeon, 2019). Two other joint models are the diffusion model (Ratcliff, 1978) and race model (Towsend &Ashby, 1978), which directly model cognitive processes using responses and RTs data but can be re-parameterized as item response models (Tuerlinckx & De Boeck, 2005). This section reviews joint models for response and RT with and without dependencies and extensions of the hierarchical framework to jointly model response and RT for cognitive diagnosis.

2.4.1 Hierarchical Framework for Modeling Speed and Accuracy

van der Linden (2007) proposed a three-level hierarchical framework for the hybrid type of test having items with varying difficulty and requiring varying amount of cognitive processing. The first level of the framework specifies distinct measurement models for response and RT for each combination of person and item. The second level models represent the relations between the parameters in the first-level models. Prior distributions for the second-level parameters or hyperparameters are specified at the third level. Developed as a "plug-and-play" approach, this framework allows for plug-ins of alternative models for the response and RT distributions as well as distributions for their parameters.

As an illustrative example, van der Linden adopted the 3PL normal-ogive model as the response model, and the lognormal model as RT model. The sampling distribution for the response vector and response-time vector $(\mathrm{U}_j, T_j), j = 1, \ldots, J$ for the items $i = 1, \ldots, I$, and the test-takers $j = 1, \ldots, J$, given conditional independence of U_j and T_j, are given as

$$f(U_j, T_j; \xi_j, \Psi) = \prod_{i=1}^{I} f(u_{ij}; \theta_j, a_i, b_i, c_i) f(t_{ij}; \tau_j, \alpha_i, \beta_i) \tag{2.43}$$

where $\xi_j = (\theta_j, \tau_j)$ denotes the vector with parameters for person j and $\psi_i = (a_i, b_i, c_i, \alpha_i, \beta_i)$ denotes the vector of parameters for item i. Second-level models describe the joint distribution of the person parameters in a population, $\mathcal{P}$, and a joint distribution for the item parameters in the domain of items, $\mathcal{I}$, as following a multivariate normal distribution. The values of ξ_j, assumed to be randomly drawn from a multivariate normal distribution, is

$$\xi_j \sim f(\xi_j; \mu_{\mathcal{P}}, \Sigma_{\mathcal{P}}) \tag{2.44}$$

where the density function is

$$f(\xi_j; \mu_{\mathcal{P}}, \Sigma_{\mathcal{P}}) = \frac{|\Sigma_{\mathcal{P}}^{-1}|^{\frac{1}{2}}}{2\pi} \exp\left[-\frac{1}{2}(\xi_j - \mu_{\mathcal{P}})^T \Sigma_{\mathcal{P}}^{-1}(\xi_j - \mu_{\mathcal{P}})\right] \tag{2.45}$$

with mean vector

$$\mu_{\mathcal{P}} = (\mu_\theta . \mu_\tau) \tag{2.26}$$

and covariance matrix

$$\Sigma_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix} \tag{2.47}$$

Likewise, parameter vector ψ_i follows a multivariate normal distribution

$$\psi_i \sim f(\psi_i; \mu_\jmath, \Sigma_\jmath) \tag{2.48}$$

with density function

$$f(\psi_i\,; \mu_\jmath, \Sigma_\jmath) = \frac{|\Sigma_\jmath^{-1}|^{\frac{1}{2}}}{(2\pi)^{\frac{5}{2}}} \exp\left[-\frac{1}{2}(\psi_i\ - \mu_\jmath)^T \Sigma_\jmath^{-1}(\psi_i\ - \mu_\jmath)\right] \tag{2.49}$$

mean vector

$$\mu_\jmath = \left(\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta\right) \tag{2.50}$$

and covariance matrix

$$\Sigma_\jmath = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{ca} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta\alpha} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix} \tag{2.51}$$

The sampling distribution for the full model is

$$f(\mathrm{u}, \mathrm{t}; \xi, \Psi) = \prod_{j=1}^{J} \prod_{i=1}^{I} f\left(\mathrm{u}_j, \mathrm{t}_j; \xi_j, \Psi_i\right) f(\,\xi_j; \mu_\mathcal{P}, \Sigma_\mathcal{P}) f(\psi_i; \mu_\jmath, \Sigma_\jmath) \tag{2.52}$$

Prior distributions for the population and item-domain models are specified as normal/inverse-Wishart distributions.

2.4.2 Diffusion Model and Race Model

The diffusion model directly models the cognitive processes involved in simple single-stage two-choice decisions as encompassing an encoding process with duration u, the decision process with duration d, and a response output process with duration w (Ratcliff, 1987; Ratcliff & McKoon, 2008). This model assumes a noisy evidence accumulation process with a starting point z and culminating in one of two response criteria or boundaries labeled a and 0, at which point a response is initiated.

This process is analogous to a random walk process between two boundaries. The rate of evidence accumulation is the drift rate (v). Nondecision components are u and w combined with mean duration T_{er}. In this model, the probability of a nonmatch (error rate) is given by

$$\gamma - (\xi) = \frac{e^{-\left(\frac{2\xi a}{s^2}\right)} - e^{-\left(\frac{2\xi z}{s^2}\right)}}{e^{-\left(\frac{2\xi a}{s^2}\right)} - 1} \tag{2.53}$$

where ξ denotes relatedness and is set equal to drift, and s^2 is the variance in the drift. The finishing time density function for a nonmatch is given by

$$g - (t, \xi) = \frac{\pi s^2}{a^2} e^{-\left(\frac{z\xi}{s^2}\right)} \sum_{k=1}^{\infty} ksin\left(\frac{\pi z k}{a}\right) e^{-\frac{1}{2}\left(\frac{\xi^2}{s^2} + \frac{\pi^2 k^2 s^2}{a^2}\right)t} \tag{2.54}$$

Setting $\xi = -\xi$ and $z = a - z$ results in equivalent expressions for a match.

The Q-diffusion model is a modification of the diffusions model and is considered an alternate to the hierarchical model (De Boeck & Jeon, 2019). This model decomposes the drift rate and boundary separation into a person and item part denoted by v^p, v^i, a^p, and a^i respectively and combines them into the diffusion parameters (van der Maas et al, 2011). The quotient function is applied to both v and a, such that $v = \frac{v^p}{v^i}$ and $a = a^p / a^i$. By extending Bock's (1972) nominal response model, a general framework for modeling item responses for simple abilities, assuming that a diffusion process produces the item responses, is given by

$$P_+ = \frac{e^{\frac{a_k^p v_k^p}{a_j^i v_j^i} - In(M_j - 1)}}{1 + e^{\frac{a_k^p v_k^p}{a_j^i v_j^i} - In(M_j - 1)}} \tag{2.55}$$

where the 2PL parameters are set to $\alpha^* = \alpha$ and $\beta^* = \beta$. As the two major latent variables in this model are cognitive efficiency, the drift rate, and cautiousness, the boundary separation, dimensions in the Q-diffusion model are considered a rotation of the ability and speed dimensions in the hierarchical model (De Boeck & Jeon, 2019).

The race models are another type of cognitive process model of response choice and RT data based upon which inferences are drawn regarding the processes and mental representations underlying information-processing. An example of this is the lognormal race model proposed by Rounder, Province, Morey, Gomez, & Heathkote (2015). This model assumes an accumulator for each response option with a boundary. The response choice and RT are determined by the first accumulator that reaches its boundary. By modeling cognitive process as a race between accumulators, this model applies to any number of response choices and accommodates other models such as IRT and cell-mean models in its accumulation rates. This model is given by

$$x_j = m \Leftrightarrow y_{mj} = \min_i(y_{ij}) \tag{2.56}$$

where x_j denotes response choice for the ith trial with $x_j = 1, \dots, n$; and $j = 1, \dots, J$. And y_{ij} denotes the finishing time for the ith accumulator on the jth trial. RT t_j is

$$t_j = \psi + \min_i(y_{ij}) \tag{2.57}$$

where ψ is the shift parameter that reflects the contribution of nondecision processes such as encoding and response execution. Each finishing time y_{ij} is modeled as log normally distributed

$$y_{ij} \sim Lognormal(\mu_i, \sigma_i^2) \tag{2.58}$$

The joint density function of choice m at time t is

$$f(m,t) = g(t - \psi;\ \mu_m, \sigma_m^2) \prod_{i \neq m} \left(1 - G(t - \psi; \mu_i, \sigma_i^2)\right) \qquad (2.59)$$

where g and G are the density and cumulative distribution functions of the two-parameters lognormal distribution. Another example is Ranger, Kuhn, Gaviria's (2015) race model for response and RT that specifies two increasing stochastic processes for representing information accumulation associated with one of the two response options. The two latent variables in this model represent information accumulation for producing the correct response and misinformation accumulation for generating the incorrect response. De Boeck and Jeon (2019) describe race models with speed and ability as latent variables as they can be re-parameterized as effective speed and effective ability. Both diffusion models and race models are analogous to the hierarchical model in working with the same two-dimensional space represented by the latent variables.

2.4.3 Local Dependency Models

Three different assumptions of conditional independence associated with modeling the relationship between response and RT in the hierarchical framework are independence between responses to difference items given ability, independence between RTs on different items given speed, and independence between response and RT on the same item given speed and ability (van der Linden, 2009; van der Linden & Glas, 2010). An additional assumption is constancy of speed and proficiency during the test (van der Linden & Glas, 2010). Fluctuation of speed and ability during test administration results in violations of local dependence, henceforth alternative

modeling approaches to accounting for conditional dependencies if occurring in a large and systematic fashion.

Local dependency joint models capture dependencies beyond higher-level correlation between overall speed and ability by including local dependency parameters or through mixture modeling of different classes of responses (De Boek & Jeon, 2019). van der Linden and Glas (2010), for example, include an additional parameter representing the shift in the expected response on item i as caused by a correct response to item k by the same test taker as the alternative model for responses, and an extra parameter denoting the correlation between the log-times on items i and k by the same test taker in the RT model. De Boeck, Chen, & Davison (2017) specify general dependency and item-specific dependency of accuracy on RT in the CAF as

$$\eta_{pi} = \theta_p + \beta_i + \omega \log RT_{pi} \tag{2.60}$$

and

$$\eta_{pi} = \theta_p + \beta_i + \omega_i \log RT_{pi} \tag{2.61}$$

where ω and ω_i are general and item-specific time-dependency parameters. Bolsinova, De Boeck, and Leuven (2017) model conditional dependence between response and RT by incorporating the effects of the residual RT on the intercept and slope parameter of the 2PL model for response accuracy.

The alternative approach to modeling local dependencies between RT and accuracy is mixture modeling of two classes of response as determined by RT (De Boek & Jeon, 2019). Different response processes may be used when subjects respond to different sets of items, giving rise to within-subject heterogeneity of the

item characteristics across the RT. Molenaar and De Boeck (2018) proposed a response mixture model for response and RT that specifies two item-specific latent classes underlying the responses of each item. In this model, class membership is regressed on RT and classes differ in their item characteristics. A mixture of two measurement models each with their distinct set of item discrimination and difficulty parameters is formulated for response probabilities given class membership, respectively as

$$logit\left[P\left(X_{pi}=1\middle|\theta_p,\alpha_{0i},\beta_{0i}\right)\right]=\alpha_{0i}\theta_p-\beta_{0i} \tag{2.62}$$

and

$$logit\left[P\left(X_{pi}=1\middle|\theta_p,\alpha_{1i},\beta_{1i}\right)\right]=\alpha_{1i}\theta_p-\beta_{1i} \tag{2.63}$$

for response probabilities given class membership $C_{pi}=0$ and $C_{pi}=1$. Class membership is then repressed on the subject and item-corrected log-RT to determine whether faster or slower RT are indicative of distinct class membership. Wang and Xu (2015) proposed a mixture hierarchical model for RT and response accuracy to account for differences in responses and RT associated with two test-taking behaviors: rapid guessing and solution behavior.

2.4.4 Joint Modeling of Responses and Response Times for Cognitive Diagnosis

Joint modeling of response and RT has been extended to incorporate cognitive diagnostic models as an alternative response accuracy model (Jiao et al, 2019; Zhan et al 2018a), and to joint testlet models that accommodate local dependencies (Zhan et al, 2018b). Based upon the hierarchical framework for modeling responses and RTs, Zhan et al (2018b) proposed a joint RTs DINA (JRT-DINA) model with the DINA

model as the response accuracy model and the lognormal model as the RT model. The first-level models are the lognormal RT model (van der Linden, 2006) expressed as

$$Log(T_{ni}) = \zeta_i - \tau_n + \varepsilon_{ni}, \qquad \varepsilon_{ni} \sim N\left(0, \sigma_{\varepsilon_i}^2\right) \tag{2.64}$$

and the reparameterized DINA model given by

$$logit\left(P(Y_{ni} = 1)\right) = \beta_i + \delta_i \prod_{k=1}^{K} \alpha_{nk}^{q_{ik}} \tag{2.65}$$

where β_i is the item intercept parameter and δ_i is the interaction parameter and

$$\beta_i = logit(g_i)$$

$$\delta_i = logit(1 - s_i) - logit(g_i). \tag{2.66}$$

The second level models specify the item parameters of the JRT-DINA model as following a trivariate normal distribution and the person parameters as following a bivariate normal distribution, respectively given by

$$\Psi_i = \begin{pmatrix} \beta_i \\ \delta_i \\ \zeta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\zeta \end{pmatrix}, \Sigma_{item}\right) \tag{2.67}$$

and

$$\Theta_n = \begin{pmatrix} \theta \\ \tau \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \Sigma_{person}\right), \Sigma_{person} = \begin{pmatrix} \sigma_\theta^2 & \rho_{\theta\tau}\sigma_\theta\sigma_\tau \\ \rho_{\theta\tau}\sigma_\theta\sigma_\tau & \sigma_\tau^2 \end{pmatrix} \tag{2.68}$$

Bayesian estimation using MCMC indicates that joint modeling of response and RT in the DINA model would improve parameter estimation and classification accuracy rates for attributes and attribute profiles.

Zhan et al (2018b) extends the JRT-DINA model to account for local item dependency and item time dependency in joint modeling of response and RT for cognitive diagnosis. In this model, testlet effects as caused by the same stimulus are

defined as paired local dependence, i.e., dependency in RT, and dependency in response accuracy. Based upon the hierarchical framework for modeling RT and response accuracy (van der Linden, 2006), the first-level models are testlet-DINA model with an additional testlet parameter to represent the interaction effect between person and items on response accuracy, and a lognormal testlet model with an additional testlet parameter to denote the local RT dependency. They are given by

$$logit\big(P(Y_{ni}=1)\big)=\beta_i+\delta_i\prod_{k=1}^{K}\alpha_{nk}^{q_{ik}}+\sum_{m=1}^{M}\mu_{im}\gamma_{nm} \tag{2.69}$$

and

$$LogT_{ni}\sim N(\zeta_i-\tau_n-\sum_{m-1}^{M}\mu_{im}\lambda_{nm},\omega_i^{-2}) \tag{2.70}$$

where μ_{im} indicates whether or not item i is part of testlet m, and γ_{nm} and λ_{nm} denote the effect of testlet m on response accuracy and RTs respectively. At the second level, the testlet effect parameters are assumed to follow a bivariate normal distribution

$$\Gamma_{nm}=\binom{\gamma_{nm}}{\lambda_{nm}}\sim N\left(\binom{0}{0},\Sigma_{testlet,m}\right) \tag{2.71}$$

where the same μ_{im} value is assumed in the response and RT models. Specification of the item and person parameters at this level are identical compared to the JRT-DINA model.

2.4.5 Summary and Discussion

This section reviews joint models of response and RT, the hierarchical framework for modeling speed and ability, the diffusion model, the race model, local dependency models, and joint models of response and RT for cognitive diagnosis.

The hierarchical framework for modeling speed and ability is a complex two-dimensional measurement model, with one dimension for response accuracy, and another dimension for RT (De Boeck & Jeon, 2019). The first-level models specify distributions for response and RT, while the second level specify multivariate distributions for the model parameters, allowing for estimation of the relation between speed and ability at the population level. Applied to educational and psychological testing, this framework simultaneously estimates IRT parameters and other parameters in this framework, which may lead to increased accuracy of the estimated parameters (van der Linden, 2007; van der Linden, Klein Entink, & Fox, 2010). RT modeling in this framework can be used to improve the design of adaptive testing, handle the issue of speededness in testing (van der Linden, Breithaupt, Chuah, & Zhang, 2007), and detect aberrant test behaviors (van der Linden & Guo, 2008).

The diffusion model and race models, on the other hand, are finer-grained process models the parameterization of which maps onto elements of the cognitive processes contributing to decision making. The diffusion model assumes information accumulation between boundaries, while the race models assume a race among different accumulators (De Boeck & Jeon, 2019). Both types of models account for speed-accuracy trade-off and represent it as a complicated function of the model parameters. In addition, both types of models are amenable to latent variable modeling, with the dimensions in the models corresponding to the ability and speed dimensions of the hierarchical model (De Boeck & Jeon, 2019; Ranger et al, 2014; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011).

Local dependency models capture extra dependency of the response and RT, beyond the relationship of their latent variables and item parameters. By specifying a general and item-specific dependency of accuracy on RT, De Boeck et al (2017) identified two kinds of speed effects: a speed-accuracy trade-off induced by imposed speed, and an opposite CAF effect associated with speed within conditions, which may occur as a result of within-person variation of the cognitive capacity. Bolsinova et al's (2017) investigation of the residual dependence between RT and accuracy identified the dependence of item properties on the speed of responses: SAT holds for more difficult items, whereas for easier items, an opposite SAT exists, with slower responses associated with a lower probability of a correct response. Slower responses are also less informative for ability as their discrimination parameters decrease with residual RT. These findings are in line with Molenaar and De Boeck's (2018) who found a similar effect, indicating local dependence of the response accuracy and RT conditioned on item properties. Using a hierarchical mixture modeling approach, Wang and Xu (2015) identify differences in RT patterns and responses attributable to two test-taking behaviors: problem-solving and rapid guessing and demonstrate that the model yields more accurate item and person parameter estimates than a non-mixture model.

The JRT-DINA model (Jiao et al, 2019; Zhan et al, 2018a) and its extension to account for dual local response and RT dependence (Zhan et al, 2018b) are joint models of response and RT for cognitive diagnostic models built upon the hierarchical framework for modeling speed and accuracy (van der Linden, 2007). By incorporating RT modeling in the modeling of responses, the JRT-DINA model

yields improved attribute and profile correct classification rates and more accurate and precise estimation of the model parameters (Zhan et al, 2018a). This finding is in line with van der Linden et al (2006), suggesting that simultaneous estimation of the DINA model parameters and other parameters in the joint model can lead to increased accuracy of the estimated parameters. As noted by Jiao et al (2019), under certain circumstances when a test is not adequately designed for cognitive diagnosis, as indicated by Q-matrix not properly verified, inadequate test lengths, or poor item quality, the effect of incorporating RT in the modeling of responses on parameter estimation is evident.

RT is one of the most widely studied response process data in psychometric modeling. Another type of response process data is ACs. This next section review AC patterns and outcomes, indices for detecting aberrant ACs, and AC modeling.

2.5 Answer Change Modeling

Answer changes (ACs), synonymous to erasures (Sinharary, 2018), response changes (Liu, Bridgeman, Gu, Xu, & Kong, 2015), or response revisions, refer to the fact that test-takers, after making an initial decision, subsequently revisit the decision and revert to an alternative option as their best choice (Jeon, De Boeck, & van der Linden, 2017; Malia, 2007). Distinctions are drawn between benign ACs and fraudulent ACs, with the former describing non-aberrant ACs and the latter suggesting test tampering by test-takers, test administrators, or educators (Sinharay, Duong, & Wood, 2017; Sinharay & Johnson, 2016). Sinharay and Johnson (2016) further distinguishes between two types of benign ACs: string-end ACs and random ACs. String-end ACs occur when examinees randomly guess on the remaining test

items due to time constraints, but subsequently revise some of the answers when additional time becomes available. Random ACs refer to the types of ACs that examinees make upon reconsidering the choices they initially make by accident (Wollack, Cohen, & Serlin, 2015). This section reviews research on AC patterns, aberrant ACs, and psychometric models of ACs and AC patterns.

2.5.1 Patterns and Outcomes of Answer Changes (ACs)

ACs research dating back to the 1920s seeks answers to issues including outcomes of ACs, the relationship between ability and ACs, and factors affecting test-takers' AC behaviors (e.g., Bridgeman, 2012; Kruger et al, 2015; Liu et al, 2015; Malia, 2007; Jeon et al, 2017). Results of ACs studies conducted at the aggregate level consistently suggest that, contrary to the belief that students should trust their first instinct and initial response choices are more accurate than subsequent responses, ACs are likely to result in score gains and improved test performance (e.g. Bridgeman, 2012; Liu et al, 2015). A majority of the students change answers during a test (Al-Hamly & Coombe, 2005; Liu et al, 2015). Although only a small portion of the test items are typically changed during a test, the majority of the students benefit from changing their answers (Al-Hamly & Coombe, 2005; Liu et al, 2015; Milia, 2007). Limited research on computer-adaptive tests (CAT) yields similar findings, i.e., although ACs occur for only a small portion of the test items, more test-takers change their answers than those who do not, and those who do usually gain scores from changing their answers (Liu et al., 2015).

The extent to which test-takers gain scores from ACs is moderated by test-takers' ability level and depends on the nature of responses test-takers change during

the test (Al-Hamly & Coombe, 2005; Jeon et al, 2017; Liu et al, 2015; McMorris, DeMers, & Schwarz, 1987; Milia, 2007). High-performing examinees were more likely to make wrong-to-right changes and fewer right-to-wrong changes (Al-Hamly & Coombe, 2005; Milia, 2007). High-performing examinees tend to gain more from ACs than low-performing examinees, as indicated by significantly higher gain-to-loss ratios by the high-performing group (Liu et al, 2015). Score gains are minimal for the low-ability test takers, as compared to minor to moderate gains for test-takers with moderate to high ability levels (Jeon et al, 2017; McMorris, et al, 1987). In addition, the nature of the responses that test-takers change affects the effect of ACs. Score gains are more likely for responses that test-takers initially and mistakenly make due to carelessness or time constraints and are able to correct subsequently, but less likely for responses resulting from misconception or confusion over alternative options on multiple-choice items (Higham & Gerrard, 2005; Liu et al, 2015).

2.5.2 Indices for Detecting Aberrant Answer Changes (ACs)

Considerable research focuses on the analysis of ACs or erasure analysis to derive indices for detecting fraudulent erasures and test tampering (e.g. Belov, 2011, 2015, 2017; Sinhary, 2018; Sinharay, Duong, & Wood, 2017; Sinharay & Johnson, 2016; Wollack, Cohen, & Eckerly, 2015). Belov (2015) for example proposed the D-index based on the Kullback-Leibler divergence (KLD; Kullback & Leibler, 1951) measure of the difference between posteriors of ability computed from responses to two subsets: one subset with ACs and one subset without ACs to detect aberrant ACs. Wallack et al (2015) suggested the erasure detection index (EDI) based on Bock's (1972) nominal response model for fraud detection at the individual level. Wallack

and Eckerly (2017) extended the EDI to detection of fraudulent erasures at the group level. Sinharay et al (2017) proposed the L-index based on the likelihood ratio test (LRT; e.g., Cox & Hinkley, 1974) of the equality of the model parameters underlying the subset of items with ACs and the subset without ACs for detection of aberrant ACs. Simulation studies of the performance of these indices demonstrate their robustness and/or usefulness for fraudulent erasures for dichotomous items.

2.5.3 Models of Answer Changes (ACs)

ACs models represent AC behaviors as a sequence consisting of an initial stage where test-takers give initial responses to test items and a final stage in which they either confirm or replace their initial responses (Jeon, De Boeck, & van der Linden, 2017; van der Linden & Jeon, 2012). van der Linden and Jeon (2012) proposed an IRT-based approach to model the probability of test-takers changing answers upon reviewing their initial choices for multiple-choice paper-and-pencil tests, based on the assumption that test-takers are allowed enough time to respond to all items on the test and review their answers upon completing a first pass. This model distinguishes three types of erasure patterns: a RW erasure that occurs when the initial correct response is replaced by an incorrect response, a WW erasure which replaces an initial incorrect response with another incorrect response, and a WR erasure which replaces the initial incorrect response with a correct response. In this model, the final stage responses, conditional on the initial responses, are given by

$$\Pr\left\{U_{ni}^{(2)} = 1 \middle| U_{ni}^{(1)} = 1\right\} = \frac{\exp\left[a_{1i}\left(\theta_n^{(1)} - b_{1i}\right)\right]}{1 + \exp\left[a_{1i}\left(\theta_n^{(1)} - b_{1i}\right)\right]} \tag{2.72}$$

the complement of which is the probability of a RW erasure, and

$$\Pr\left\{U_{ni}^{(2)} = 1 \middle| U_{ni}^{(1)} = 0\right\} = \frac{\exp\left[a_{0i}\left(\theta_n^{(1)} - b_{0i}\right)\right]}{1 + \exp\left[a_{0i}\left(\theta_n^{(1)} - b_{0i}\right)\right]} \tag{2.73}$$

which is the probability of a WR erasure and the complement of which is a compound event of a WW erasure or a test-taker confirming an incorrect response given at the initial stage. In this model the item parameters a_i and b_i are free but the final stage ability parameters are set to equal to their initial values such that $\theta_n^{(2)} = \theta_n^{(1)}$.

Jeon et al (2017) adopted a similar approach in an application of the generalized IRT tree model to model AC behaviors. The leaves in the IRT tree in this model represents four possible outcomes of AC behavior: WW (both initial and final responses are wrong), WR (the initial response is wrong and the final response is right), RW (the initial response is right and the final response is wrong), and RR (both the initial and final responses are right). Three nodes represent three latent abilities contributing to the four AC patterns, with the node at the top defining the ability to correctly respond to an item when initially reviewed, and the two nodes in the middle representing two different abilities, the ability to make a correct change when the initial response is wrong and propensity to make no change when the initial response is right. Node-specific response probabilities denoted by $Y_{pi}^{(1)}$, $Y_{pi}^{(2)}$, $Y_{pi}^{(3)}$ are given by the 2PL IRT model specifying them as a function of three distinct sets of ability and item parameters. The following T matrix shows how the outcomes denoted as Z_{pi} relate to the node-specific response probabilities

$$\begin{array}{c} \\ Z_{pi}=1\ (WW) \\ Z_{pi}=2\ (WR) \\ Z_{pi}=3\ (RW) \\ Z_{pi}=4\ (RR) \end{array} \begin{array}{c} \begin{array}{ccc} Y_{pi}^{(1)} & Y_{pi}^{(2)} & Y_{pi}^{(3)} \end{array} \\ \begin{bmatrix} 0 & 0 & NA \\ 0 & 1 & NA \\ 1 & NA & 0 \\ 1 & NA & 1 \end{bmatrix} \end{array} \tag{2.74}$$

The probabilities of the observed outcomes are then computed as the product of the node-specific probabilities as follows:

$$\Pr\left(Z_{pi}=1\middle|\theta_p\right)=\Pr\left(Y_{pi}^{(1)}=0|\theta_p^{(1)}\right)\Pr\left(Y_{pi}^{(2)}=0|\theta_p^{(2)},Y_{pi}^{(1)}=0\right)$$

$$\Pr\left(Z_{pi}=2\middle|\theta_p\right)=\Pr\left(Y_{pi}^{(1)}=0|\theta_p^{(1)}\right)\Pr\left(Y_{pi}^{(2)}=1|\theta_p^{(2)},Y_{pi}^{(1)}=0\right)$$

$$\Pr\left(Z_{pi}=3\middle|\theta_p\right)=\Pr\left(Y_{pi}^{(1)}=1|\theta_p^{(1)}\right)\Pr\left(Y_{pi}^{(3)}=0|\theta_p^{(3)},Y_{pi}^{(1)}=0\right)$$

$$\Pr\left(Z_{pi}=4\middle|\theta_p\right)=\Pr\left(Y_{pi}^{(1)}=1\middle|\theta_p^{(1)}\right)\Pr\left(Y_{pi}^{(3)}=1\middle|\theta_p^{(3)},Y_{pi}^{(1)}=1\right) \tag{2.75}$$

Simpler models can be specified by constraining the item and person parameters to be the same across the three nodes.

2.5.4 Summary and Discussion

This section reviews AC patterns and outcomes, indices for detecting aberrant ACs, and AC models. As described in this section, the majority of the students change their answers during a test and the majority of the students gain scores from changing them (Al-Hamly & Coombe, 2005; Liu et al, 2015; Milia, 2007). Further, AC outcomes are associated with test-takers' ability level, with high-performing students gaining from making more wrong to right changes (Al-Hamly & Coombe, 2005; Milia, 2007). Indices for detecting aberrant ACs are the D-index based on the KLD, the EDI based on Bock's (1972) nominal response model, and the L-index based on the LRT. IRT-based approach and IRT tree model are used to model the

probability of AC patterns: WW, WR, RW, and RR. AC patterns and outcomes are closely associated with test-takers' ability level. In addition, AC outcomes directly contribute to changes in response patterns. ACs as a response process data can provide more information for the estimation of person ability and the assessment of students' mastery status on the attributes of interest. Thus incorporating ACs as a process data in the joint model of responses and RT can provide more information about the estimation of students' ability level, resulting in improved attribute and attribute profile classification accuracy and more accurate and precise estimation of the ability parameter.

2.6 Model Estimation

2.6.1 Bayesian Inference

Bayesian inferences regarding a parameter θ are drawn in terms of probability statements that are conditional on the observed value of y denoted as $p(\theta|y)$ (Almond, Mislevy, Steinberg, Yan, &Williamson, 2015; Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2014; Levy & Mislevy, 2016). The level of conditioning on observed data distinguishes Bayesian inferences from the alternative approach to statistical inference which retrospectively evaluates the procedure used to estimate θ over the distribution of possible y values conditional on the true unknown value of θ (Gelman et al, 2014). Applying the basic property of conditional probability, defined as Bayes' rule, the posterior density is expressed as

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{2.76}$$

where the joint probability mass or density function $p(\theta, y)$ is written as the product of the prior distribution $p(\theta)$ and the sampling distribution $p(y|\theta)$, and $p(y) = \sum_\theta p(\theta)p(y|\theta)$ is summed over all possible values of θ. In response data modeling, the posterior density of the parameter θ translates into probability beliefs about the parameters based on prior and response data information (Fox, 2010; Levy & Mislevy, 2016). As the factor $p(y)$ does not depend on θ and can be considered a constant, the unnormalized posterior density omitting the factor $p(y)$ can be written as

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{2.77}$$

As such, applications of Bayesian inferences are primarily concerned with modeling $p(\theta, y)$ and summarizing $p(\theta|y)$ in appropriate ways (Gelman et al, 2014).

2.6.2 Markov Chain Monte Carlo

Markov chain simulation, also called Markov Chain Monte Carlo (MCMC), is a method that draws values of θ from approximate distributions and then corrects them to better approximate the target posterior distribution, $p(\theta|y)$ (Gelman et al, 2014). In this method, sampling is done sequentially, and the resulting distribution of the sampled draws depends only on the value last drawn, hence forming a Markov chain. Markov chain simulation is used when sampling θ directly from $p(\theta|y)$ is not possible, as in many hierarchical models where marginal posteriors are intractable and the dimensionality of the problem results in difficulties in sampling θ directly from $p(\theta|y)$ (Fox, 2010; Gelman et al, 2014).

One of the Markov chain algorithms often used in many multidimensional problems is the Gibbs sampler (Fox, 2010; Gelman et al, 2014). The Gibbs sampler

partitions the parameter vector θ into d subvectors, $\theta = (\theta_1, \ldots, \theta_d)$. Each iteration of the Gibbs sampler consists of d steps cycling though the d subvectors and drawing each subset conditional on the values of all the others (Gelman et al, 2014). At each iteration t, the d subvectors are ordered and each subvector is updated conditional on the latest values of the other subvectors of θ, resulting in the iteration t values for the subvectors already updated and the iteration $t-1$ values for the others.

The Metropolis-Hastings algorithm refers to a family of Markov chain simulation methods for sampling from Bayesian posterior distributions and is a generalization of the basic Metropolis algorithm (Gelman et al, 2014; Levy & Mislevy, 2016). The Metropolis algorithm, defined as an adaptation of a random walk, computes acceptance/rejection probabilities for mixing a proposal distribution and a jumping distribution and cycling through the process until convergence to the target distribution is reached (Gelman et al, 2014). As described by Gelman et al (2014; see also Levy & Mislevy, 2016), the algorithm consists of an initial draw of a starting point θ^0 from a starting distribution $p_0(\theta)$ and subsequently sampling a proposal θ^* from a symmetric jumping distribution or proposal distribution at time t, $J_t(\theta^*|\theta^{t-1})$, at which point the ratio of densities

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)} \tag{2.78}$$

is computed. Specifying the acceptance/rejection rule as

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(\mathrm{r}, 1) \\ \theta^{t-1} & \text{otherwise} \end{cases} \tag{2.79}$$

the algorithm generates the transition distribution $T_t(\theta^t|\theta^{t-1})$ as a weighted jumping distribution $J_t(\theta^t|\theta^{t-1})$ that adjusts for the acceptance rate.

The Metropolis-Hastings algorithm generalizes the Metropolis algorithm by allowing asymmetric jumping distributions and correcting for the asymmetry in the jumping rule with a reformulation of the ratio r as a ratio of ratios:

$$r = \frac{\frac{p(\theta^*|y)}{J_t(\theta^*|\theta^{t-1})}}{\frac{p(\theta^{t-1}|y)}{J_t(\theta^{t-1}|\theta^*)}} \tag{2.80}$$

Asymmetric jumping rule in the Metropolis-Hasting algorithm increases the speed of the random walk and improves computational efficiency (Gelman et al, 2014).

2.6.3 Convergence Assessment

Iterative simulations can yield significant underrepresentation of the target distribution if convergence is not reached. Serial correlation, although not necessarily problematic at convergence, can cause computational inefficiencies. Assessment of convergence in iterative simulations involves checking mixing and stationarity by comparing within- and between-sequence variation (Gelman et al, 2014; Levy & Mislevy, 2016). For quantities with normal marginal posterior distributions, Gelman et al (2014) recommend assessing convergence by estimating the factor by which the scale of the distribution for the scalar estimand ψ might be reduced if the simulations were continued to infinity. For simulations ψ_{ij} with m chains each of length n, the potential scale reduction factor is

$$\hat{R} = \sqrt{\frac{\widehat{var}^+(\psi|y)}{W}} \tag{2.81}$$

where W is the within-sequence variance, and $\widehat{var}^+(\psi|y)$ is a weighted average of the within-sequence variance W and between-sequence variance B given by

$$\widehat{var}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B \tag{2.82}$$

The potential scale reduction factor decreases to 1 as n approaches infinity. If $\hat{R}$ is close to 1, the inference that iterations approximate the target distribution is justified. For extreme quantiles or for parameters with multimodal posterior distributions, Gelman et al (2014) recommend also monitoring the extreme quantiles of the between and within sequences.

2.7 Summary of Literature Review

This chapter is a comprehensive review of current approaches to cognitive diagnostic modeling and the modeling of testlet effects, RT, and ACs, and frameworks for modeling speed and accuracy, which provides the context and theoretical background for the joint model of responses, RT, and ACs in testlet-based assessments for cognitive diagnosis proposed in this research. This chapter additionally introduces the model estimation method to be used in this research: fundamentals of Bayesian inference, the Monta Carlo simulation method, and diagnostics for assessing model convergence. Chapter 3 proposes the joint model, describing its formulation and parameterization, the design of a simulation study to investigate the impact of the manipulated factors on model performance and parameter estimates, and a description of an empirical study to evaluate and validate the proposed model.

Chapter 3: Methods

This chapter proposes a joint model of responses, RTs, and ACs in testlet-based assessments for cognitive diagnosis, the components of which are drawn from the modeling approaches and methodologies reviewed in Chapter 2. Section 3.1 presents the overall framework for modeling the responses, RT and AC patterns in testlet-based cognitive diagnostic assessment, followed by a description of model specification and parameterization for each of the componential measurement models in this framework. Section 3.2 specifies the prior distributions of the model parameters and hyperparameters discussed in Section 3.1 and methods for estimating them from the Bayesian inference perspective. Section 3.3 delineates the design of the simulation studies and introduces the fixed versus manipulated factors and criteria for evaluating model fit and parameter recovery. This chapter concludes with a description of the empirical data and analytic procedures employed to evaluate and validate the performance of the proposed model.

3.1 The Proposed Model

The research study proposed in this chapter adopts and extends van der Linden's (2007) hierarchical framework for the modeling and analysis of response accuracy, RT, and an additional source of information about test-takers, ACs, in testlet-based assessment for cognitive diagnosis. As reviewed in Chapter 2, the hierarchical framework is the prototypical model in the joint modeling of response

and RT for tests typically administered in computer-based testing programs (De Boeck & Jeon, 2019) and has been used extensively in joint models of response and RT (e.g., Klein Entink, Fox, & van der Linden, 2009; Loeys et al, 2011; Zhan et al, 2018a; Zhan et al, 2018b). By accounting for dependencies between the item and person parameters in a higher-level structure, the hierarchical framework is flexible in allowing alternative models for the distributions of responses, RT, and their parameters and the modeling of the relationship between speed and accuracy and between the time and response parameters of the items at the population level (van der Linden, 2007).

The first level models in the proposed joint model are distinct models for cognitive diagnosis, RT, and ACs: the DINA model (Junker & Sijtsma, 2001; Macready & Dayton, 1977), the lognormal RT model (van der Linden, 2006), and partial credit model for ACs (Masters, 1982). Following Zhan et al (2018b), testlet parameters are incorporated in the response and RT model to specifically address the testlet effects. The following subsections describes in detail formulation of the model and specification of the model parameters for the three first-level models and for the second level models that capture the relations between person parameters, item parameters, and testlet parameters.

3.1.1 Higher-Order Latent Trait DINA Model for Testlet-Based Assessment

The first-level response model is the higher-order latent trait DINA model (de la Torre & Douglas, 2004; Junker & Sijtsma, 2001; Macready & Dayton, 1977). As reviewed in Chapter 2, the DINA model is a non-compensatory parsimonious cognitive diagnostic model that specifies response probabilities as a function of two

parameters: a slipping parameter and a guessing parameter, for each item (e.g., Junker & Sijtsma, 2001; Macready & Dayton, 1977). The latent response variables ξ_{ij} is a binary function of binary inputs and functions as the "and" gate component combining deterministic input $\alpha_{ik}^{Q_{jk}}$ where α_{ik} indicates whether examinee i possesses attribute k and Q_{jk} indicates whether attribute k is required for task or item j. Following Zhan et al (2018a), the IRF for a given item can be reexpressed as

$$P(Y_{ij}=1) = g_j + (1 - s_j - g_j)\prod_{k=1}^{K}\alpha_{ik}^{Q_{jk}} \tag{3.1}$$

and, using the logit scale, reparameterized as

$$logit\left(P(Y_{ij}=1)\right) = \beta_j + \delta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{ik}} \tag{3.2}$$

where

$$\beta_l = logit(g_j) \tag{3.3}$$

and

$$\delta_j = logit(1 - s_j) - logit(g_j) \tag{3.4}$$

This reformulated IRF can be easily extended to incorporate a testlet parameter to account for the contextual effects of testlets on items (Zhan et al, 2018b; see also Im, 2017). Following Zhan et al (2018b), the DINA model for testlet-based assessment is given by

$$logit\left(P(Y_{ij}=1)\right) = \beta_j + \delta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{ik}} + \sum_{d=1}^{D}\chi_{jd}\gamma_{id(j)} \tag{3.5}$$

where χ_{jd} is a 1/0 variable that indicates whether item j is an item nested within testlet d and $\gamma_{id(j)}{\sim}N(0,\sigma^2_{\gamma_{d(j)}})$ denotes the testlet effect of item j to person i nested

within testlet $d(j)$. $\sigma^2_{\gamma_{d(j)}}$ represents the magnitude of the testlet effects and is allowed to vary across testlets (Wang et al., 2002; Wang & Wilson, 2005). All testlet effects $\gamma_{id(j)}$s are assumed to be independent of each other.

Assuming attributes and their acquisition as related to a more-broadly defined latent construct of general intelligence or aptitude denoted as θ, the higher-order latent structural model specifies the probability for attribute α conditional on θ as a logistic regression model with latent covariate θ (de la Torre & Douglas, 2004). Using the logit scale, it can be reformulated as

$$logit\big(P(\alpha_{ik} = 1|\theta_i)\big) = \kappa_k\theta_i - \iota_k \tag{3.6}$$

where κ_k and ι_k are the slope and intercept for attribute k (Zhan et al., 2018a). This implies an estimation of $2K$ parameters, which greatly reduces the complexity of the higher-order structural model. Further, this high-order structure model generates an estimate $\hat{\theta}$ beyond the attribute profiles yielded by classification of α_{ik}.

3.1.2 The Lognormal RT Model for Testlet-based Assessment

As reviewed in Chapter 2, the lognormal RT model is a flexible model for fitting RT data generated by different item types in computer-based tests. It specifies RT distributions for a fixed person as determined by the person speed parameter τ_i, the time intensity of item j denoted as ζ_j, and the discrimination parameter ω_j modifying the relationship between time t_{ij} and its mean (van der Linden, 2006). This model assumes conditional independence of the RT given person speed at the level of a fixed person, i.e., person speed and ability are constant, and once a person's choice of ability and speed level is made, only person speed accounts for the RT

distributions. A second level of modeling captures the dependence between speed and ability at the population level.

In this research the RT model is the lognormal RT testlet model which extends the lognormal RT model to specifically account for local RT dependence (Im, 2017; Zhan et al, 2018b). This model is given by

$$T_{ij}\sim f\left(t_{ij};\ \tau_i,\omega_j,\zeta_j,\lambda_{id(j)}\right)=\frac{\omega_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\omega_j\left(\log t_{ij}-\left(\zeta_j-\tau_i-\sum_{d=1}^{D}\chi_{jd}\lambda_{id(j)}\right)\right)\right]^2\right\} \tag{3.7}$$

which is equivalent to

$$\log T_{ij}\sim N\left(\zeta_j-\tau_i-\sum_{d=1}^{D}\chi_{jd}\lambda_{id(j)},\omega_j^2\right) \tag{3.8}$$

where $\log t_{ij}$ is the log RT, χ_{jd} indicates testlet membership, and $\lambda_{id(j)}\sim N(0,\sigma_{\lambda_d}^2)$ is the testlet parameter representing the effect for person i on testlet d. The variance of the testlet effect $\sigma_{\lambda_d}^2$ indicates its magnitude. Further, all testlet effects $\lambda_{id(j)}$s are assumed to be independent of each other.

3.1.3 Partial Credit AC Model

In the proposed model, partial credit model is chosen to fit the AC data. The partial credit model is a polytomous item response model that applies Rasch's model for dichotomies to each pair of adjacent categories in an ordered sequence (Master, 1982, 2018; Masters & Wright, 1996). As a Rasch family model, PCM features separable person and item parameters and sufficient statistics, which allows objective comparisons of persons and items (Masters, 2018; Rasch, 1977). Two sets of parameters in the model, one for persons and one for items, represent locations on the

underlying measurement variable. PCM is applied to tests using items with two or more ordered response categories and is easy to implement in practice due to simplicity of its formulation (Masters, 2018).

The reviewed studies distinguish four AC behaviors based on a comparison of the initial and final response: WW (both the initial and final responses are wrong), WR (the initial response is wrong and the final response is right), RW (the initial response is right and the final response is wrong), and RR (both the initial and final response are right) (Jeon at al., 2017; van der Linden & Jeon, 2012). ACs studies reviewed in Chapter 2 additionally associate AC outcomes with test-takers' ability level: high-performing are more likely to make WR ACs and fewer RW ACs and they benefit more from ACs compared to low-performing examinees (e.g., Jeon et al, 2017, Liu et al, 2015; Milia 2007). By adopting the PCM as the AC model, the proposed research assumes that AC patterns follow a categorical sequence ordered as WW, RW, WR, and RR, with RR indicating a final confirmation of an initially right answer and WW suggesting either a change from one wrong answer to another or a final confirmation of an initially wrong answer. This model further assumes that the probability of a given AC pattern is a function of test-takers' latent ability and item step difficulty (Jiao et al, 2020).

Assuming the response categories followed an intended order $0 < 1 <, \dots, < A_j$, in the PCM the conditional probability of scoring a a rather than a $a-1$ using Rasch's model of dichotomies is given by

$$\frac{P\{U_{ij}=a\}}{P\{U_{ij}=a-1\}+P\{U_{ij}=a\}}=\frac{\exp(\theta_i-b_{ja})}{1+\exp(\theta_i-b_{ja})} \tag{3.9}$$

This can be re-expressed as the unconditional probability of each possible outcome of person i responding to item j given by

$$P\{U_{ij} = a\} = \frac{\exp \sum_{k=0}^{a} (\theta_i - b_{jk})}{\sum_{h=0}^{A_i} \exp \sum_{k=0}^{h} (\theta_i - b_{jk})} (a = 0, 1, \ldots, A_i) \tag{3.10}$$

where

$$\sum_{k=0}^{0} (\theta_i - b_{jk}) = 0 \text{ and } \sum_{k=0}^{h} (\theta_i - b_{jk}) = \sum_{k=1}^{h} (\theta_i - b_{jk}) \tag{3.11}$$

θ_i is the ability of person i as in the response model, and b_{jk} is the item step parameter for item j getting a score category of a. The higher-order ability parameter θ connects the attributes and AC patterns. The item step parameter b_{jk} is reparametrized into an item location parameter b_j and the threshold parameter b_{ja} for $a - 1$ score categories.

3.1.4 Specification of the Second-level Models

Following van der Linden's (2007) hierarchical framework, subsections 3.1.1 through 3.1.3 present the first-level models. This subsection presents the second-level models specifying the joint distributions of the person, item, and testlet parameters. These models describe the relations between the person, item, and testlet parameters in the first level models for response, RT, and AC patterns. In the first model, the person parameters are assumed to follow a bivariate normal distribution:

$$\xi_i = \begin{pmatrix} \theta_i \\ \tau_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \Sigma_{\text{person}} \right), \Sigma_{\text{person}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix} \tag{3.12}$$

The second model describes the relations between the item parameters, which are assumed to follow a multivariate normal distribution

$$\Psi_j = \begin{pmatrix} \beta_j \\ \delta_j \\ \zeta_j \\ b_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\zeta \\ \mu_b \end{pmatrix}, \Sigma_{item} \right) \tag{3.13}$$

The residual error variance $\sigma_{\varepsilon_j}^2$ is assumed to be independently distributed and is not included in Ψ_j. The third model captures the relations between testlet parameters in testlet d, which are assumed to follow a bivariate normal distribution

$$\Gamma_{ij(d)} = \begin{pmatrix} \gamma_{id(j)} \\ \lambda_{id(j)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_{testlet,d} \right), \Sigma_{testlet,d} = \begin{pmatrix} \sigma_\gamma^2 & \sigma_{\gamma\lambda} \\ \sigma_{\gamma\lambda} & \sigma_\lambda^2 \end{pmatrix} \tag{3.14}$$

To establish identifiability, the following constraints are set

$$\mu_\theta = 0, \sigma_\theta^2 = 1, \mu_\tau = 0 \tag{3.15}$$

The first two constraints are similar to constraints for the θ parameters in higher-order latent trait models and IRT models and function to identify the scale between θ_i and τ_i and between θ_i and b_j. The third constraint fixes the zero of τ_i and removes the tradeoff between ζ_j and τ_i. Fixing the location of τ_i identifies the scale between τ_i and ζ_j. The proposed model further assumes independence of the attributes given θ_i, independence of the responses given α_i and $\gamma_{id(j)}$, independence of the RTs given τ_i and $\lambda_{id(j)}$, and independence between responses and RTs for a given item given person parameters and testlet parameters.

3.2 Model Parameter Estimation

This research uses the Bayesian approach to estimate the parameters in the join model of response, RT, and ACs in testlet-based assessment for cognitive diagnosis. Parameter estimation will be implemented using the program "Just

Another Gibbs Sampler" Version 4.3.0 (JAGS; Plummer, 2017). The JAGS program interfaces with the R program via the package R2jags (Su & Yajima; 2020).

3.2.1 Specification of the Priors and Hyper Priors

The proposed model assumes conditional independence of response, RT, and ACs. Specification of the priors and hyper priors follows Jiao et al (2020; see also Zhan et al, 2018a; Zhan et al, 2018b). Prior distributions for the attributes, responses, RT, and ACs are specified as

$$\begin{aligned} Y_{ij} &\sim Bernoulli\ (P(Y_{ij}=1) \\ logT_{ij} &\sim N\left(\zeta_j - \tau_i - \sum_{d=1}^{D}\chi_{jd}\lambda_{id(j)}, \omega_j^{-2}\right) \\ \alpha_{ik} &\sim Bernoulli\ \left(P(\alpha_{ik}=1)\right) \\ U_{ij} &\sim Categorical\ \left(P\left(U_{ij}=0,1,2,3\right)\right) \end{aligned} \tag{3.16}$$

In the second-level models, the priors of the person parameters are specified as

$$\begin{pmatrix}\theta_i \\ \tau_i\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_{\text{person}}\right) \tag{3.17}$$

The variance of θ_i is constrained to 1 for identification purposes. Using the Chelosky decomposition, Σ_{person} is reparameterized as $\Delta_{person}\Delta'_{person}$(Zhan et al, 2018a). Δ_{person} is shown as

$$\Delta_{person} = \begin{pmatrix}1 & 0\\ \varphi & \psi\end{pmatrix} \tag{3.18}$$

and Δ'_{person} is its conjugate transpose. The prior elements in the Δ_{person} are set as $\varphi \sim N(0,1)$ and $\psi \sim Gamma(1,1)$.

Priors for the slope and intercept parameter in the higher-order structural model are specified as

$$\kappa_k \sim N(0, 0.25), \iota_k \sim N(0, 0.25) I(\iota_k > 0) \tag{3.19}$$

The priors of the item parameters are assumed to follow a multivariate normal distribution

$$\begin{pmatrix} \beta_j \\ \delta_j \\ \zeta_j \\ b_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\zeta \\ \mu_b \end{pmatrix}, \Sigma_{item} \right), \sigma_{\varepsilon_j}^2 \sim invGamma(1,1) \tag{3.20}$$

Hyper priors for the parameters in this distribution are specified as following normal/inverse-Wishart distributions:

$$\mu_{b_j} \sim N(0, 1)$$

$$\mu_\beta \sim N(-2.197, 2)$$

$$\mu_\delta \sim N(4.394, 2) I(\mu_\delta > 0)$$

$$\mu_\zeta \sim N(3, 2)$$

$$\Sigma_{item} \sim InvWishart(R, 4). \tag{3.21}$$

On a logit scale, μ_β at -2.197 translates into a mean guessing probability of 0.1 and μ_δ indicates a mean slipping probability of 0.1. With a variance 2, the simulated mean guessing probabilities range from 0.026 to 0.314 and the range of the simulated mean slipping probability is from 0.007 to 0.653. On a natural log scale, μ_ζ at 3 with a variance of 2 indicates a mean RT of 20.086 and a range from 4.883 to 82.617 for the simulated RT means. Item step parameters in the ACs model are assumed to follow a normal distribution:

$$b_{j1} \sim N(0,1), b_{j2} \sim N(0,1), b_{j3} = -sum(b_{j1}, b_{j2}) \quad (3.22)$$

The priors of the testlet parameters for a given testlet are specified as

$$\begin{pmatrix} \gamma_{id(j)} \\ \lambda_{id(j)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right), \Sigma_{testlet,d} \quad (3.23)$$

where

$$\Sigma_{testlet,d} \sim InvWishart\left(R_{testlet,d}, 2\right) \quad (3.24)$$

For the specification of prior distributions and hyper priors, the joint posterior distribution of the parameters is given by

$$f\left(\boldsymbol{\xi}, \boldsymbol{\psi}, \boldsymbol{\Gamma}, \mu_{person}, \mu_{item}, \mu_{testlet}, \Sigma_{person}, \Sigma_{item}, \Sigma_{testlet} \middle| \boldsymbol{y}, \boldsymbol{log}(\boldsymbol{T}), \boldsymbol{u}\right)$$

$$\propto \prod_{i=1}^{I} \prod_{j=1}^{J} f\left(y_{ij}; \alpha_i, \beta_j, \delta_j, \gamma_d\right) f\left(\log(T); \zeta_j, \tau_i, \lambda_d\right) f(u_{ij}; \theta_i, b_j) f(\alpha_i; \theta_i, \iota_k, \kappa_k)$$

$$\times f(\xi_i; \mu_{person}, \Sigma_{person}) f(\psi_j; \mu_{item}, \Sigma_{item}) f(\Gamma_d; \mu_{testlet}, \Sigma_{testlet})$$

$$f\left(\mu_{person,} \Sigma_{person}\right) f(\mu_{item}, \Sigma_{item}) f(\mu_{testlet}, \Sigma_{testlet}) f(\iota_k) f(\kappa_k) \quad (3.25)$$

3.2.2 Implementation of Markov Chain Monte Carlo

This research uses the Markov chain simulation method, the Gibbs sampler, to implement Bayesian estimation of model parameters. Bayesian inferences are based on the assumption that the distributions of the simulated values are close to the target distribution and use iterative simulation draws from $p(\theta|y)$ to summarize the posterior density. As stated in Chapter 2, the Markov chain simulation draws values of θ from approximate distributions and subsequently corrects those draws to better approximate the target posterior distribution (Gelman et al, 2014). The Gibbs sampler is an appropriate method as it can treat the parameters in the domain-specific models as blocks of parameters and iterate through draws from the conditional distributions

of one block of parameters conditional on all remaining parameters. As suggested by Gelman et al (2014), this research simulates a minimum of two chains to allow effective monitoring of convergence. Further, it runs a sufficient number of iterations until convergence is reached. Within each chain, the first half iterations of the simulated runs are discarded to minimize the influence of the starting values. Each of the remaining chains are split up into the first and second half to allow simultaneous testing of mixing and stationarity. Convergence of the iterative simulation is diagnosed by the potential scale reduction factor $\hat{R}$. $\hat{R}$ close to 1 is an indication that convergence has reached, at which point posterior density will be summarized.

3.3 Simulation Design

The simulation study proposed in this research addresses the fit of the proposed model and the degree to which model parameters are adequately recovered. This section describes the design of the simulation study conducted in this research, factors that are fixed versus manipulated, and criteria used to evaluate model fit and parameter recovery, followed by a description of the methodology adopted for the analysis of empirical data in section 3.4.

3.3.1 Fixed Factors

The simulation study is designed to evaluate model performance and parameter estimation under simulated conditions and across three models differentiated on the inclusion or exclusion of testlet parameters or the AC component. The simulation design fixes specific factors to create conditions under which the fit of the different models and the precision with which model parameters

are estimated can be compared. Factors fixed in the simulation study are: a) test design including the Q-matrix design and the number of test items; and b) specific elements of the distributions for the priors and hyper priors that are not manipulated.

This research specifies a higher-order DINA model for the response data. An essential component of this model is the Q-matrix. To allow comparison across the models, a uniform test design and Q-matrix are used in the simulation study for data generation, model fitting, and parameter estimation. The test is a portion of a standardized large-scale Math Test featuring 25 item math test that measure five attributes and was used in an empirical analysis of the response, RT, and ACs data for cognitive diagnosis (Jiao, Ding, & Yin, 2020). Table 3.1 shows the Q-matrix for the portion of the test chosen for this research study. Items A1, A2, A5, A7, and A8 depend on Attribute AF-1; items B2, B4, and B9 load on Attribute AF-10; items B3, B4, B19, and B28 measure Attribute AF-3; five items measure Attribute AF-5: items A9, A11, B10, B11, and B19; and the last attribute, Attribute DA-1, is measured by 9 items: items B1, B5, B7, B12, B13, B15, B17, B22 and B23. For comparison purposes, this test design and Q-matrix are used for data generation and model estimation.

Table 1
Q-Matrix for the Simulation Study

Item	AF-1	AF-10	AF-3	AF-5	DA-1
A1	1	0	0	0	0
A2	1	0	0	0	0
A3	0	0	1	0	0
A4	0	0	1	0	0
A5	1	0	0	0	0
A7	1	0	0	0	0
A8	1	0	0	0	0
A9	0	0	0	1	0

A11	0	0	0	1	0
B1	0	0	0	0	1
B2	0	1	0	0	0
B4	0	1	0	0	0
B5	0	0	0	0	1
B7	0	0	0	0	1
B9	0	1	0	0	0
B10	0	0	0	1	0
B11	0	0	0	1	0
B12	0	0	0	0	1
B13	0	0	0	0	1
B15	0	0	0	0	1
B17	0	0	0	0	1
B19	0	0	1	1	0
B22	0	0	0	0	1
B23	0	0	0	0	1
B28	0	0	1	0	0

Specific components of the distributions for the priors and hyper priors described in Chapter 2 are likewise fixed to compare models including or excluding the ACs model or the testlet parameters. As stated in Chapter 2,

the person parameters in the proposed model are assumed to follow a bivariate normal distribution

$$\begin{pmatrix}\theta_i \\ \tau_i\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \Sigma_{\text{person}}\right) \tag{3.26}$$

Person parameters will be generated from a normal distribution where $\theta_i \sim N(0,1)$ and $\tau_i \sim N(0, 0.25)$. Higher-order structural parameters are fixed at $\kappa_k = 1.5$ for all attributes and $\iota_k = (= (-0.8, 0, 0.8, -0.8, 0.8)$, which indicates moderate correlations among attributes (Zhan et al, 2018b). Person attribute mastery parameter α_{ik} will be generated from a Bernoulli distribution $\alpha_{ik} \sim Bernoulli\ \left(P(\alpha_{ik} = 1)\right)$, as specified in Chapter 2. The multivariate normal distribution from which item parameters will be generated is fixed as

$$\begin{pmatrix}\beta_j\\ \delta_j\\ \zeta_j\\ b_j\end{pmatrix}\sim N\left(\begin{pmatrix}-2.197\\ 4.394\\ 4.000\\ 0\end{pmatrix},\begin{pmatrix}1.00 & & & \\ -0.8 & 1.00 & & \\ -0.25 & 0.15 & 0.25 & \\ -0.5 & 0.16 & 0.36 & 1.00\end{pmatrix}\right) \quad (3.27)$$

And $\sigma_{\varepsilon_i} = .05$ for all items. Item step parameters are fixed as

$$b_{j1}\sim N(I,0,1), b_{j2}\sim N(I,0,1), b_{j3} = -sum(b_{j1}, b_{j2}) \quad (3.28)$$

RA and RT testlet parameters will be generated from the same bivariate normal distribution

$$\Gamma_{ij(d)} = \begin{pmatrix}\gamma_{id(j)}\\ \lambda_{id(j)}\end{pmatrix}\sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \Sigma_{testlet,d}\right) \quad (3.29)$$

where $\sigma_{\gamma\lambda} = -0.25$. Parameters $\rho_{\theta\tau}$, σ_γ^2, and σ_λ^2 are manipulated factors in this simulation study, as will be discussed in the next subsection.

3.3.2 Manipulated Factors

Factors manipulated in the simulation study are a) sample size; b) the variance of the testlet effects; and 3) the correlation between the speed parameter and the ability parameter. Different sample sizes and levels of the variances of the testlet effects and of the correlations between the speed and ability parameters are configured to examine the effects of the variation on the degree to which the proposed model and the comparison models recover the parameters of interest in this research study.

One of the most often manipulated factors in simulation studies of model performance and parameter estimation is sample size, the choice of which is contingent upon model specification and parameters of interest. Sample sizes vary across the simulation studies reviewed in Chapter 2, with some studies having a fixed

sample size, and others varying sample sizes to examine the conditions under which parameters can be optimally recovered. Studies with a fixed sample size often opt for 1,000 as in studies specifying multilevel testlet models (Jiao et al, 2012; Jiao et al 2013) and polytomous multilevel testlet models (Jiao & Zhang, 2015). Bolsinova and Tijmstra (2019) similarly fix the sample size at 1,000 when fitting a model differentiating RTs for correct and incorrect responses. In joint modeling of responses and RTs for cognitive diagnosis, Zhan et al. (2018a) and Zhan et al. (2018b) fix the sample size at 1,000. Other studies specify different sample sizes to examine the extent to which sample size affects model performance. Wang and Wilson (2005), for instance, compare recovery of the parameters in the Rasch testlet model for sample sizes of 200 and 500. Their simulation study suggests that as sample sizes increased, the root mean square errors of the parameter estimates decreased to an acceptable level. Fox et al. (2020) similarly set the sample size at 200, 500, and 1,000 in a simulation study that compares the Bayesian covariance structure model (BCSM) for testlets and the random effects models for testlets. Their study suggests that parameter estimation is more accurate for BCSM when sample size is small. In a simulation study of the performance of the 4PLRT model, Wang and Hanson (2005) opt for three different sample sizes, 1000, 2000, and 4000. Their study shows that increasing sample size consistently reduces standard error and root mean square error but does not necessarily result in smaller bias. Based on the review of the sample sizes set in these studies, this research sets sample size at 200 and 500 to examine the extent to which sample sizes affect parameter estimation in the proposed model.

The levels of testlet variance are indicative of the magnitude of the testlet effects and are one of the most-often examined factors in the modeling and estimation of testlet response models (e.g., Bradley et al, 1999; Fox et al, 2020; Jiao et al, 2012; Jiao et al, 2013; Jiao & Zhang, 2015; Wang et al, 2002; Wang & Wilson, 2005). In studies of testlet effects, varying sets of variances have been used to represent the magnitude to the testlet effects. Wang and Wilson (2005), for instance, set the variances of the random testlet variables at 0.25, 0.50, 0.75, and 1.00, presenting small to large effects of the testlets. Jiao et al (2012) used two levels of variances, 0.25 and 1.00 to indicate low and moderate local item dependence and person clustering effects. Jiao et al (2013) specify four levels of testlet variance at 0, 0.25, 0.5625, and 1, representing testlets effecting ranging from none to large. Fox et al (2020) focus on small testlet variances and set them at 0.1, 0.05, and 0.01. The review of relevant literature suggests that while the studies similarly use 0.25 and 0.5/0.5625 to indicate small and moderate testlet effects, the labeling of a large testlet effect seems to be at the researchers' discretion. In this dissertation study, local response and RT testlet variances are set at 0.25, 0.5, and 1.00 to represent small, moderate, and large testlet effects to simulate conditions under which model performance and parameter estimation can be examined.

The third factor for which different levels are chosen in this research study is the correlation between ability and speed. In examining the basic issues in RT modeling, van der Linden (2009) equates SAT in reaction time research with the speed-ability tradeoff in testing and concludes that the two are related through a distinct function $\theta = \theta(\tau)$ for each test-taker, which models of responses and RT do

not need to incorporate but require fixed parameters for the effective speed and ability of the test-takers. Empirical evidence suggests that the correlations between the two parameters can be positive or negative, with studies reporting them to range from -0.65 to 0.30 and suggesting that more capable students may have better time-management skills and strategically speed up or slow down to meet the time constraint of the tests (van der Linden, 2009). In simulation studies of response and RT, various levels of the correlations between speed and ability are chosen. Molenaar and De Boeck (2018) set the correlation at 0.4 in response mixture modeling that accounts for heterogeneity in item characteristics across response times. Bolsinova and Tijmstra (2019), in their model differentiating RTs for correct and incorrect response, specify two different levels for correlations between speed and ability, 0, and 0.5, with 0 representing the baseline condition and 0.5 indicating that response speed provides collateral information for the estimation of ability. Zhan et al. (2018a) compared parameter estimates for four different levels of the correlation: -0.5, -0.3, 0.3, and 0.5. To simulate testing conditions with varying time constraints and the varying degree to which response speed provides collateral information for the estimation of ability, this research uses four levels for the correlations between speed and ability, -0.5, -0.3, 0.3, and 0.5 as a facet of the manipulated conditions under which model performance and parameter recovery can be examined.

The three factors manipulated in this study intersect with the proposed model and comparison models to create a total of 24 conditions under which model performance and parameter estimates can be compared. 30 replications are run under each condition for each of the three model, yielding a total of 2,160 datasets.

3.3.3 Data Generation Procedure

Data generation consists of the generation of item, person, and testlet parameters specified in the measurement models for response, RT, and ACs presented in sections 3.1.1 through 3.1.3, followed by plugging them into the models to generate item response, RT, and ACs datasets. The following steps comprise the procedure taken to generate the datasets.

Simulation of the Item and Item Step Parameters. The initial step in the data generation process is the generation of the item and item step parameters. As is stated under 3.1.4, the item parameters are β_j, δ_j, ζ_j, and b_j; they are generated from the multivariate distribution specified in Equation 3.27. ω_j are generated from $N(2, 0.25)$. Item step parameters are generated from the normal distributions specified in Equation 3.28. They are combined with the item parameters as true item parameters to provide a point of reference for determining the extent to which estimated item parameters diverge from them.

Simulation of the Person Parameters. The next step in this procedure is the generation of person parameters and higher-order structural parameters. As is stated under 3.1.4, the person parameters are θ_i and τ_i; they were generated from the bivariate normal distribution specified in Equation 3.17. As is seen in this equation, the means of θ_i and τ_i are constrained to 0. σ_θ^2 is fixed as 1 and σ_τ^2 is fixed as 0.25. $\rho_{\theta\tau}$ is a manipulated factor that takes on the values of -0.5, -0.3, 0.3, and 0.5, resulting in the corresponding levels in $\sigma_{\theta\tau}$.

Simulation of the Testlet Parameters. A total of 5 testlets are presumed to underly the 25 items presented in Table 1, each consisting of 5 items. As is presented

in Section 3.1.4, the testlet parameters are the response testlet effect parameter, $\gamma_{id(j)}$ and the RT testlet parameter, $\lambda_{id(j)}$. The five pairs of $\gamma_{id(j)}$ and $\lambda_{id(j)}$ are generated from the same bivariate normal distribution specified in equations 3.23, and 3.24. $\sigma_{\gamma\lambda}$is fixed at -0.25. $\sigma^2_{\gamma_{id_{(j)}}}$ and $\sigma^2_{\lambda_{id_{(j)}}}$ are constrained to be the same, the magnitude of which is manipulated to take on three values as discussed under Section 3.3.2.

Simulation of the Attribute Patterns. As is described under Section 3.3.1, the higher-order structural parameters are fixed at $\kappa_k = 1.5$ for all attributes and $\iota_k = (-0.8, 0, 0.8, -0.8, 0.8)$. These are plugged into Equation 3.6 in Section 3.1.1 to generate the attribute pattern matrix indicating the probability of every examinee's mastery status on the five attributes specified in the Q-matrix. Elements of the pattern matrix are specified as following a binomial distribution, serving as points of reference for calculating attribute and attribute profile classification accuracy.

Simulation of the Response Data. The response data are simulated by plugging in the generated item parameters β_j and δ_j, the generated response testlet parameter $\gamma_{id(j)}$ into Equation 3.5 to generate the probabilities of the examinees giving a correct response to the items. The latent response variable in this equation $\prod_{k=1}^{K} \alpha_{ik}^{q_{ik}}$ is computed as the product of the generated attribute patterns and the Q-matrix. The response data are specified as following a binomial distribution.

Simulation of the RT Data. The RT time data are similarly simulated by plugging in the generated item time intensity parameter ζ_j, ω_j, the person speed parameter τ_i, and the RT testlet parameter $\lambda_{id(j)}$ into Equation 3.8. They are specified as following a lognormal distribution.

Simulation of the ACs Data. The ACs data are likewise simulated by plugging in the generated item difficulty parameter b_j and the item step parameters into Equation 3.10. Answer changes data set and the item response dataset follow the same dimension. The responses data are binary and consist of 0s and 1s; the answer change data are categorical and consist of 1s, 2s, 3s, and 4s. Conditional dependence is established through generating patterns 1 and 2 for responses that are 0 by generating the probability for category 1 and category 2. They are then scaled up by a constant of 1 if the responses are 1. This results in the correspondence between the response and ACs data with a response of 0 only having an AC pattern of 1 or 2, and a response of 1 only having an AC pattern of 3 or 4.

3.3.4 Evaluation Criteria

This dissertation research uses three sets of indices to evaluate the accuracy and precision with which parameters of interest are estimated by the Monta Carlo simulation study, comparative model fit, and classification accuracy for the proposed model and comparison models. The first set of indices compares parameter estimates and their true values generated by the proposed model and the comparison models. The level of estimation precision is determined by two indices, bias and standard error (SE) of the estimate, respectively given by

$$Bias(y) = \frac{\sum_{r=1}^{R}(\hat{y}_r - y_{true})}{R} \tag{3.30}$$

$$SE(y) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{y}_r - \frac{\sum_{r=1}^{R}\hat{y}}{R}\right)^2} \tag{3.31}$$

where R denotes the total number of replications, y_{true} is the true value of parameter if the parameter of interest, and $\hat{y}_r$ is an estimate of the parameter y for replication r. $Bias(y)$ is the systematic error indicating the extent to which estimated values deviate from the true value of the parameter across replications, and $SE(y)$ is the random error indicating the variability among parameter estimates without referencing the true value of the parameter.

Indices used to evaluate and compare the fit of the proposed model and comparison models are the Akaike information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978), and the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). Information criteria are measures of predicative accuracy and are typically based on the deviance, $-2\log p\left(y|\hat{\theta}\right)$ (Gelman et al, 2014). AIC corrects for the increase in predictive accuracy caused by the fitting of k parameters by subtracting k from the log predicative density given the maximum likelihood estimate, $\log p\left(y|\hat{\theta}_{mle}\right)$, and multiplies it by -2. AIC is given by

$$AIC = -2\log p\left(y|\hat{\theta}_{mle}\right) + 2k \tag{3.32}$$

where $\log p\left(y|\hat{\theta}_{mle}\right)$ is the log predicative density given the maximum likelihood and k is the number of fitted parameters. BIC replaces the maximum likelihood estimate $\hat{\theta}$ with the posterior mean $\hat{\theta}_{Bayes} = E(\theta|y)$ and k with effective number of parameters. DIC is given by

$$DIC = -2\log p\left(y|\hat{\theta}_{Bayes}\right) + 2p_{DIC} \tag{3.33}$$

where $\hat{\theta}_{Bayes}$ is the posterior mean and p_{DIC} is the effective number of parameters. BIC corrects for the increase in predictive accuracy by a penalty that increases with the sample size n and is given by

$$BIC = -2\log p\left(y\middle|\hat{\theta}\right) + k\log n \tag{3.34}$$

which penalizes large datasets more than AIC and thus performs better for simpler models.

Two indices are used to compare the attribute-level and pattern-level classification accuracy for the proposed model and comparison models: attribute correct classification rate (ACCR), and pattern correct classification rate (PCCR). ACCR evaluates attribute level classification rate and is given by

$$ACCR = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I} W_{ik}}{R \times I} \tag{3.35}$$

where R is the number of replications, I denotes items, and $W_{ik} = 1$ if $\hat{\alpha}_{ik} = \alpha_{ik}$ and $W_{ik} = 0$ if otherwise. PCCR is the pattern-level classification accuracy and is given by

$$PCCR = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I}\prod_{k=1}^{K} W_{ik}}{R \times I} \tag{3.36}$$

where K denotes attributes. Both indices are computed for the proposed model and the comparison models to determine the effect of AC data and testlet effects on the rate at which attributes and attribute patterns are correctly classified.

Within and across each simulated condition, this research uses bias and SE to evaluate and compare the level of precision with which parameters are estimated by the proposed model and comparison models. The proposed research compares model

fit indices AIC, BIC, and DIC for the proposed model and one of the comparison models to identify the best-fitting model.

3.4 Empirical Data Analysis

This research used the proposed model and comparison models to fit and analyze data from a large-scale math test, which included binary data response data, RTs data, and ACs data for 71 respondents. The portion of math items used in the empirical data analysis includes a total of 58 items measuring five attributes. The Q-matrix for this dataset is described in detail in section 4.2.

Two chains and 10000 iterations were run in the analysis of the empirical dataset. Within each chain, the first half of the iterations was discarded as burn-ins. Convergence was assessed by the potential scale reduction factor $\hat{R}$. Model fit indices AIC, DIC, and BIC described in Subsection 3.3.3 were used to evaluate and compare relative fit for the proposed model and comparison models.

Analyses of the empirical data set resulted in estimates of the person, item, and testlet parameters in the best-fitting model, which were be summarized as the mean vector and variance covariance matrices for the three sets of parameters. Estimated higher-order structure parameters and the posterior mixing proportions of the attribute patterns resulting from the analyses were also summarized.

Chapter 4: Results

4.1 Results of the Simulation Studies

The simulation study conducted in this research purport to examine 1) the fit of the proposed joint model of responses, RT, and AC patterns for testlet-based cognitive diagnostic assessments as compared to the two alternative models; 2) the impact of this modeling approach on parameter recovery and classification accuracy; and 3) the effect of three manipulated factors on model performance and parameter estimation for the proposed model.

The proposed model accounting for dual dependency of response and RT and including AC pattern as an additional data source is evaluated in the context of model comparison with two alternative models: 1) the JRT-DINA-R/RT/AC model neglecting testlet effects in responses and RT; and 2) the Joint Testlet-DINA model excluding AC patterns in model specification. The three models are compared on outcome measures including model fit, classification accuracy at the attribute and attribute profile level, and recovery of item and person parameters, higher-order structural parameters, and variance/covariance matrices for item, person, and testlet parameters. Table 2 summarizes and compares the specification of the three data-fitting models in the simulation study.

Three factors manipulated in the simulation studies are: the sample size, the correlation between latent ability and latent speed, and the magnitude of the testlet effects. These factors constitute a total of 24 simulated conditions. 30 replications were run for each simulated condition, resulting in a total of 720 replications. Bayesian estimation of model parameters was implemented simulating two chains

and running 10,000 iterations per chain. Within each chain, the first 5,000 iterations were discarded as burn-in. Convergence of the iterative simulation was determined by the potential scale reduction factor $\hat{R}$, which was close to 1 across the replications in all simulated conditions. Parameter estimation was summarized based on 10,000 iterations. Throughout the simulation conditions and the replications, $\hat{R}$ for the model parameters was ≤ 1.1. Estimation of all three models used around 5 hours per replication for a sample of 200 examinees and 6.5 hours for a sample of 500 examinees.

Table 2
Specification of the Data-Fitting Models in the Simulation Study

Model	Distinction in Model Specification	
	Dependency of Responses and Response Time	Modeling of Answer Change Patterns
Joint Testlet-DINA	√	x
JRT-DINA-R/RT/AC	x	√
JRT-AC-DINA for Testlets	√	√

Note: √ indicates presence; x indicates absence.

The following sections present the results of the simulation study and summarize them by the criteria used to evaluate and compare the three models. The first section presents and compares model fit indices for the two models that use the same set of data. The remaining sections examine the effects of accounting for testlet effects and of including AC patterns as an additional data source on parameter estimates by comparing the classification accuracy for the attributes and attribute profiles and the biases and SEs for the parameters estimated by the three models at the levels of the manipulated factors. Table 3 summarizes the types of parameters estimated by each model in the simulation study.

Table 3

Summary of Parameters of Interest Estimated by the Models in the Simulation Study

Categories	Notation	Description	Presence/Absence in the Models		
			JAD-TT	JAD	JD-TT
Item Parameter	β	Item Intercept	√	√	√
	δ	Item Interaction	√	√	√
	ζ	Item Time Intensity	√	√	√
	b	Item Difficulty	√	√	×
Person Parameter	θ	Person Ability	√	√	√
	τ	Person Speed	√	√	√
Higher-Order Structural Parameter	ι	Attribute Easiness	√	√	√
	κ	Attribute Discrimination	√	√	√
Item Mean Vector	μ_β	Item Intercept	√	√	√
	μ_δ	Item Interaction	√	√	√
	μ_ζ	Item Time Intensity	√	√	√
	μ_b	Item Difficulty	√	√	×
Item Variance and Covariance Matrix	σ_β^2	Variance of Item Intercept	√	√	√
	σ_δ^2	Variance of Item Interaction	√	√	√
	σ_ζ^2	Variance of Item Time Intensity	√	√	√
	σ_b^2	Variance of Item Difficulty	√	√	×
	$\sigma_{\beta\delta}$	Covariance: Item Intercept & Interaction	√	√	√
	$\sigma_{\beta\zeta}$	Covariance: Item Intercept & Time Intensity	√	√	√
	$\sigma_{\beta b}$	Covariance: Item Intercept & Difficulty	√	√	×
	$\sigma_{\delta\zeta}$	Covariance: Item Interaction & Time Intensity	√	√	√
	$\sigma_{\delta b}$	Covariance: Item Interaction & Difficulty	√	√	×
	$\sigma_{\zeta b}$	Covariance: Item Time Intensity & Difficulty	√	√	×
Person Variance and Covariance Matrix	σ_τ^2	Variance of Person Speed	√	√	√
	$\sigma_{\theta\tau}$	Covariance of Person Ability & Speed	√	√	√
Testlet Variance	σ_γ^2	Variance of Response Testlet Effect	√	×	√
	σ_λ^2	Variance of Response Time Testlet Effect	√	×	√

4.1.1 Performance of the Model Fit Indices

Model performance was evaluated by using AIC, BIC, and DIC to compare the fit of the proposed model, JRT-AC-DINA for Testlets, with the alternative model that uses the same dataset: JRT-DINA-R/RT/AC. As described in Chapter 3, both information criteria are measures of predicative accuracy and are typically based on the deviance $-2\log p\left(y\middle|\hat{\theta}\right)$ (Gelman et al, 2014). For a given simulated condition, comparative model fit is determined by comparing the AIC, BIC, and DIC for the two data-fitting models and summarizing the number of replications by which the smallest values of AIC and BIC are identified.

Table 4 summarizes the number of replications by which the smallest values of AIC and BIC are identified. Across all simulated conditions, AIC and BIC values for the proposed model, JRT-AC-DINA for Testlets, are consistently smaller compared to the JRT-DINA-R/RT/AC model, suggesting that the proposed model has better fit than the JRT-DINA-R/RT/AC model. The DIC values, however, are higher for the proposed model than for the JRT-DINA-R/RT/AC model, suggesting that the JRT-DINA-R/RT/AC model has a better model fit. More detailed discussion regarding the usability of the three indices and the caution that needs to be taken when using and interpreting them is presented in Subsection 5.1.2.

Table 4

Number of Replications in Identifying the Best-Fitting Model in the Simulation Study

Condition No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	AIC		BIC		DIC	
				JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC
1	200	-0.5	0.25	30	0	30	0	0	30
2			0.5	30	0	30	0	0	30
3			1	30	0	30	0	0	30
4		-0.3	0.25	30	0	30	0	0	30
5			0.5	30	0	30	0	0	30
6			1	30	0	30	0	0	30
7		0.3	0.25	30	0	30	0	0	30
8			0.5	30	0	30	0	0	30
9			1	30	0	30	0	0	30
10		0.5	0.25	30	0	30	0	0	30
11			0.5	30	0	30	0	0	30
12			1	30	0	30	0	0	30
13	500	-0.5	0.25	30	0	30	0	0	30
14			0.5	30	0	30	0	0	30
15			1	30	0	30	0	0	30
16		-0.3	0.25	30	0	30	0	0	30
17			0.5	30	0	30	0	0	30
18			1	30	0	30	0	0	30
19		0.3	0.25	30	0	30	0	0	30
20			0.5	30	0	30	0	0	30
21			1	30	0	30	0	0	30
22		0.5	0.25	30	0	30	0	0	30
23			0.5	30	0	30	0	0	30
24			1	30	0	30	0	0	30

4.1.2 Recovery of the Person Parameters

Person parameters evaluated in this study are the latent ability parameter, θ_i, and latent speed parameter, τ_i. Inferences are drawn regarding individuals' latent ability and speed based on information provided by the person parameters. To evaluate the degree to which accounting for testlet effects and including AC pattern as an additional data source affect estimation of person parameters and classification accuracy at the attribute and attribute profile level, this section summarizes and compares the ACCR and PCCR for each of the estimation models and presents mixed-effect ANOVA results on the effects of the model type and manipulated factors on the biases and SEs of θ_i and τ_i.

Attribute Mastery Status. Examinees' attribute mastery status is indicated by ACCR and PCCR. These are summarized for each of the data-fitting models, compared under all simulated conditions and presented in Tables A.1.1-4. As is indicated by the tables, across all the 24 simulated conditions, ACCRs for all five attributes for the proposed model, JRT-AC-DINA for Testlets, are > 0.90, and PCCR is > 0.74, suggesting that overall and for each attribute, the attribute mastery status of more than 90% of the simulated examinees are correctly classified using the proposed model and the attribute profile for over 74% of the simulated examinees are correctly recovered by the proposed model.

Further, compared with the PCCRs of the JRT-DINA-R/RT/AC model, the PCCRs for the proposed model are only slightly higher under all simulated conditions. The ACCRs for α_4 and α_5 for the proposed model are slightly higher than or equal to those for the JRT-DINA-R/RT/AC model in all 24 simulated

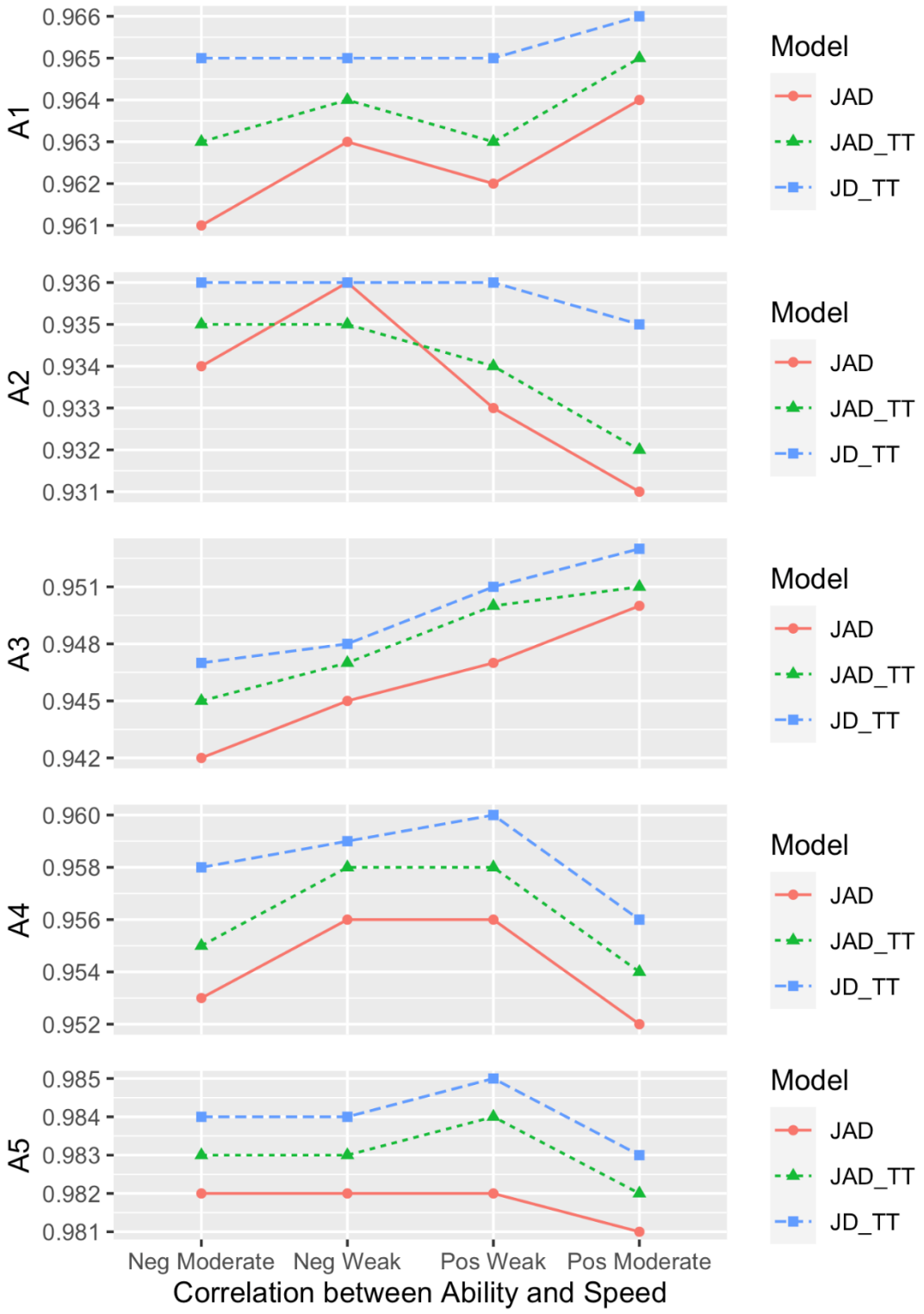

Figure 1 Marginal mean attribute correct classification rates (ACCRs) at each level of the correlation between higher-order person ability and speed. A1 to A5 indicates Attribute 1 to Attribute 5.

conditions, and in 23 out of the 24 conditions for α_3, 21 out of 24 for α_1, and 18 out of 24 for α_2. When compared with the Joint Testlet-DINA model, however, the ACCRs and PCCRs for the proposed model are slightly less or equal across all 24 simulated conditions. The differences, however, are small and can be considered negligible.

Figures 1 through 4 show the marginal mean ACCRs and PCCRs for the three models being compared at each level of the three manipulated factors. Sample size and correlation between latent speed and ability do not appear to result in differences in ACCRs and PCCRs for the three models: they are similar across the levels of the two factors. The magnitude of the testlet effects does have an effect as both ACCRs and PCCRs decrease as the variance of the testlet effects parameters increases from 0.25 to 1 and are the smallest for conditions that feature large testlet effects. The indicates that an increase in the magnitude of the testlet effects corresponds to reduced accuracy rate at both the attribute and attribute profile level.

These results indicate the proposed model successfully recovers the attributes and attribute profiles. They further suggest that while accounting for dual dependency in responses and RT in the joint model of responses and RT slightly improves classification accuracy for attributes and attributes profiles compared with the model that ignores this dependency, when the testlet effects are accounted for, including AC patterns in the joint model does not necessarily lead to improved attribute and attribute profile correct classification rates.

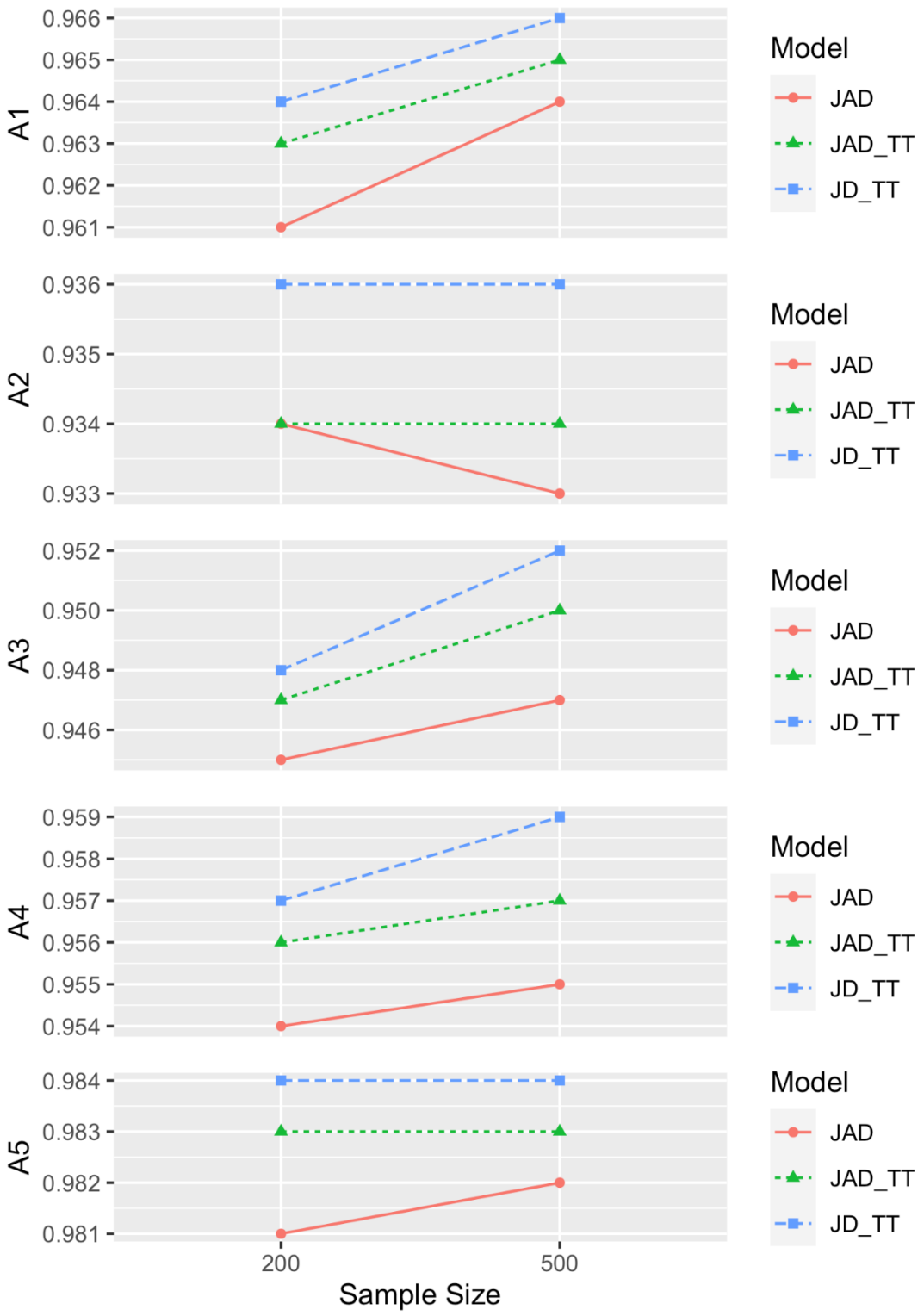

Figure 2 Marginal mean attribute correct classification rates (ACCRs) at each level of the sample size. A1 to A5 indicates Attribute 1 to Attribute 5.

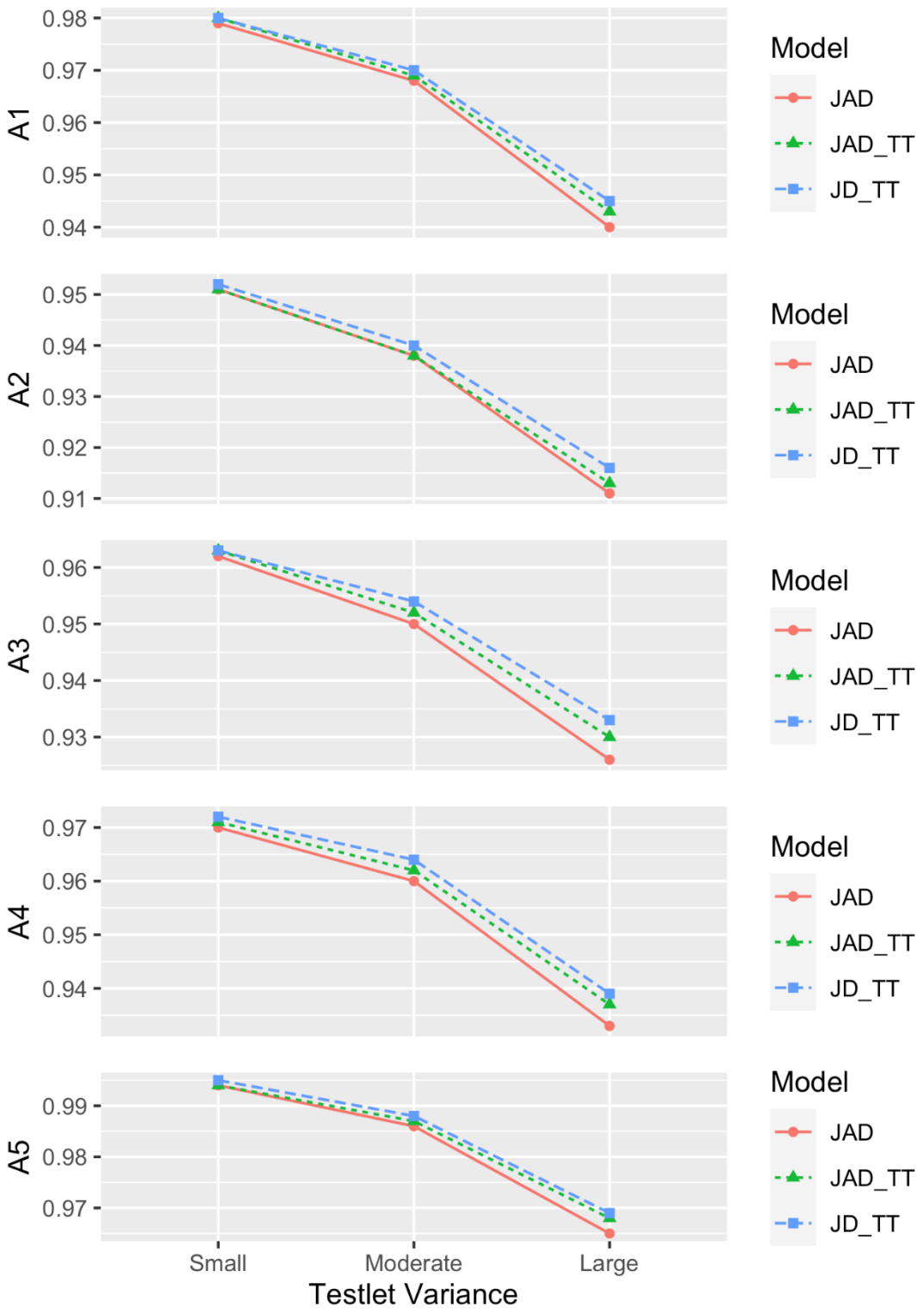

Figure 3 Marginal mean attribute correct classification rates (ACCRs) at each level of the testlet variance. A1 to A5 indicates Attribute 1 to Attribute 5.

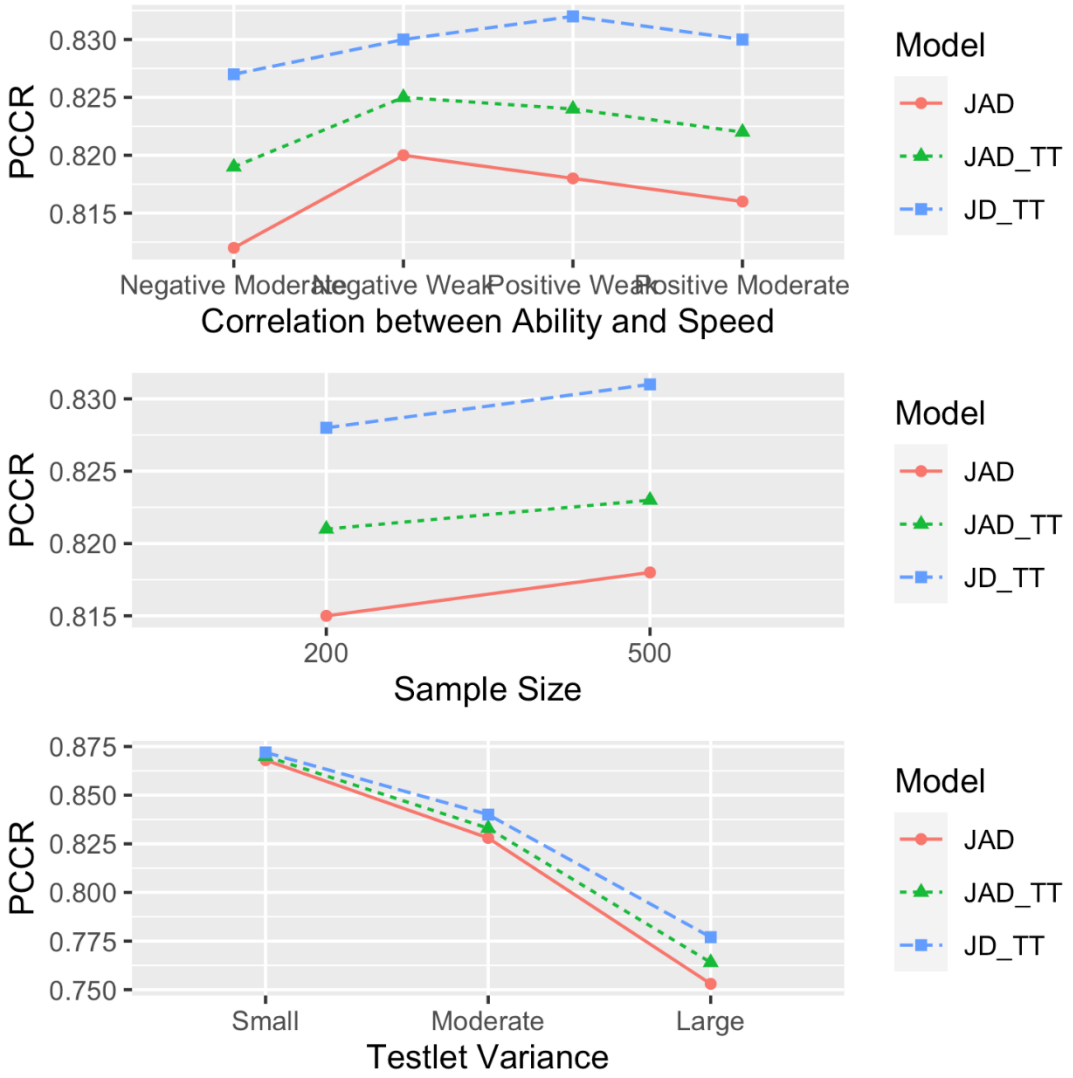

Figure 4 Marginal mean attribute profile classification rates (PCCRs) at each level of the testlet variance.

Correlation between true and estimated person parameters. Table A.2 in Appendix A presents the correlation between true and estimated higher-ability and person speed parameters for the three models under the 24 simulated conditions. As is shown in the table, the correlation is > 0.79 for the higher-ability parameter and > 0.98 for the person speed parameter as estimated by the proposed model, indicating strong correlation between the true and generated person parameters.

Of the three models being evaluated in the simulation study, the correlation for the higher-order ability parameter yielded by the proposed model is the strongest compared to the other two models across the 24 simulated conditions. Further, although overall the correlation computed using person ability parameter estimated by the proposed model is only slightly higher than that for the JRT-DINA-R/RT/AC model, compared to the Joint Testlet-DINA model, the correlation computed by using estimates from the two models are stronger by up to 5%. This indicates that the inclusion of AC pattern in the joint model of responses and RT contributes to improved correlation between true and estimated higher-ability parameter, and additionally accounting for testlet effects in the responses and RT slightly improves this correlation.

The proposed modeling approach, however, has little effects on correlation for the other person parameter, person speed parameter: across the simulated conditions correlation for this parameter is identical or only slightly different for the three models being evaluated. Thus, all three models yield estimated person speed parameter that correlates strongly with the true parameter and including AC pattern and accounting for testlet effects do not necessarily improve this correlation.

While the manipulated factors have little effects on the correlation for the person speed parameter, the variance of the response and RT testlet effect parameters is related to the correlation for the higher-order ability parameter. Across all three models, as the variance of the testlet effect parameters increases from 0.25 to 1, the correlation decreases, and is the smallest for the conditions that have a large variance of 1. This indicates that an increase in the magnitude of the testlet effects corresponds

to weakened correlation between true higher-order ability parameter and its estimates by all three models.

These results indicate that the proposed model yields estimates of the higher-order ability parameter that correlate more strongly with the true parameter compared with the other two models. Thus, modeling AC patterns in addition to responses and RT and accounting for the testlet effects lead to stronger correlation for the higher-order person ability parameter. Further, consistent with the impact of the variance of the testlet effect parameter on the ACCRs and PCCRs, increasing magnitude of the testlet effects leads to weaker correlation between the generated higher-order ability parameter and its estimates yielded by all three models.

Higher-Order Ability and Person Speed Estimates. Person ability parameters include the higher-order ability parameter θ_i, the person speed parameter τ_i, and their corresponding mean vector and variance-covariance matrix. θ_i and τ_i are individual-specific first-level parameters, and their corresponding mean vector and variance-covariance matrix are population-specific second-level parameters. This section presents the bias and SE for the two parameters to evaluate their recovery.

Mixed-effect ANOVAs were employed to examine the effects of the data-fitting model type and the manipulated factors on the recovery of θ_i and τ_i. Specifically, identifying the effects of the data-fitting model allows for inferences regarding the impact of including the modeling of AC patterns and testlet effects on the recovery of the two model parameters. Mixed-effects ANOVAs were performed separately for the two different sample sizes to ensure the robustness of the analyses to violation of the homogeneity of residual variances assumption. In the analyses, the higher-order ability parameter and person speed parameter were treated as subjects and their biases and SEs

were treated as the dependent variable. The within-subject factor was the model type, and the between-subject factors were the two other manipulated factors: correlation between θ_i and τ_i and testlet variance.

The following sections report the statistically significant effects and their effect sizes. Only the highest-order significant interaction or main effects with at least a small effect size were reported as the interpretation of lower-order interaction or main effect would be misleading if higher-order interaction effects are significant. Table 5 summarizes the highest-order significant effects with at least a small effect size and their effect sizes identified in the mixed-effect ANOVAs.

Table 5
Summary of Effect Sizes of the Highest-Order Significant Effects from the Mixed-Effect ANOVA on the Recovery of the Higher-Order Ability and Person Speed Parameter

N	Effect	Higher-Order Ability θ_i		Person Speed τ_i	
		Bias	*SE*	*Bias*	*SE*
200	Model*Testlet Variance		0.011		
	Model*Correlation*Testlet Variance			0.156	
500	Model				0.932
	Model*Testlet Variance		0.019		
	Model*Correlation		0.030		
	Model*Correlation*Testlet Variance			0.134	

Note: Effect Size is classified as follows: Small (0.01≤partial η^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

At the sample size of 200, two manipulated factors, correlation between θ_i and τ_i and testlet variance, interact with model to affect the bias of τ_i with a large effect size of 0.156. The SE of θ_i is significantly affected by the interaction between testlet variance and model type, the effect size for which is small at 0.011. At the sample size of 500, significant effect on the SE of θ_i stems from the interaction of the testlet variance and model type which results in a small effect size of 0.019. The two

manipulated factors interact with the correlation between θ_i and τ_i to affect the bias of τ_i , the effect size for which is medium at 0.134. Model type has a significant main effect on the SE of τ_i with a large effect size of 0.932. An additional significant effect on SE of θ_i is the interaction between model type and correlation between θ_i and τ_i , having a small effect size of 0.030.

Higher-Order Ability Estimates Table 6 further details the significant main and interaction effects on the bias and SE of θ_i with at least a small effect size at the sample size level of 200. Model type has a significant main effect on the SE of θ_i with a large effect size of 0.860. Model type also interact with testlet variance to significantly affect the SE of θ_i, resulting in an affect size of 0.011.

Table 6
Effect Sizes in the Mixed-Effect ANOVA Results of the Bias and SE of the Higher-Order Ability Estimates (N=200)

Source	Bias of θ			SE of θ		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment						
Model				14702.996	0.000	0.860
Model*Testlet Variance				12.990	<0.001	0.011

Note: Effect Size is classified as follows: Small (0.01≤partialη^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

Table 7 presents the descriptive statistics for the SE of θ_i at each testlet variance level. The proposed model yields better SE of θ_i compared with the models neglecting testlet effects and excluding the modeling of AC patterns when testlet variance equals 0.5 and 1. When testlet variance equals 0.25, the proposed model and the JRT-DINA-R/RT/AC model yield a smaller SE for θ_i than the Joint Testlet-DINA model excluding the modeling of AC patterns.

Table 7
Means and SD for the SE of the Higher-Order Ability Estimates by Model Type and Testlet Variance (N=200)

	JAD-TT		JAD		JD-TT	
$\sigma_\gamma^2/\sigma_\lambda^2$	Mean	SD	Mean	SD	Mean	SD
0.25	.7215	.0933	.7215	.0932	.7799	.0796
0.5	.7230	.0989	.7234	.0990	.7764	.0831
1	.7392	.0998	.7395	.1001	.7771	.0765

Table 8 is a more detailed presentation of the significant effects in the SE of θ_i with at least a small effect size for the sample size of 500. As is shown in the table, Model has a significant main effect on the SE of θ_i. The effect of the model is large at 0.462. Additional significant interaction effects on the SE of θ_i are attributable to the interaction between model and correlation between θ_i and τ_i and between model and testlet variance, both having a small effect size of less than 0.060.

Table 8
Effect Sizes in the Mixed-Effect ANOVA Results of the Bias and SE of the Higher-Order Ability Estimates (N=500)

Source	Bias of θ			SE of θ		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment						
Model				5149.775	0.000	0.462
Model*Correlation				61.570	<0.001	0.030
Model*Testlet Variance				58.443	<0.001	0.019

Note: Effect Size is classified as follows: Small (0.01≤partialη^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

Table 9 presents the descriptive statistics for the SE of θ_i at each testlet variance level. Across the levels, the SE of θ_i is slightly smaller for the proposed model than for the JRT-DINA-R/RT/AC model, which, in turn, is smaller than the Joint Testlet-DINA model. These results indicate that the proposed model yields slightly better SE of θ_i than the model neglecting testlet effects and evidently better SE than the model excluding the modeling of AC patterns.

Table 9
Means and SD for the SE of the Higher-Order Ability Estimates by Model Type and Testlet Variance (N=500)

	JAD-TT		JAD		JD-TT	
$\sigma_\gamma^2/\sigma_\lambda^2$	Mean	SD	Mean	SD	Mean	SD
0.25	0.7184	0.0910	0.7185	0.0910	0.7833	0.0805
0.5	0.7230	0.0943	0.7232	0.0943	0.7793	0.0809
1	0.7352	0.0981	0.7355	0.0981	0.7799	0.0799

Table 10 presents the descriptive statistics for the SE of θ_i yielded by the three models at each correlation level. The SE of θ_i is slightly better for the proposed model than for the JRT-DINA-R/RT/AC model at the correlation levels of 0.3, 0.5, and -0.5. Across the levels, the SE of θ_i as estimated by the Joint Testlet-DINA model is the largest compared to the other two models. These results indicate that the proposed model and the JRT-DINA-R/RT/AC model yield better SE of θ_i than the model excluding the modeling of AC patterns across all correlation levels. When correlation equals 0.3, 0.5, and -0.5, the proposed model yields a slightly smaller random error for θ_i than the JRT-DINA-R/RT/AC model.

Table 10
Means and SD for the SE of the Higher-Order Ability Estimates by Model Type and Correlation between the Higher-Order Ability and Person Speed (N=500)

	JAD-TT		JAD		JD-TT	
	Mean	SD	Mean	SD	Mean	SD
0.3	.7304	.0972	.7306	.0972	.7729	.0778
0.5	.7236	.0955	.7237	.0955	.7896	.0847
-0.3	.7260	.0938	.7260	.0938	.7734	.0768
-0.5	.7222	.0925	.7224	.0924	.7874	.0808

Person Speed Estimates. Table 11 further details the significant effects in the bias and SE of τ with at least a small effect size at the sample size of 200. Testlet variance and correlation between θ_i and τ_i have a significant main effect on the bias of τ_i. Model has a main effect on the SE of τ_i. Also significant in their effects on the

bias of τ_i are two-way interaction effects of model and correlation between θ_i and τ_i, of model and testlet variance, and of correlation between θ_i and τ_i and testlet variance, all having a medium effect size of less than 0.14. The highest-order interaction effect is the interaction of all three factors on the bias of τ_i, having a large effect size of 0.156.

Table 11
Effect Sizes in the Mixed-Effect ANOVA Results of the Bias and SE of the Person Speed Estimates (N=200)

Source	Bias of τ_i			SE of τ_i		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment						
Model*Correlation	95.765	<0.001	0.107			
Model*Testlet Variance	124.898	<0.001	0.095			
Model*Correlation*Testlet Variance	73.580	<0.001	0.156			
Between-Subject Effects						
Correlation	22.642	<0.001	0.028			
Testlet Variance	31.938	<0.001	0.026			
Correlation*Testlet Variance	56.878	<0.001	0.125			

Note: Effect Size is classified as follows: Small (0.01≤partialη^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

Figure 5 is a visual presentation of the interaction between model type and testlet variance on the bias of τ_i for each ability-speed correlation level and at the sample size level of 200. The bias of τ_i yielded by all three models varies by the levels of the testlet variance and of the correlation between θ_i and τ_i. At the

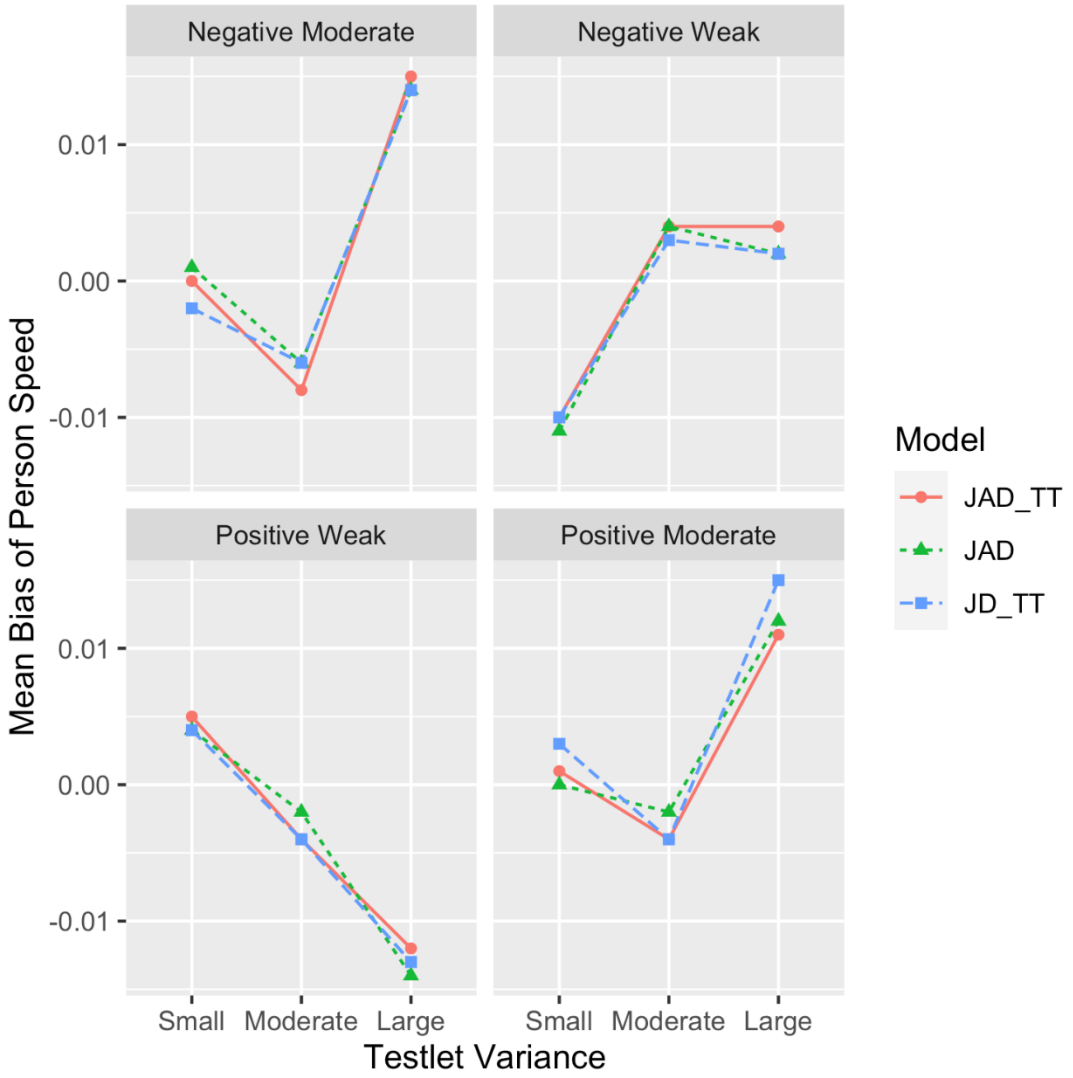

Figure 5. Significant two-way interaction of testlet variance and model type on the bias for τ_i at all correlation levels and for a sample size of 200.

correlation level of 0.3, the absolute value of the bias for τ_i is the highest in conditions having a large testlet effect and lowest in conditions having a moderate testlet effect. Model effect is not consistent across the variance levels: in conditions having a large testlet effect and having an ability-speed correlation of 0.3 and 0.5, the proposed model yields a bias smaller than the other two models; in conditions having a moderate testlet effect, the JRT-DINA-R/RT/AC model yields reduced systematic error compared to the proposed model and the Joint Testlet-DINA model. Similar

patterns are observed for conditions having an ability-speed correlation of -0.3 and -0.5. At the correlation level of -0.3, bias is highest in conditions having a small testlet variance and in conditions having a moderate and a large testlet variance the bias is comparable. The proposed model yields reduced systematic error compared with the Joint Testlet-DINA model and/or the JRT-DINA-R/RT/AC model only in conditions having a small testlet variance.

Table 12 is a detailed presentation of the significant effects in the bias and SE of τ_i with at least a small effect size for the sample size of 500. Model has a significant main effect on the bias of τ_i. Model additionally has a main effect on the SE of τ_i, the effect size for which is large at 0.932. Three two-way interaction effects involving model, testlet variance, and correlation between θ_i and τ_i on the bias of τ_i are also significant. The highest-order interaction effect is the interaction effect of all three factors on the bias of τ_i, having a medium effect size of 0.134.

Table 12
Effect Sizes in the Mixed-Effect ANOVA Results of the Bias and SE of the Person Speed Estimates (N=500)

Source	Bias of τ_i			SE of τ_i		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment						
Model	321.141	<0.001	0.051	81996.888	0.000	0.932
Model*Correlation	633.417	0.000	0.241			
Model*Testlet Variance	240.427	<0.001	0.074			
Model*Correlation*Testlet Variance	154.152	<0.001	0.134			
Between-Subject Effects						
Correlation*Testlet Variance	20.711	<0.001	0.020			

Note: Effect Size is classified as follows: Small (0.01≤partialη^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

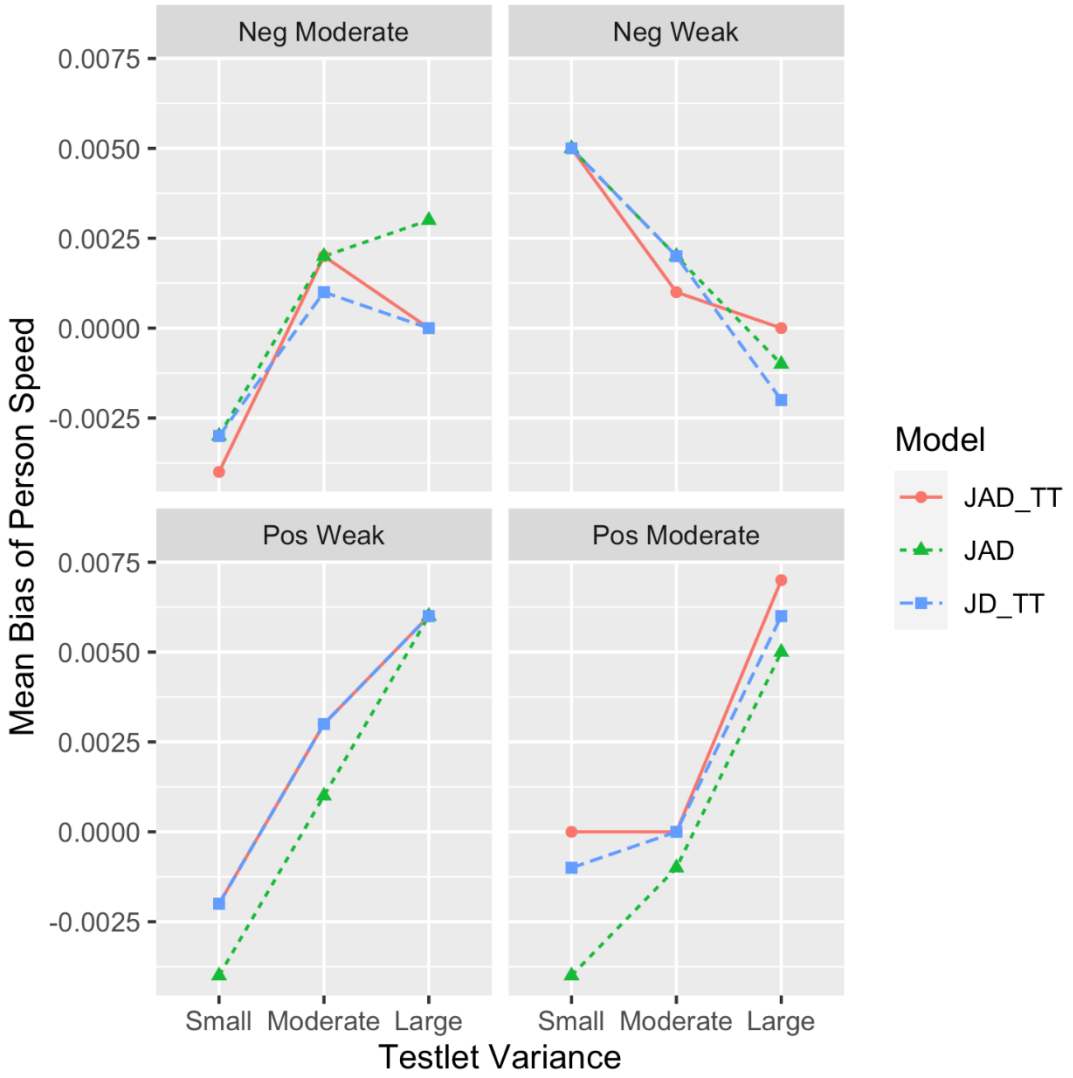

Figure 6. Significant three-way interaction of testlet variance, correlation between θ_i and τ, and model type on the bias for τ at the sample size level of 500.

Figure 6 is a visual presentation of the interaction of testlet variance and model on the bias for τ_i for each level of the correlation between θ_i and τ_i. When the correlation is negative, the proposed model yields slightly reduced bias for τ_i compared to the other two models when testlet variance is large. When the correlation is positive week at 0.3, biases of τ_i yielded by the proposed model and the Joint Testlet-DINA model are identical and are better than by the JRT-DINA-R/RT/AC model ignoring the testlet effects. At a positive moderate correlation of 0.5, the

proposed model yields slightly reduced systematic error when the testlet variance is small at 0.25, compared to the other two models.

Table 13 presents the descriptive statistics for the SE of τ_i for the three models at the sample size of 500. The proposed model and the Joint Testlet-DINA model yield identical SEs that are smaller than the SE yielded by the JRT-DINA-R/RT/AC model. This indicates that the proposed model and the Joint Testlet-DINA model accounting for testlet effects result in reduced random error compared to the JRT-DINA-R/RT/AC model ignoring testlet effects.

Table 13
Means and SD for the SE of the Person Speed Estimates by Model Type (N=500)

	JAD-TT		JAD		JD-TT	
	Mean	SD	Mean	SD	Mean	SD
SE of τ	.4698	.0624	.4781	.0635	.4698	.0624

4.1.3 Recovery of the Person Variance/Covariance Matrix

Table A.3 and Table A.4 in Appendix A summarize the bias and SE for the variance of the person speed parameter and covariance between person speed and higher-order ability under the 24 simulated conditions. Across the conditions, the absolute value of the bias for the two parameters is < 0.04 and the SE is < 0.05.

Compared with the JRT-DINA-R/RT/AC model, the absolute value of the bias of σ_τ^2 for the proposed model is smaller under all 24 simulated conditions, by a very small margin. Only under specific simulated conditions, conditions 1, 8, and 13, is the absolute value of the bias for $\sigma_{\theta\tau}$ slightly reduced for the proposed model. Under the rest of the simulated conditions, the absolute value of the bias for $\sigma_{\theta\tau}$ is either slightly higher than or equal to the bias for the JRT-DINA-R/RT/AC model. Similarly, SEs for σ_τ^2 and $\sigma_{\theta\tau}$ are better for the proposed model, but not across all

simulated conditions. Under 16 out of the 24 simulated conditions is the SE for σ_τ^2 for the proposed model slightly smaller and 13 out of 24 for the SEs for $\sigma_{\theta\tau}$. Compared with the Joint Testlet-DINA model, however, the absolute value of the bias of σ_τ^2 for the proposed model is higher across all simulated conditions whereas the bias of $\sigma_{\theta\tau}$ is smaller in conditions 3, 6, 9, 12, 18, 21, and 24, all having a testlet effect size of 1. Similarly, the SE for both σ_τ^2 and $\sigma_{\theta\tau}$ for the proposed model is slightly higher than for the Joint Testlet-DINA model.

Figures 7 through 10 are visual representations of the marginal mean bias and SE of the estimates of σ_τ^2 and $\sigma_{\theta\tau}$ by levels of the manipulated factors. The proposed model yields marginal mean bias of the estimates of σ_τ^2 that are smaller than for the JRT-DINA-R/RT/AC model but are larger than for the Joint Testlet-DINA model. The marginal mean bias of the estimates of $\sigma_{\theta\tau}$ and the SE of both σ_τ^2 and $\sigma_{\theta\tau}$ for the proposed model and the JRT-DINA-R/RT/AC model are comparable and are larger than for the Joint Testlet-DINA model. Variation by the levels of the sample size is seen in the marginal mean bias of the estimates of $\sigma_{\theta\tau}$ and the SE of both σ_τ^2 and $\sigma_{\theta\tau}$, with a larger sample size of 500 corresponding to reduced values in these estimates. An increase in the variance of the testlet effects corresponds to reduced marginal mean bias and SE of the estimates of σ_τ^2 but increased absolute value of the marginal mean bias of the estimates of $\sigma_{\theta\tau}$. The marginal mean SE of the estimates of $\sigma_{\theta\tau}$, however, is the lowest at the testlet variance level of moderate, increases at the level of large, and is the largest at the level of small. Variation by levels of $\rho_{\theta\tau}$ differs for σ_τ^2 and $\sigma_{\theta\tau}$: whereas the marginal mean bias and SE of the estimates of σ_τ^2 is larger for moderate $\rho_{\theta\tau}$ than for weak $\rho_{\theta\tau}$, the marginal mean bias of $\sigma_{\theta\tau}$ is comparable for

the moderate and weak $\rho_{\theta\tau}$ levels. The marginal mean SE of $\sigma_{\theta\tau}$, by contrast, increases as $\rho_{\theta\tau}$ progresses from negative moderate to positive moderate for the proposed model and the JRT-DINA-R/RT/AC model.

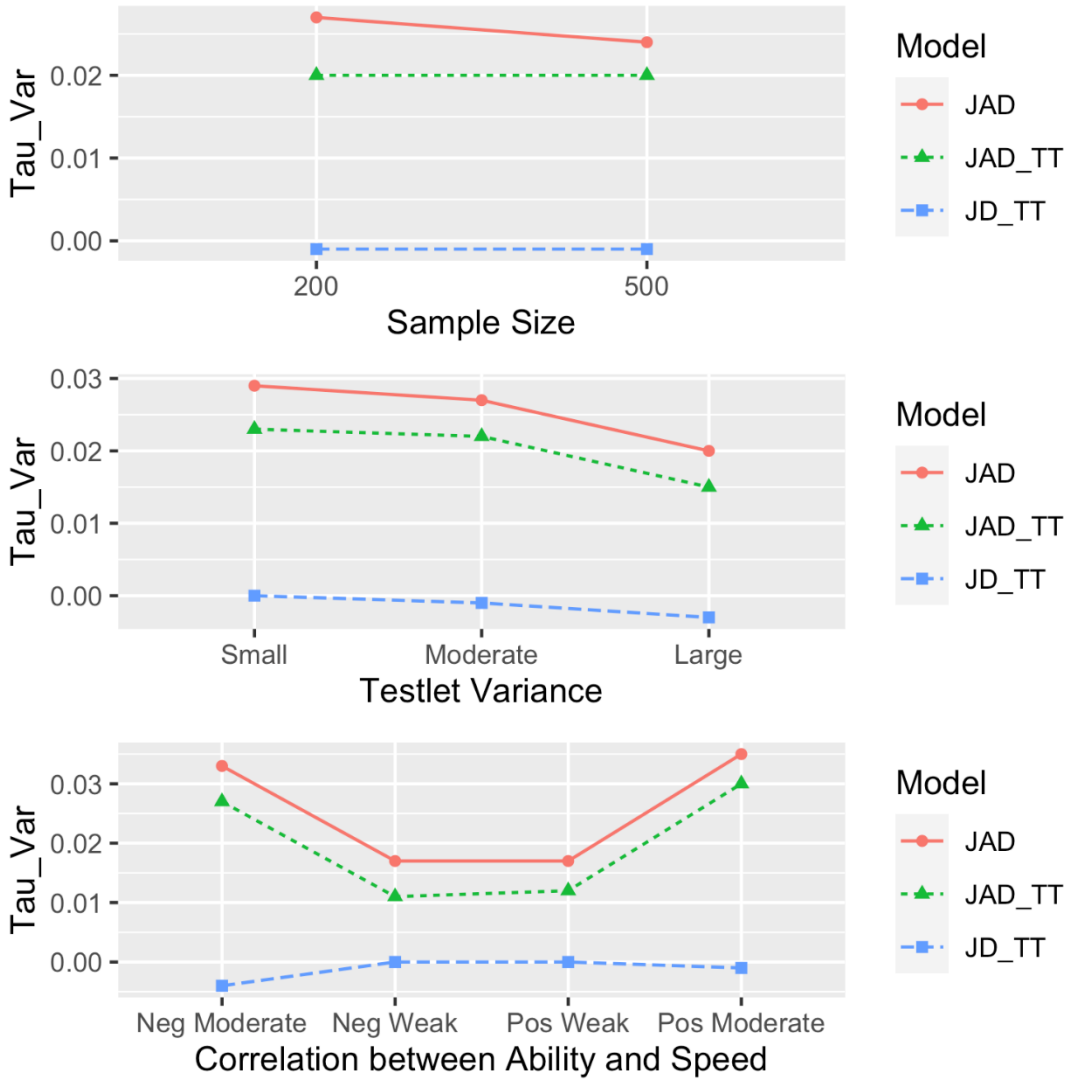

Figure 7 Marginal mean bias of the estimates of the variance of person speed τ at all levels of the manipulated factors.

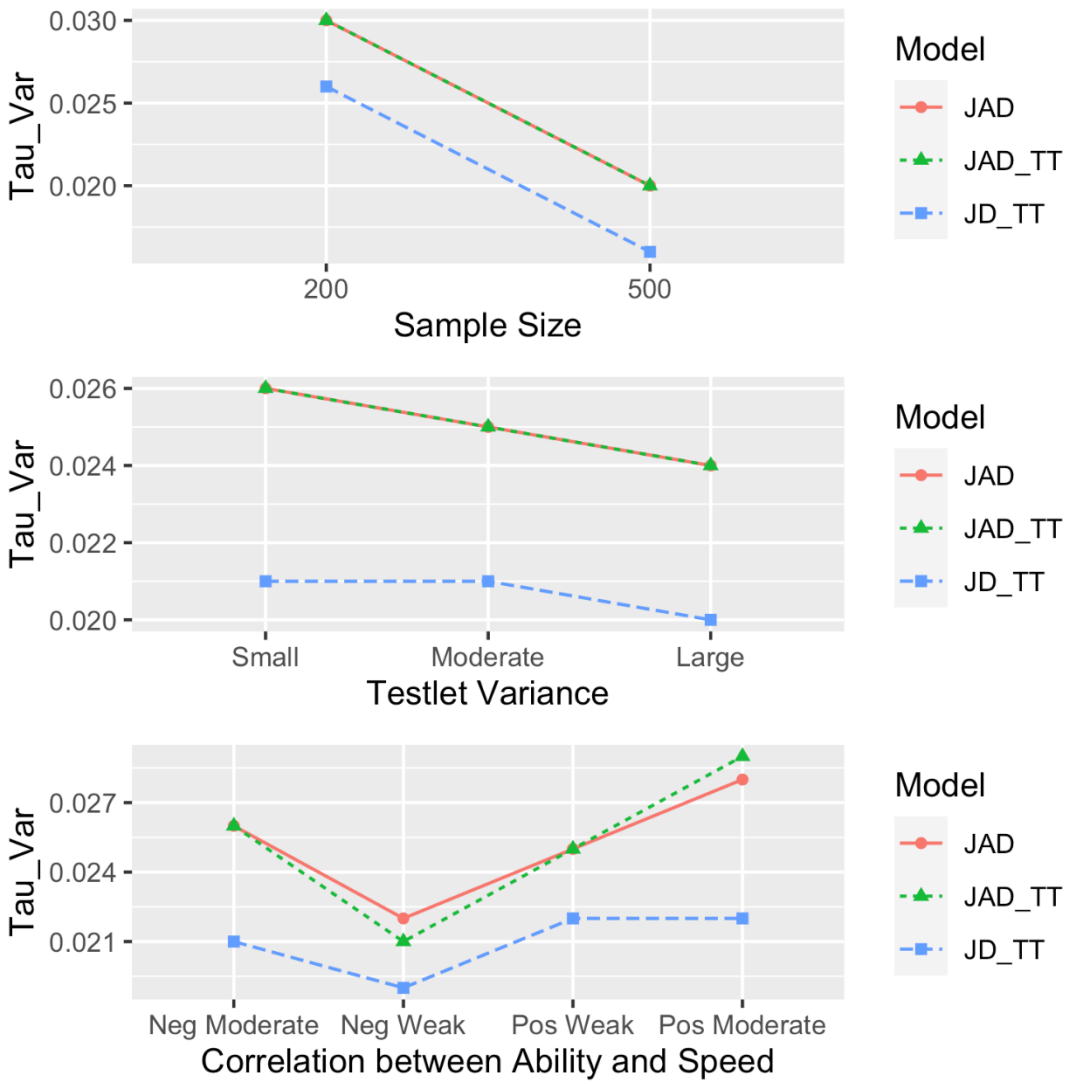

Figure 8 Marginal mean SE of the estimates of the variance of person speed τ at all levels of the manipulated factors.

Thus, based upon the marginal means, the proposed model and the JRT-DINA-R/RT/AC model both including AC patterns in the joint model of responses and RT yield comparable random error for estimating for σ_τ^2 and $\sigma_{\theta\tau}$ and comparable systematic error for estimating $\sigma_{\theta\tau}$. Compared with the Joint Testlet-DINA model, however, their biases and SEs are both larger. Thus, when testlet effects are accounted for, additionally modeling AC patterns does not necessarily lead to improved bias for estimating σ_τ^2 and SE for both σ_τ^2 and $\sigma_{\theta\tau}$.

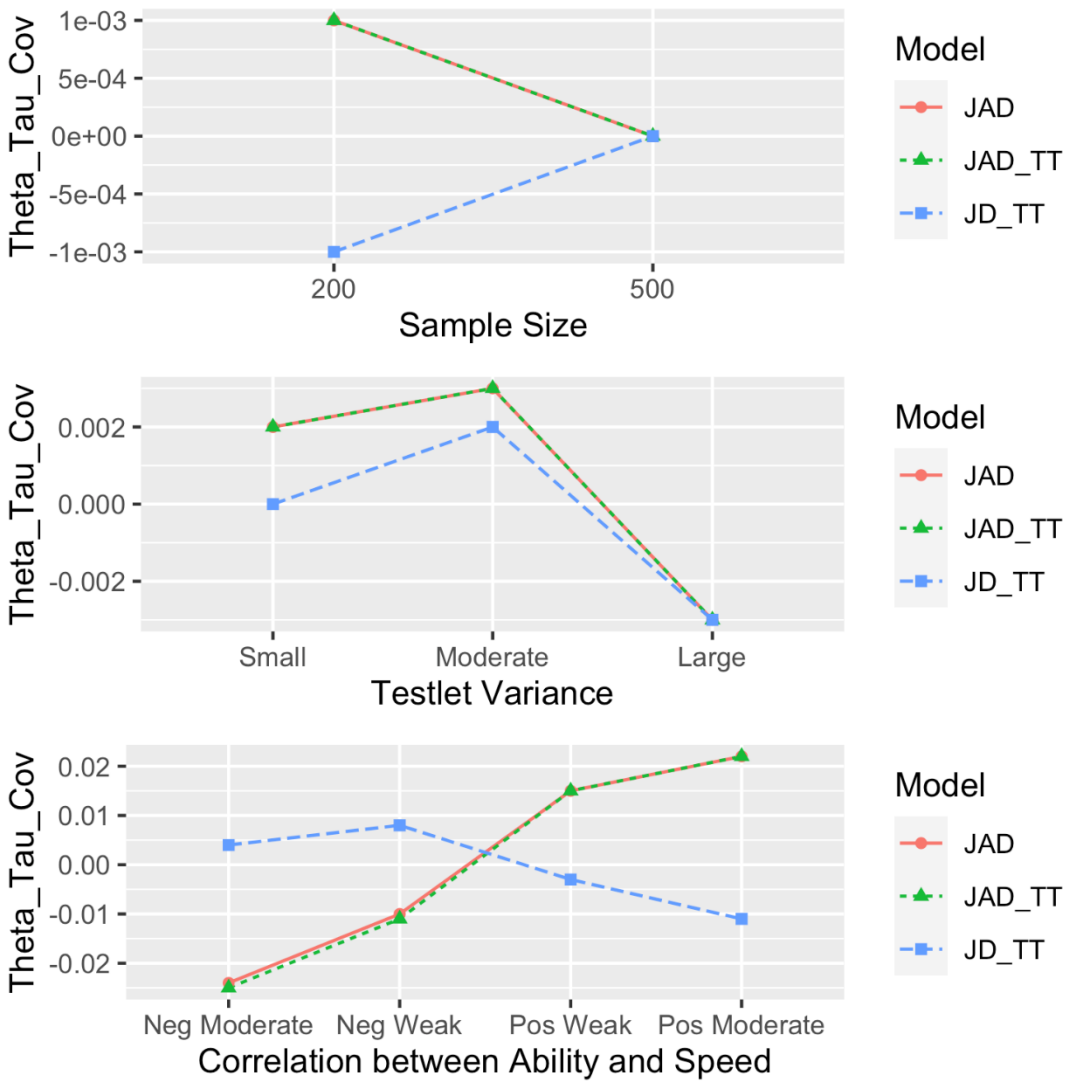

Figure 9 Marginal mean bias of the estimates of the covariance of person ability θ and person speed τ at all levels of the manipulated factors.

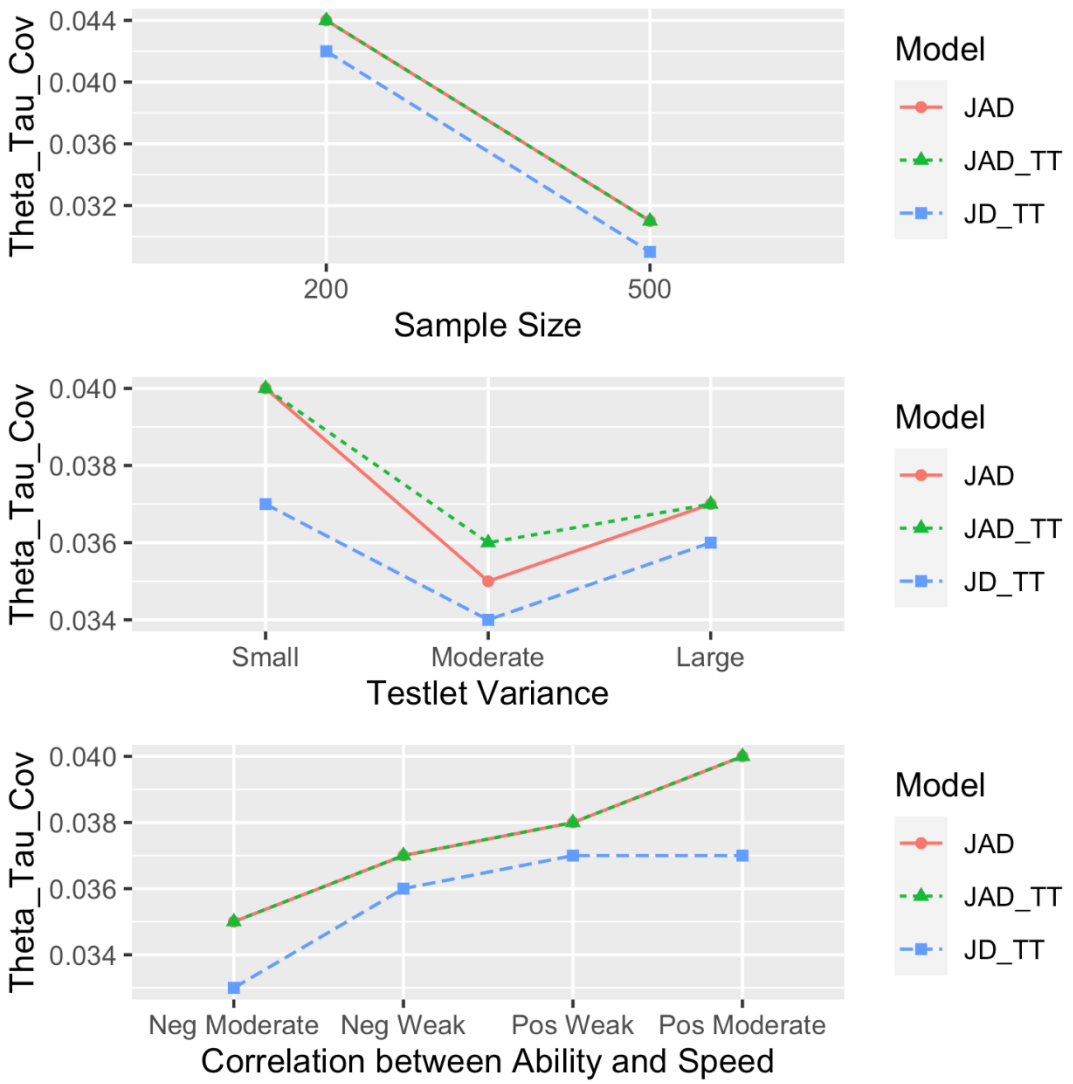

Figure 10 Marginal mean SE of the estimates of the covariance of person ability θ and person speed τ at all levels of the manipulated factors.

4.1.4 Recovery of the Higher-Order Structural Parameters

The higher-order structural parameters are the attribute easiness parameters (ι_k) and the attribute discrimination parameters (κ_k) that specify the relationship between latent ability and attribute mastery status in the higher-order structure. This section summarizes the bias and SE for ι_k and κ_k for each of the five attributes specified in the proposed model. These are summarized for each of the data-fitting models and all 24 simulated conditions and presented in Tables A.5.1-4 and A.6.1-4 in Appendix A.

Attribute Easiness. Figures 11-15 are visual representations of the marginal mean biases and SEs of ι_k for the five attributes at all levels of the manipulated factors. The impact of model specification on the marginal mean bias of the ι_k is not consistent across the attributes. The marginal mean biases of ι_k for attributes 1, 2, 4, and 5 are comparable as estimated by the proposed model and the JRT-DINA-R/RT/AC model and are the lowest as estimated by the Joint Testlet-DINA model. For attribute 3, however, they are comparable as estimated by the proposed and the Joint Testlet-DINA model and are larger as estimated by the JRT-DINA-R/RT/AC model. Further, for attribute 3 they vary by the levels of two of the manipulated factors: they are larger for larger testlet variance, and for stronger $\rho_{\theta\tau}$. For the other attributes, they do not exhibit much variation across the levels.

The impact of model specification on the marginal mean SE of ι_k are similarly inconsistent across the attributes. For attributes 1 and 5 and across the levels of the manipulated factors, the marginal mean SEs of ι_k are comparable for the proposed model and the JRT-DINA-R/RT/AC model and are the smallest as estimated by the

Joint Testlet-DINA model. For attributes 2 and 3 and across the levels of the manipulated factors, they are comparable as estimated by the proposed model and the Joint Testlet-DINA model and are the smallest as estimated by the JRT-DINA-R/RT/AC model. For attribute 4, they are similar for the proposed model and the JRT-DINA-R/RT/AC model and are smaller than the estimates generated by the Joint Tesltlet-DINA model at the levels of $\sigma_{\theta\tau}$ but are larger at the levels of sample size and testlet variance. They additionally increase as testlet variance increases from small to large. Variation by levels of $\rho_{\theta\tau}$ and testlet variance are not consistent across the attributes: attributes 4 and 5 see least variation whereas attributes 1, 2, and 3 see attribute-specific variation patterns. The only factor for which consistency across the attributes is shown is sample size: as it increases from 200 to 500, the marginal mean SEs of the attribute easiness parameters decrease.

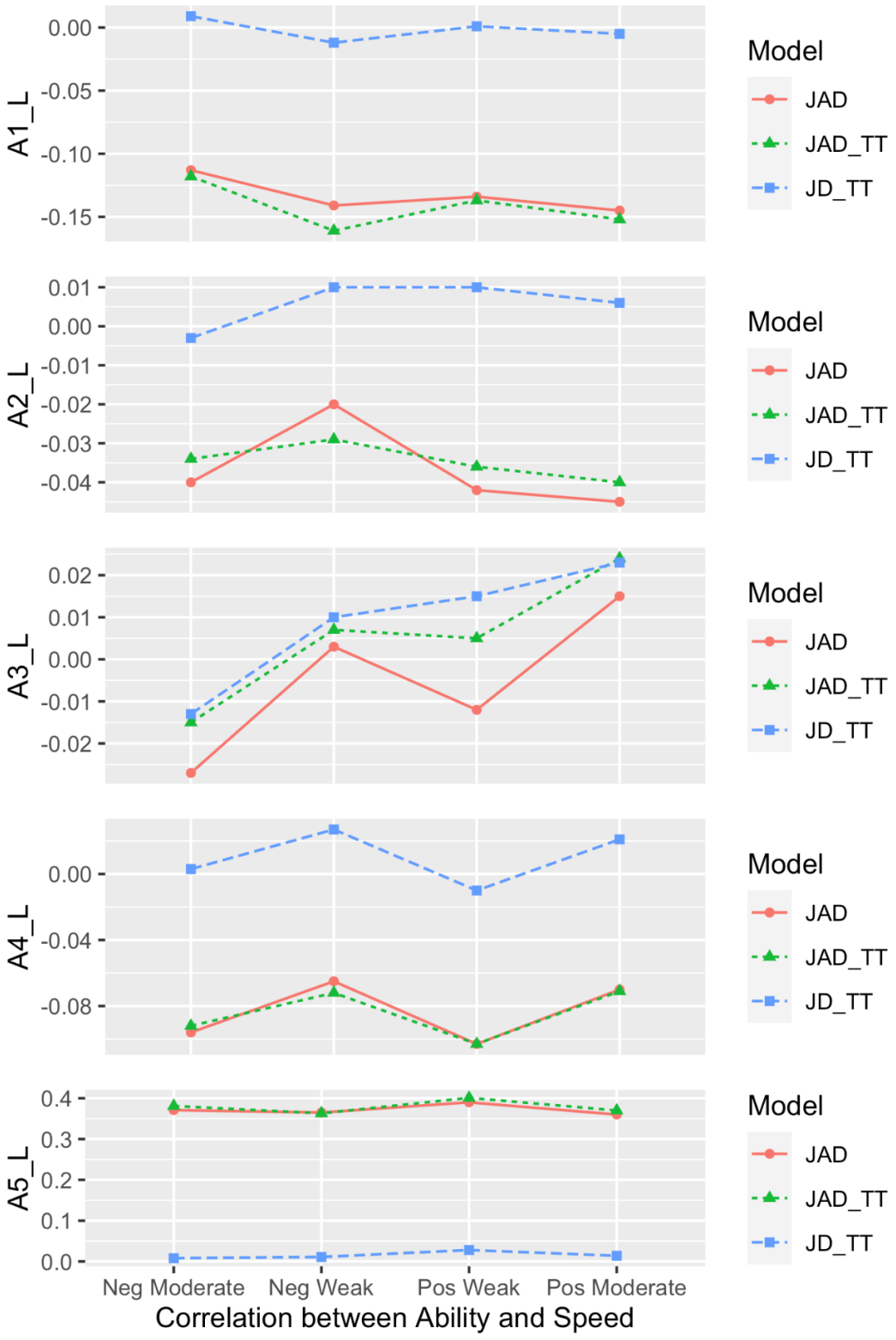

Figure 11. Marginal mean bias of the high-order attribute easiness estimates for each of the five attributes at the four correlation levels.

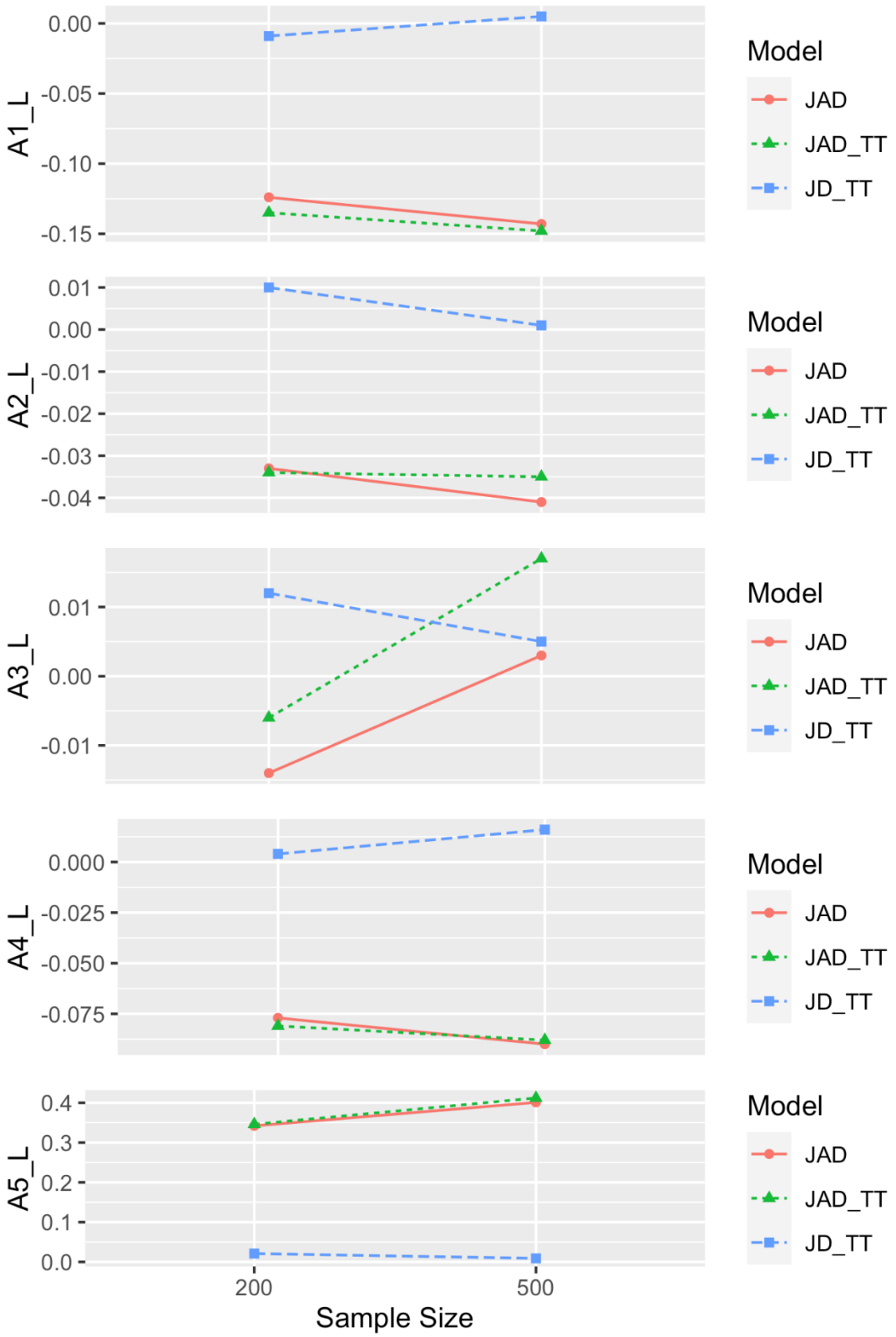

Figure 12. Marginal mean bias of the high-order attribute easiness estimates for each of the five attributes at the two sample sizes.

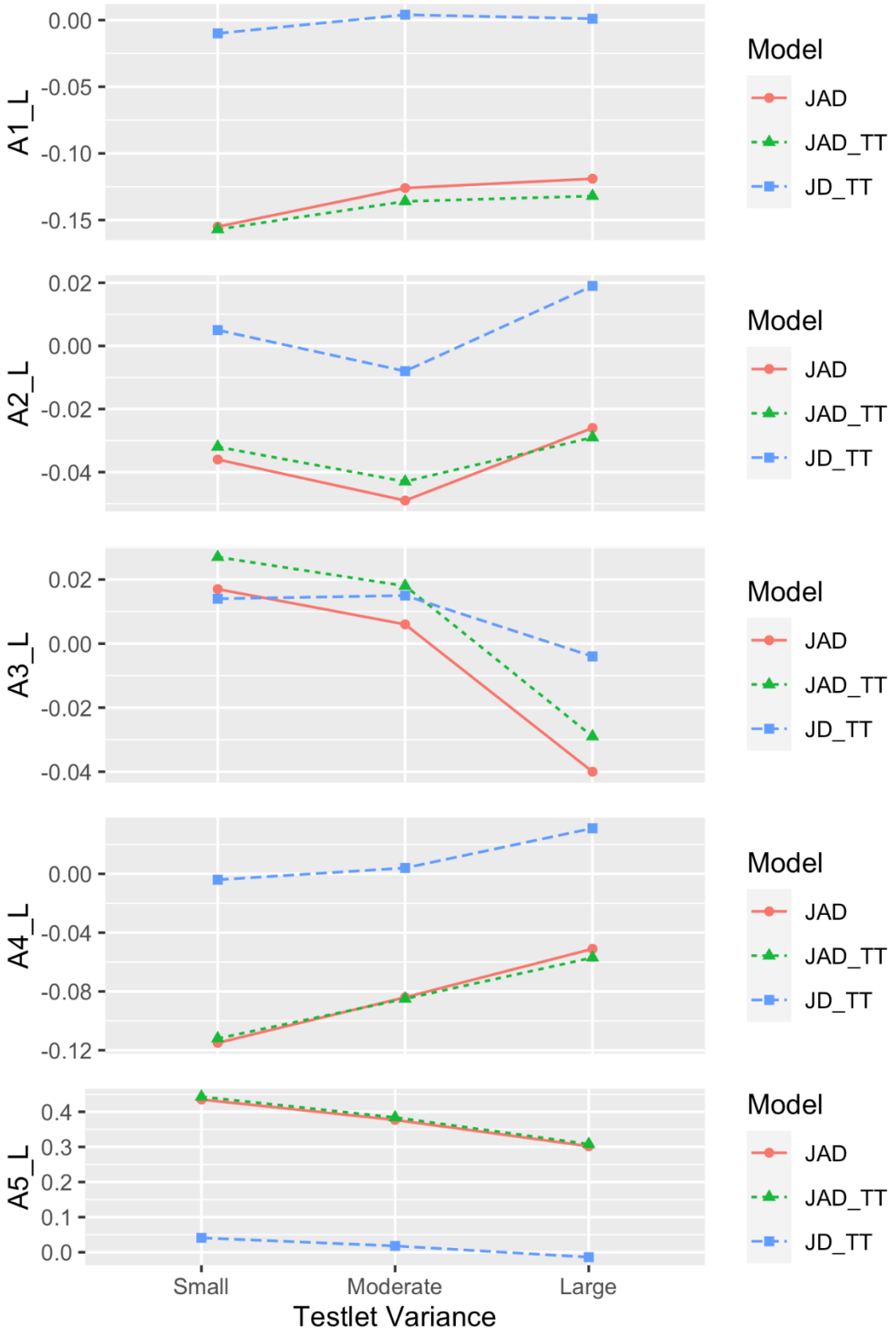

Figure 13. Marginal mean bias of the high-order attribute easiness estimates for each of the five attributes at the three testlet variance levels.

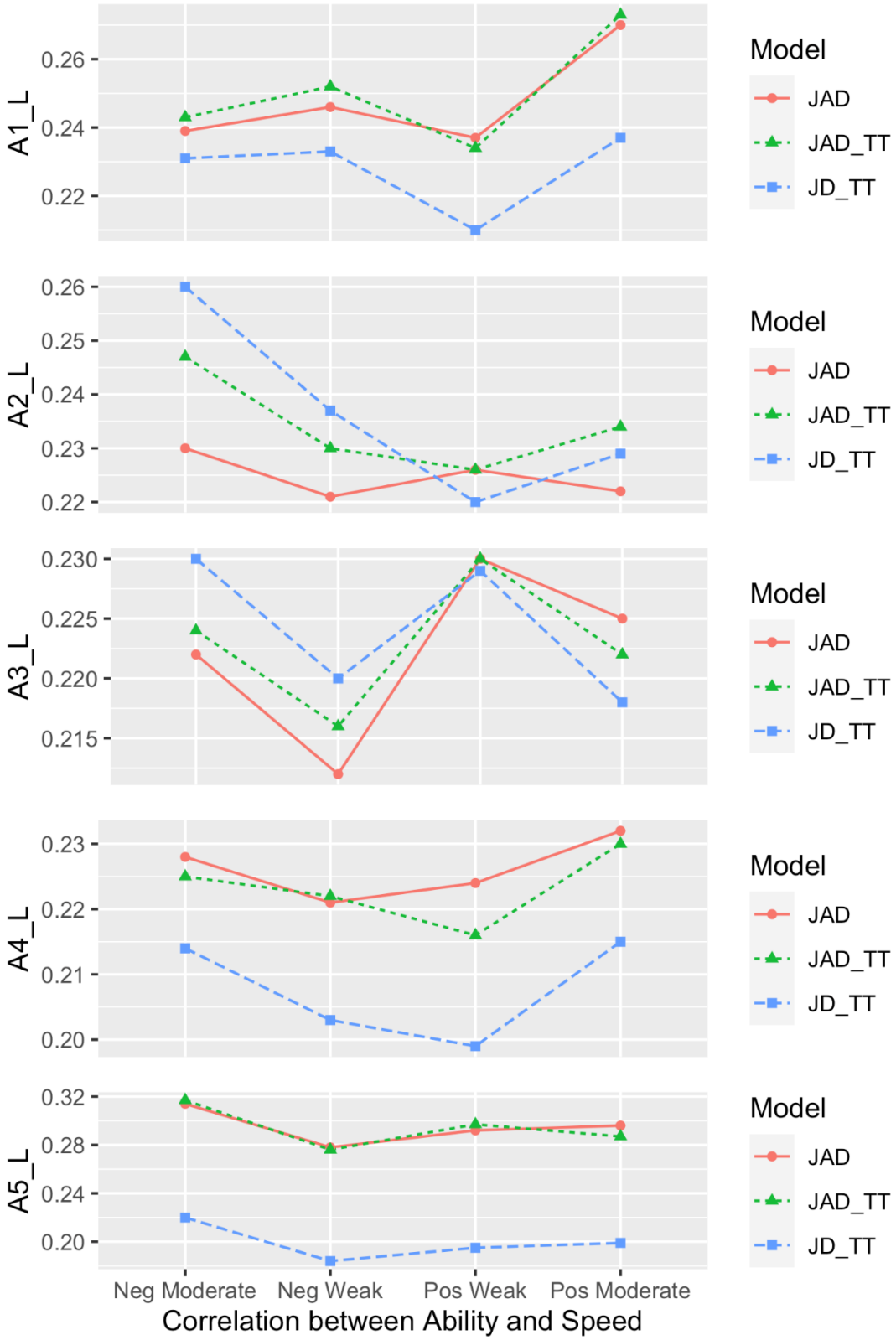

Figure 14 Marginal mean SE of the high-order attribute easiness estimates for each of the five attributes at the four correlation levels.

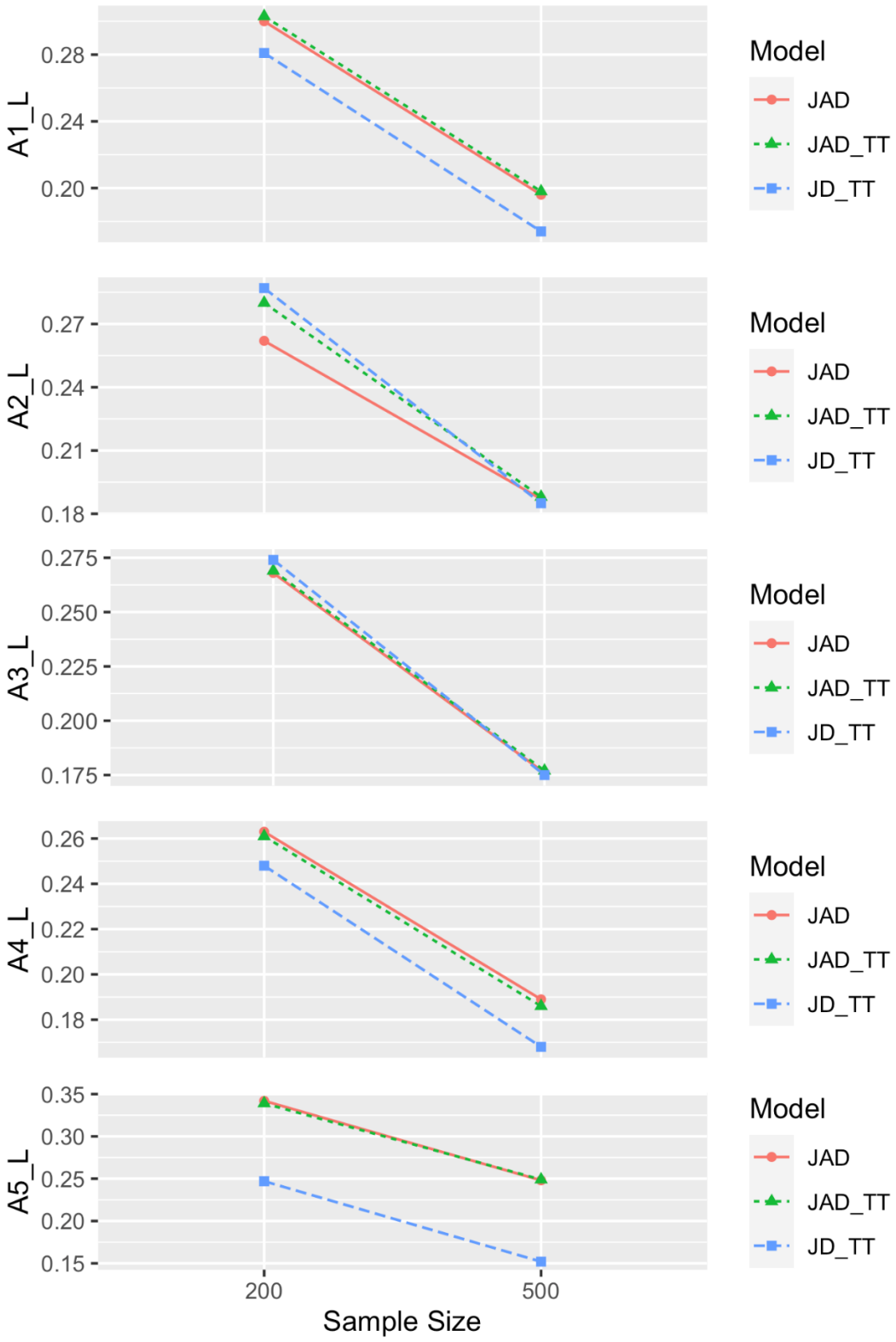

Figure 15. Marginal mean SE of the high-order attribute easiness estimates for each of the five attributes at two sample size levels.

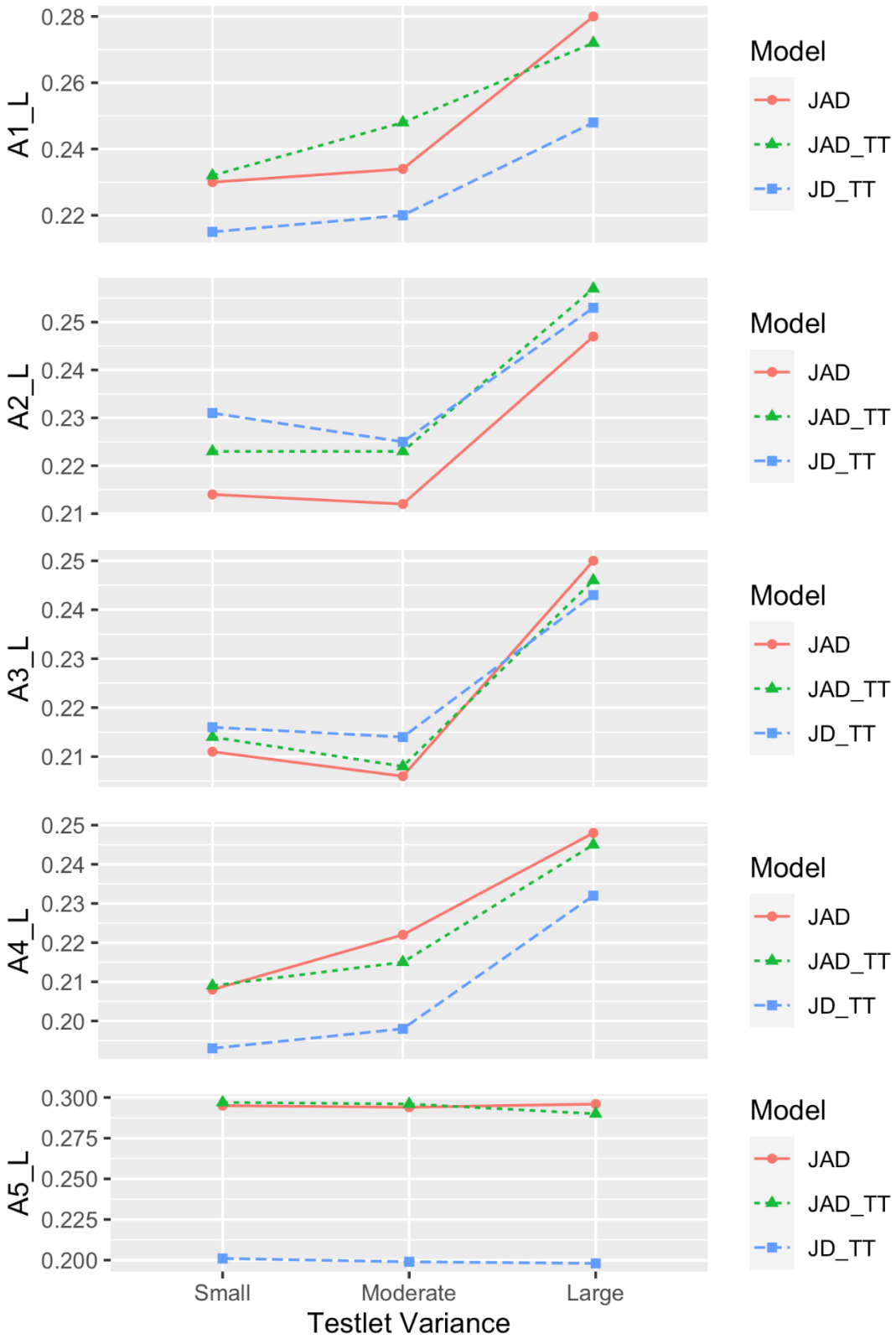

Figure 16 Marginal mean SE of the high-order attribute easiness estimates for each of the five attributes at three testlet variance levels.

Attribute Discrimination. Figures 17 through 22 display the biases and SEs of the high-order discrimination estimates for the five attributes at all levels of the manipulated factors. Consistent across the levels of the manipulated factors and the five attributes, marginal mean biases of the estimates generated by the proposed model and the JRT-DINA-R/RT/AC model are comparable and are larger than those generated by the Joint Testlet-DINA model. They additionally show little variation by levels of the sample size and $\rho_{\theta\tau}$, suggesting that the two factors have little impact on them. Increased testlet variance corresponds to increased marginal mean biases for two attributes: attribute 2 and 3. Attributes 1, 4, and 5 show little variation by the levels of this factor.

The marginal mean SEs of the estimates of the high-order discrimination parameters, however, show evident variation across the attributes and by levels of the manipulated factors. Similar to patterns observed for the marginal mean bias of the discrimination estimates, SEs of the estimates are comparable for the proposed model and the JRT-DINA-R/RT/AC model across the levels of the manipulated factors. For attribute 5 they are larger than the estimates generated by the Joint Testlet-DINA model. For attributes 1 and 4, they are larger than those yielded by the Joint Testlet-DINA model at specific levels of the testlet variance and $\rho_{\theta\tau}$: at negative $\rho_{\theta\tau}$ or small testlet variance for attribute 4 and at positive moderate $\rho_{\theta\tau}$ or moderate testlet variance for attribute 1. At the other levels of $\rho_{\theta\tau}$ or of testlet variance, the proposed model and the JRT-DINA-R/RT/AC model yield reduced SEs compared with the Joint Testlet-DINA model. Marginal mean SEs for attributes 2 and 3 are consistently

smaller for estimates generated by the proposed model and the JRT-DINA-R/RT/AC model than for those generated by the Joint Testlet-DINA model.

The only manipulated factor that sees consistency across attributes is sample size: SEs are consistently smaller at the sample size of 500 than at 200. Variation by the levels of the other two manipulated factors is attribute- and model-specific. For instance, attribute 5 sees least variation in the marginal mean SEs by the levels of $\rho_{\theta\tau}$ or testlet variance whereas for attribute 3 and across the models, its marginal mean SEs increase as testlet variance increases.

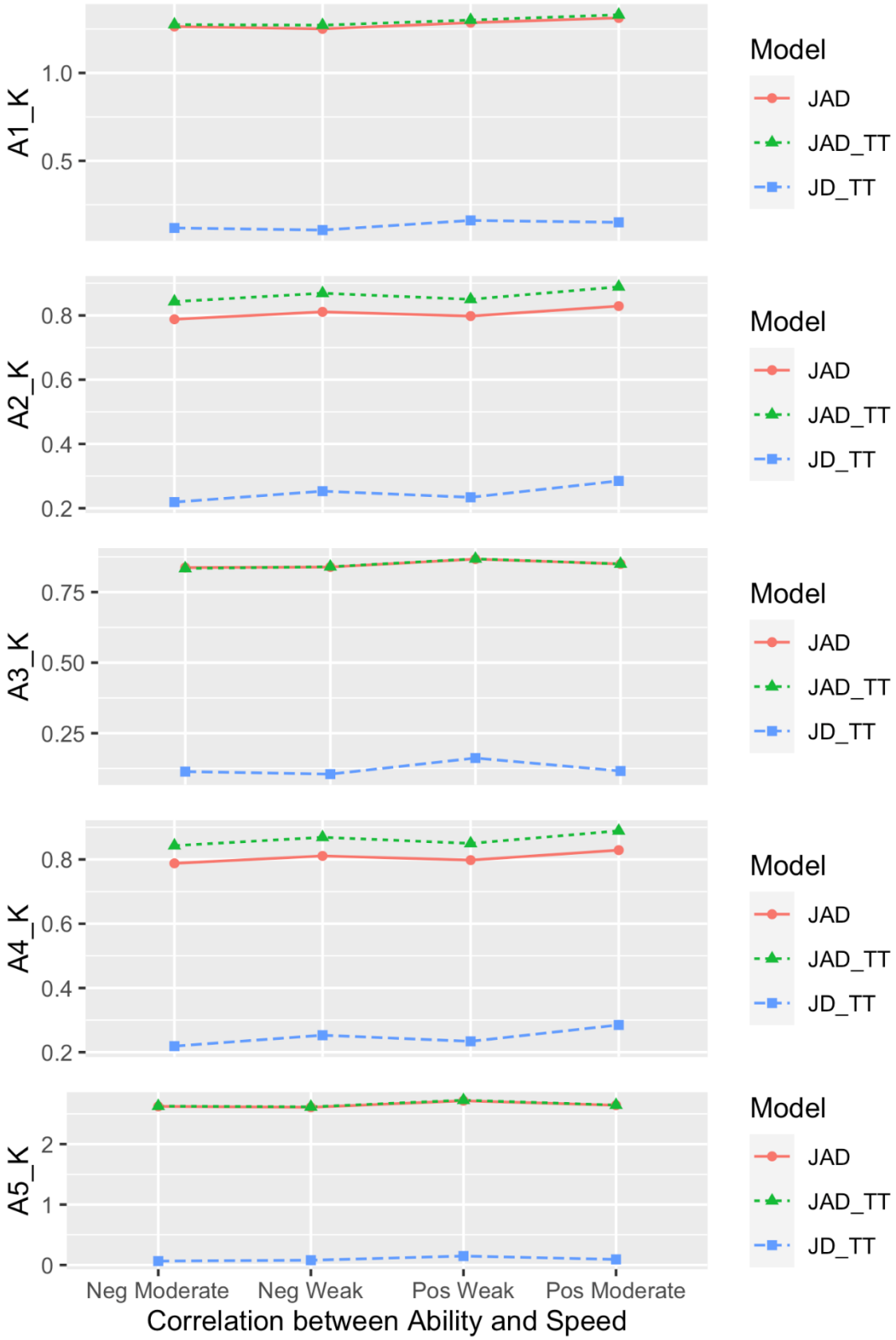

Figure 17. Marginal mean bias of the high-order attribute discrimination estimates for each of the five attributes at four correlation levels.

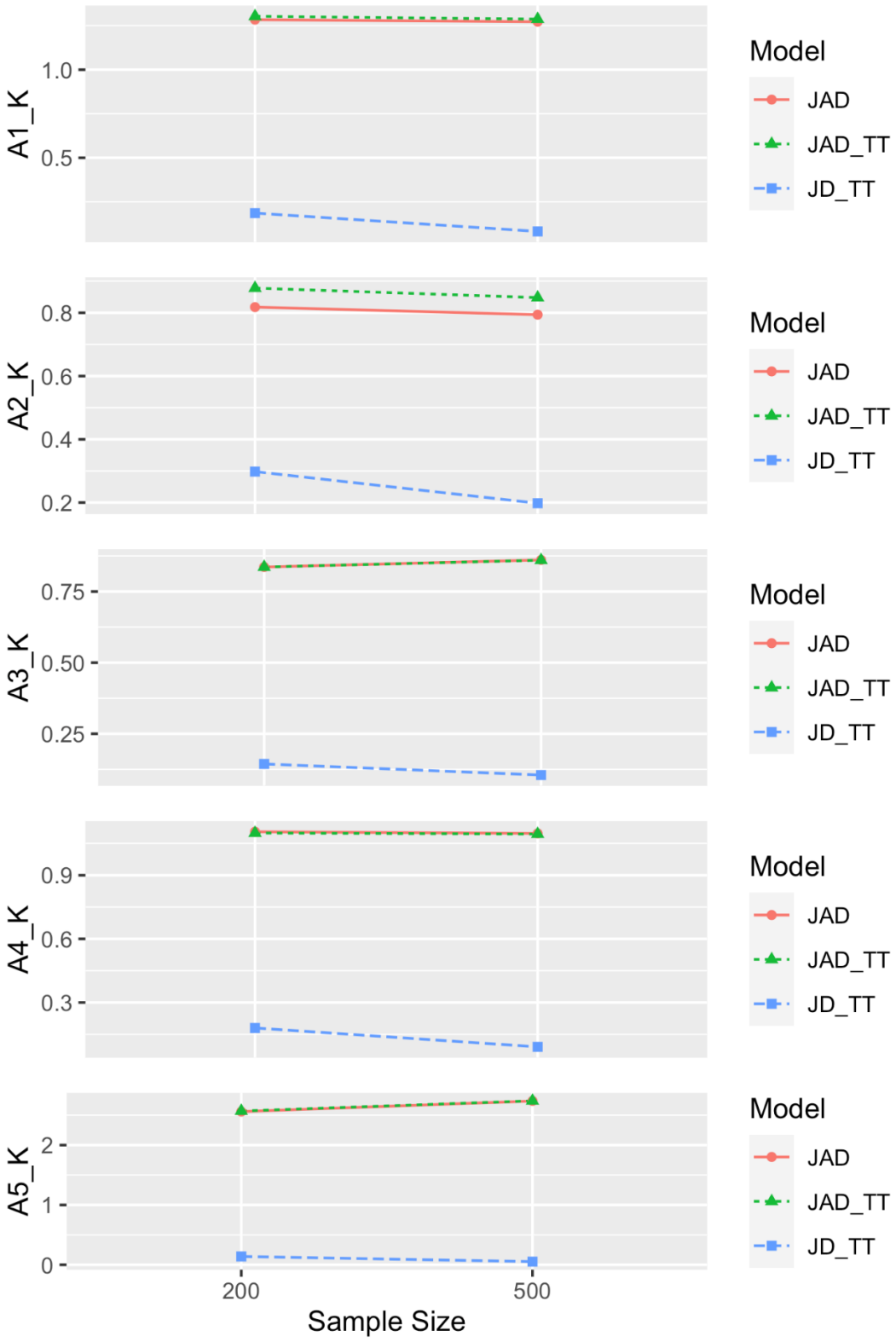

Figure 18. Marginal mean bias of the high-order attribute discrimination estimates for each of the five attributes at the two sample sizes.

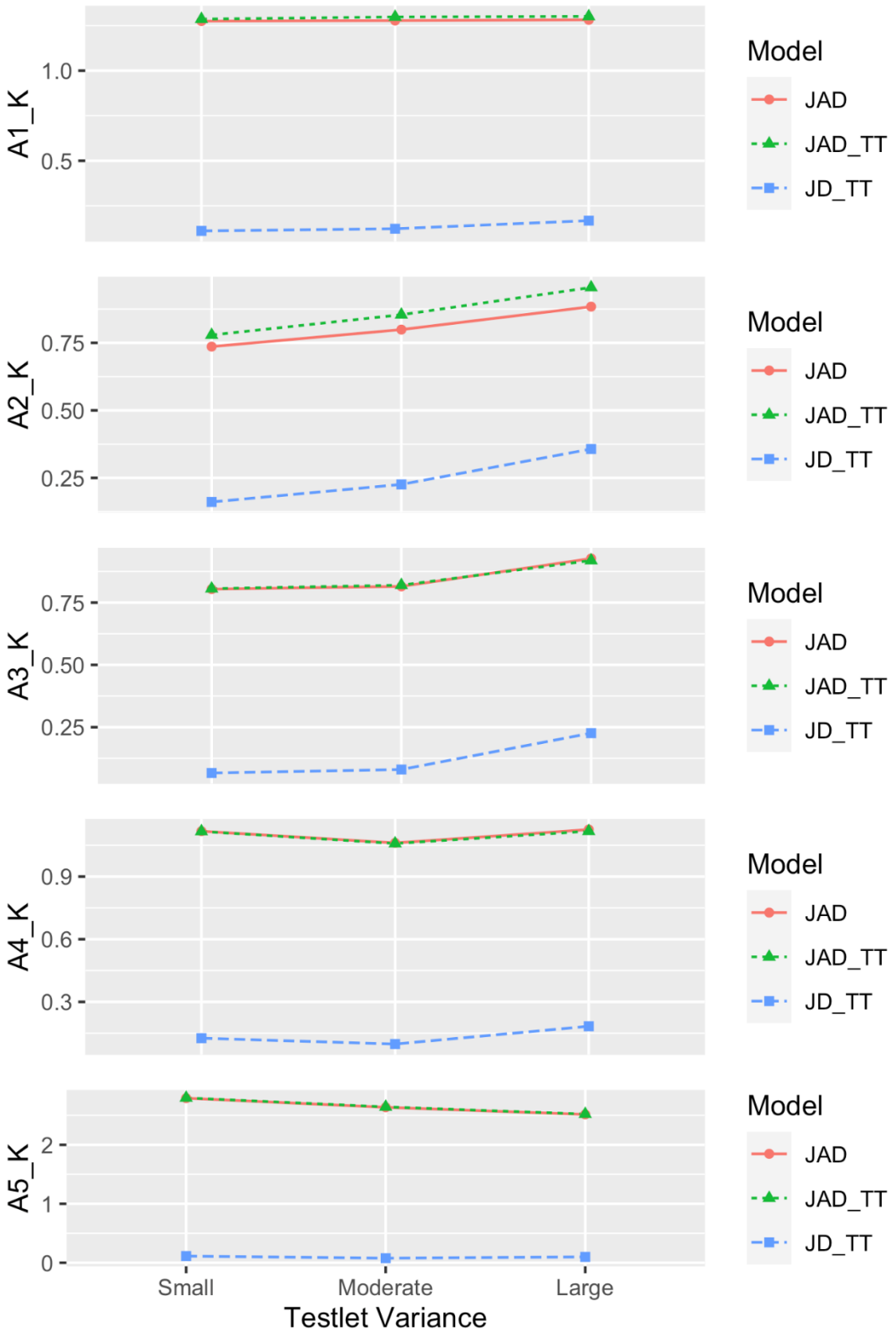

Figure 19. Marginal mean bias of the high-order attribute discrimination estimates for each of the five attributes at three testlet variance levels.

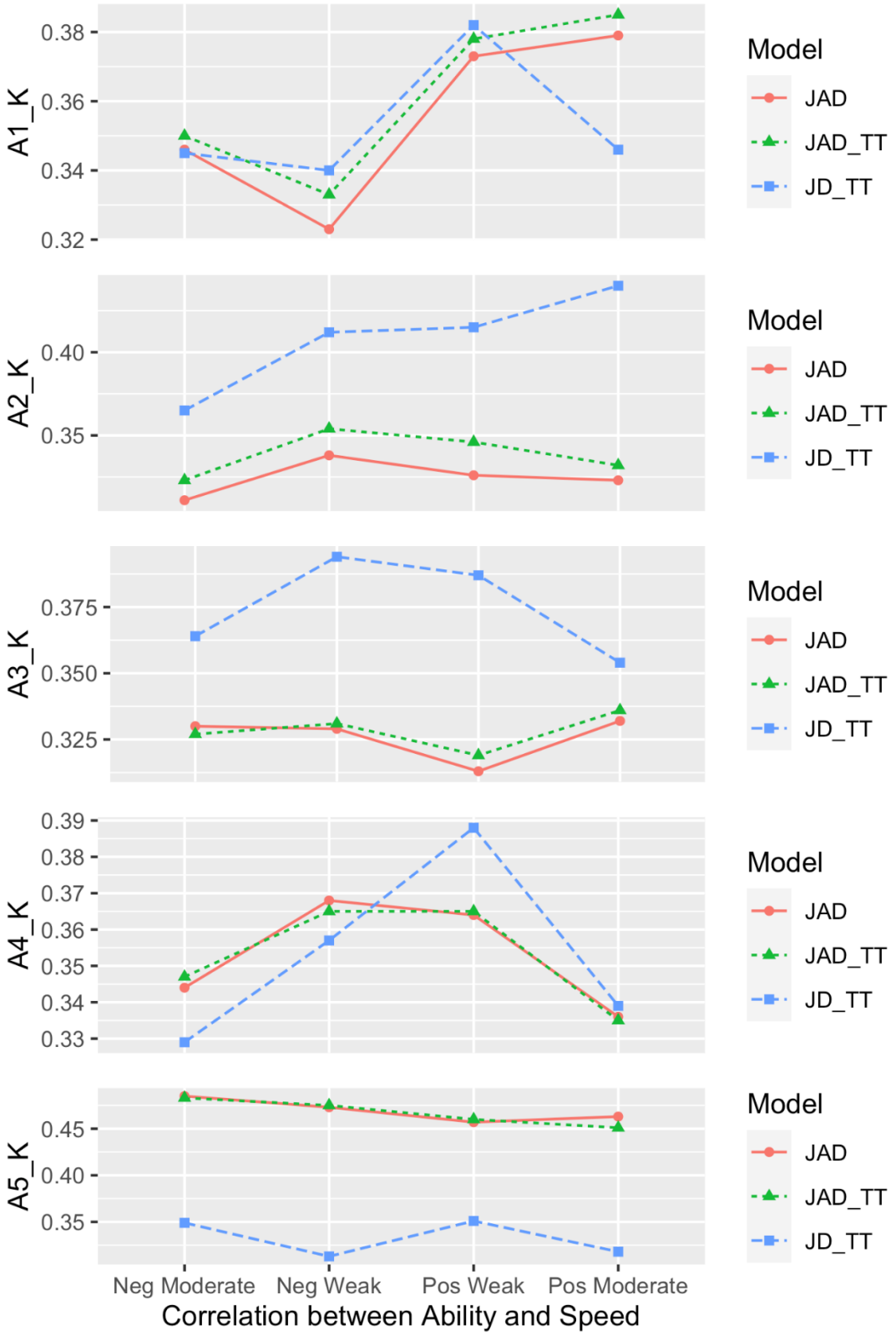

Figure 20. Marginal mean SE of the high-order attribute discrimination estimates for each of the five attributes at the four correlation levels.

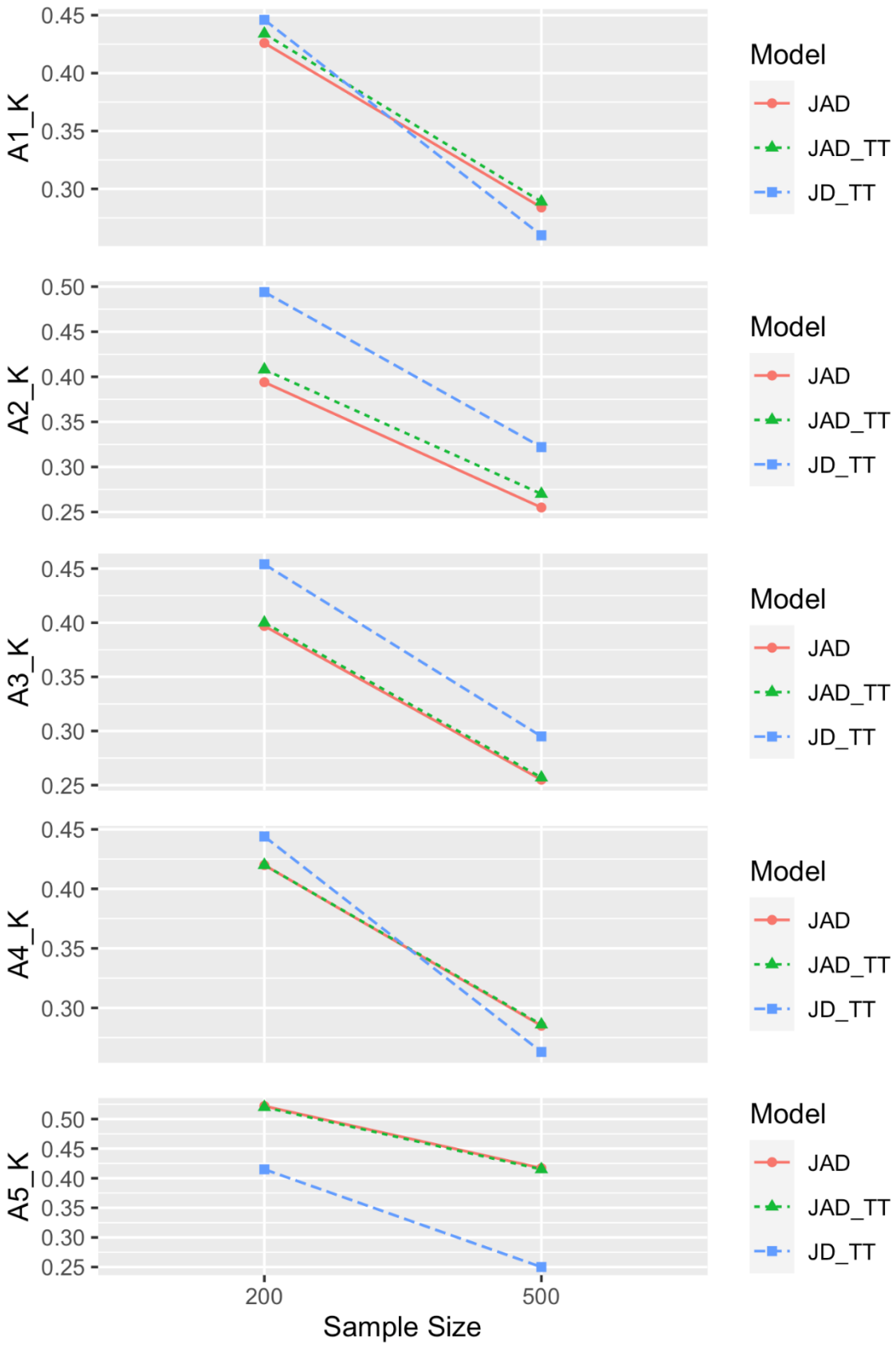

Figure 21 Marginal mean SE of the high-order attribute discrimination estimates for each of the five attributes at two sample size levels.

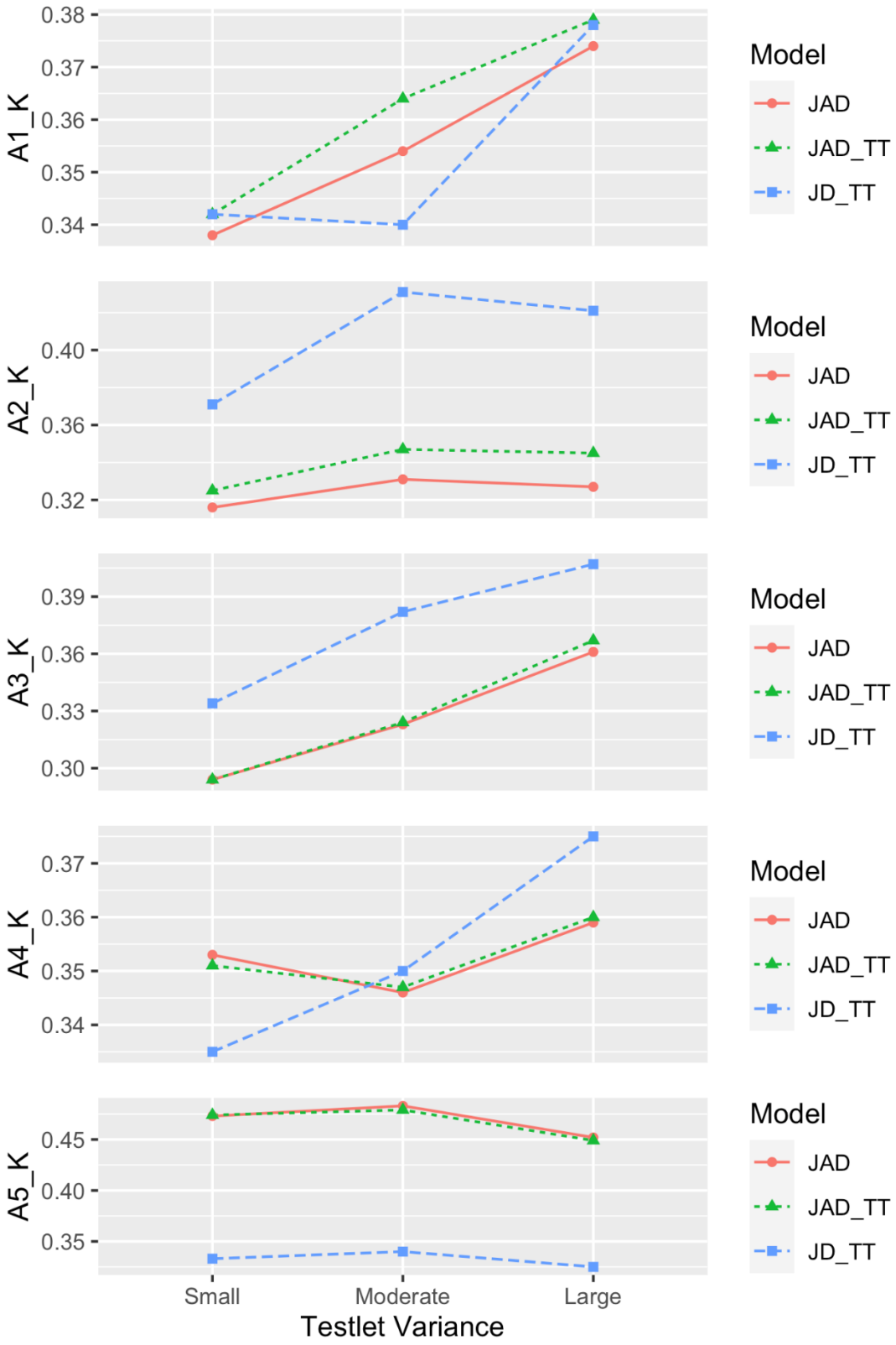

Figure 22 Marginal mean SE of the high-order attribute discrimination estimates for each of the five attributes at the three testlet variance levels.

4.1.5 Recovery of the Item Parameters

Item parameters evaluated in this study are β_j, the item intercept parameter, δ_j, the item interaction parameter, ζ_j the item time intensity parameter, and b_j, the item difficulty parameter. To evaluate the degree to which the proposed model successfully recovers these parameters as compared to the other two models, this section presents and summarizes the correlation between the true item parameters and their estimates generated by each model. Mixed effect ANOVA results on the bias and SE of β_j, δ_j, and ζ_j are presented to examine the extent to which the three models and the three manipulated factors affect the recovery of these parameters. The ANOVA results do not include b_j as it is not one of the parameters in the Joint Testlet-DINA model.

Correlation between true and estimated item parameters. Table A.7 presents the correlation between true and estimated person parameters for each model under the 24 simulated conditions. As is shown in in the table, correlation for the proposed model is ≥ 0.90 for β_j, ≥ 0.80 for δ_j, close to 1 for ζ_j, and ≥ 0.70 for b_j, suggesting that estimates for these parameters yielded by proposed model correlate well with the true parameters. The correlations for b_j and δ_j are weaker than the correlations for β_j and ζ_j.

The correlations for the three models are identical or comparable under the 24 simulated conditions. Correlation for ζ_j, the item time intensity parameter, is identical for the three models under all simulated conditions. Slight differences in the correlation for β_j, δ_j, and b_j are observed between the proposed model and JRT-

DINA-R/RT/AC. In 14 out of the 24 simulated conditions, the correlation for β_j is slightly stronger for the proposed model. All 14 conditions have a testlet variance of 0.5 or 1. In 15 out of the 24 conditions the correlation for δ_j is slightly stronger for the proposed model. Examples are conditions 7, 8, and 9, all having a sample size of 200 and a value of 0.3 for $\rho_{\theta\tau}$, and conditions 13, 14, and 15 with a sample size of 500 and a value of -0.5 for $\rho_{\theta\tau}$. Correlations for b_j for the proposed model and the JRT-DINA-R/RT/AC are comparable, although in 7 out of the 24 conditions, its value is slightly stronger for the proposed model. Compared with the Joint Testlet-DINA model, however, the proposed model yields a weaker correlation for β_j and δ_j under most of the simulated conditions. Only in a couple of simulated conditions are the correlations for the proposed model stronger: conditions 7 for β_j and conditions 12 and 18 for δ_j.

Thus, while under specific conditions, accounting for testlet effects in addition to modeling the answer change patterns can lead to slight improvement in the correlation between true and estimated β_j, δ_j, and ζ_j, when the testlets effects are accounted for, as in the Joint Testlet-DINA model, additionally modeling AC patterns in the joint model of response and RT does not result in a stronger correlation for the three item parameters. Further, correlation for ζ_j is identical for the three models being compared, indicating no impact of the modeling approach on the estimation of this parameter.

Item Intercept Parameter. Table 14 summarizes highest-order significant effects on the recovery of item parameters. The four-way highest-order interaction of model and the three manipulated factors: sample size, $\rho_{\theta\tau}$, and testlet variance, has a

significant effect on the bias of ζ_j with a large effect size of 0.774, and the SE of ζ_j and δ_j with a medium and small effect size. The interaction of sample size, $\rho_{\theta\tau}$, and testlet variance has a significant effect on the bias of β_j and δ_j , both having a small effect size, whereas the interaction of model, $\rho_{\theta\tau}$, and testlet variance and of model, sample size, and testlet variance has a significant effect on the SE of β_j and the bias of δ_j, similarly having a small effect size.

Table 14
Summary of Effect Sizes of the Highest-Order Significant Effects from the Mixed- Effect ANOVA on the Recovery of Item Parameters

Effect	β_j		δ_j		ζ_j	
	Bias	*SE*	*Bias*	*SE*	*Bias*	*SE*
Model						
Model*Sample Size*Testlet Variance		0.013	0.011			
Model*Correlation*Testlet Variance		0.023	0.031			
Model*Sample Size*Correlation*Testlet Variance				0.024	0.774	0.222
Sample Size*Correlation*Testlet Variance	0.042		0.025			

Note: Effect Size is classified as follows: Small (0.01≤partial η^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

Table 15 summarizes the significant main and interaction effects on the recovery of β_j identified in the mixed-effect ANOVA. Model, sample size, and testlet variance all have a significant main effect on the bias and SE of β_j. Significant two-way interaction effects on the bias and SE of β_j stem from the interaction of model and sample size and of model and testlet variance. Two three-way interaction of model, sample size, and $\rho_{\theta\tau}$, and of model, $\rho_{\theta\tau}$, and testlet variance significantly affect the SE of β_j. Significant three-way interaction effect on the bias of β_j is from the interaction of sample size, $\rho_{\theta\tau}$, and testlet variance.

Table 15
Effect Sizes in the Mixed-Effect ANOVA Results of the Bias and SE of the Item Intercept Estimates

Source	Bias of β_j			SE of β_j		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment						
Model	8900.325	0.000	0.939	3536.772	0.000	0.860
Model* Sample Size	98.651	<.001	0.146	27.601	<0.001	0.046
Model*Testlet Variance	52.311	<.001	0.154	32.207	<0.001	0.101
Model*Sample Size*Correlation				2.537	0.032	0.013
Model*Correlation*Testlet Variance				2.258	0.018	0.023
Between-Subject Effects						
Sample Size	6.588	0.011	0.011	8.538	0.004	0.015
Testlet Variance	172.400	<.001	0.374	33.001	<0.001	0.103
Sample Size*Correlation*Testlet Variance	4.252	<.001	0.042			

Note: Effect Size is classified as follows: Small (0.01≤partial η^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

Figure 23 displays the highest-order interaction effect of $\rho_{\theta\tau}$ and testlet variance on the marginal mean bias of β_j as estimated by the proposed model at the two sample sizes. Consistent across the two sample sizes, bias is the lowest at the level of small testlet variance, increases at the level of moderate testlet variance, and is the highest when testlet variance is large. Variation by $\rho_{\theta\tau}$ is specific to the testlet variance level and the sample sizes. At the sample size of 200, for example, the lowest mean bias for the levels of small and moderate testlet variance is at positive weak $\rho_{\theta\tau}$. At 500, the lowest mean bias for the same testlet variance levels is at weak $\rho_{\theta\tau}$.

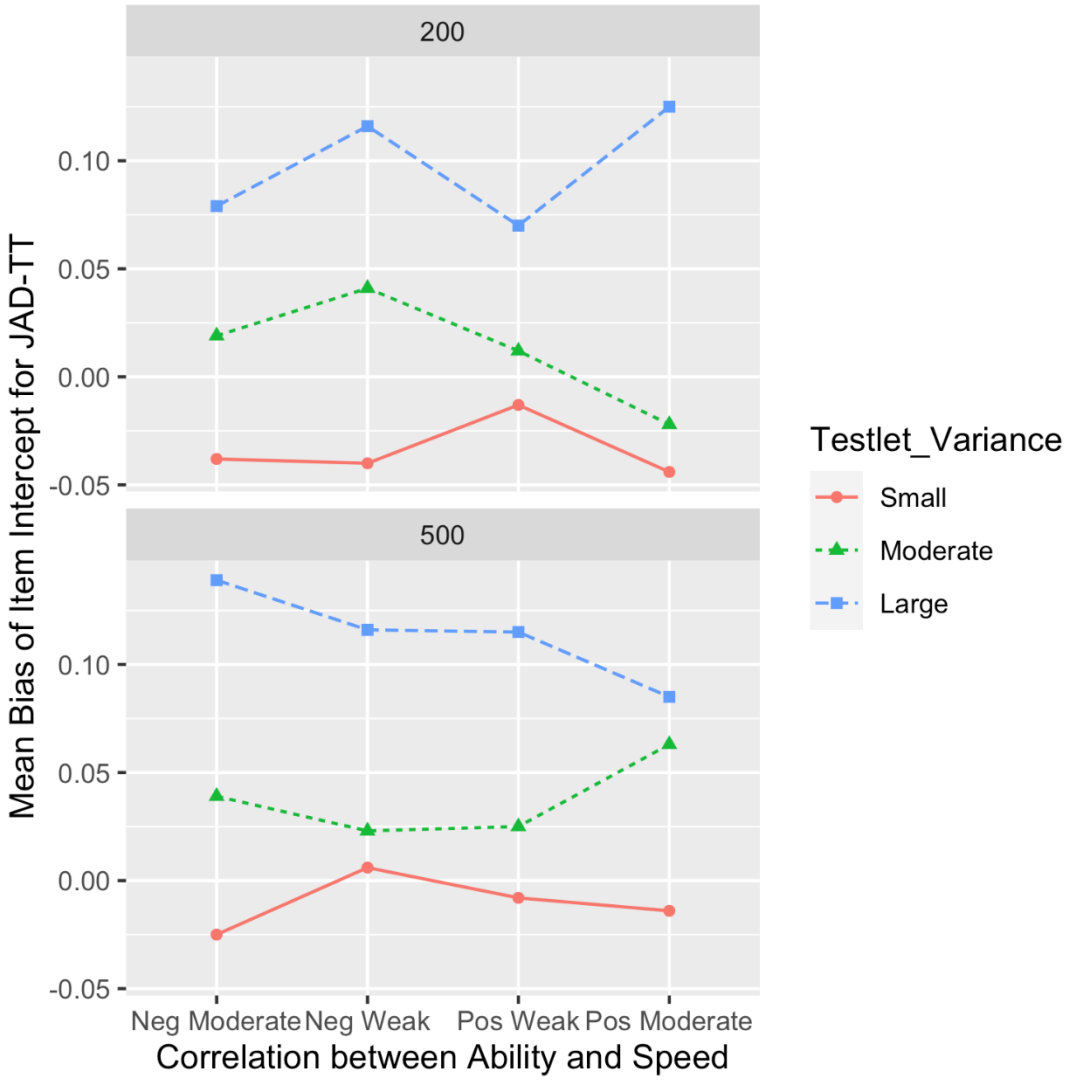

Figure 23 Significant three-way interaction of sample size, correlation, and testlet variance on the mean bias of β_j for the proposed model.

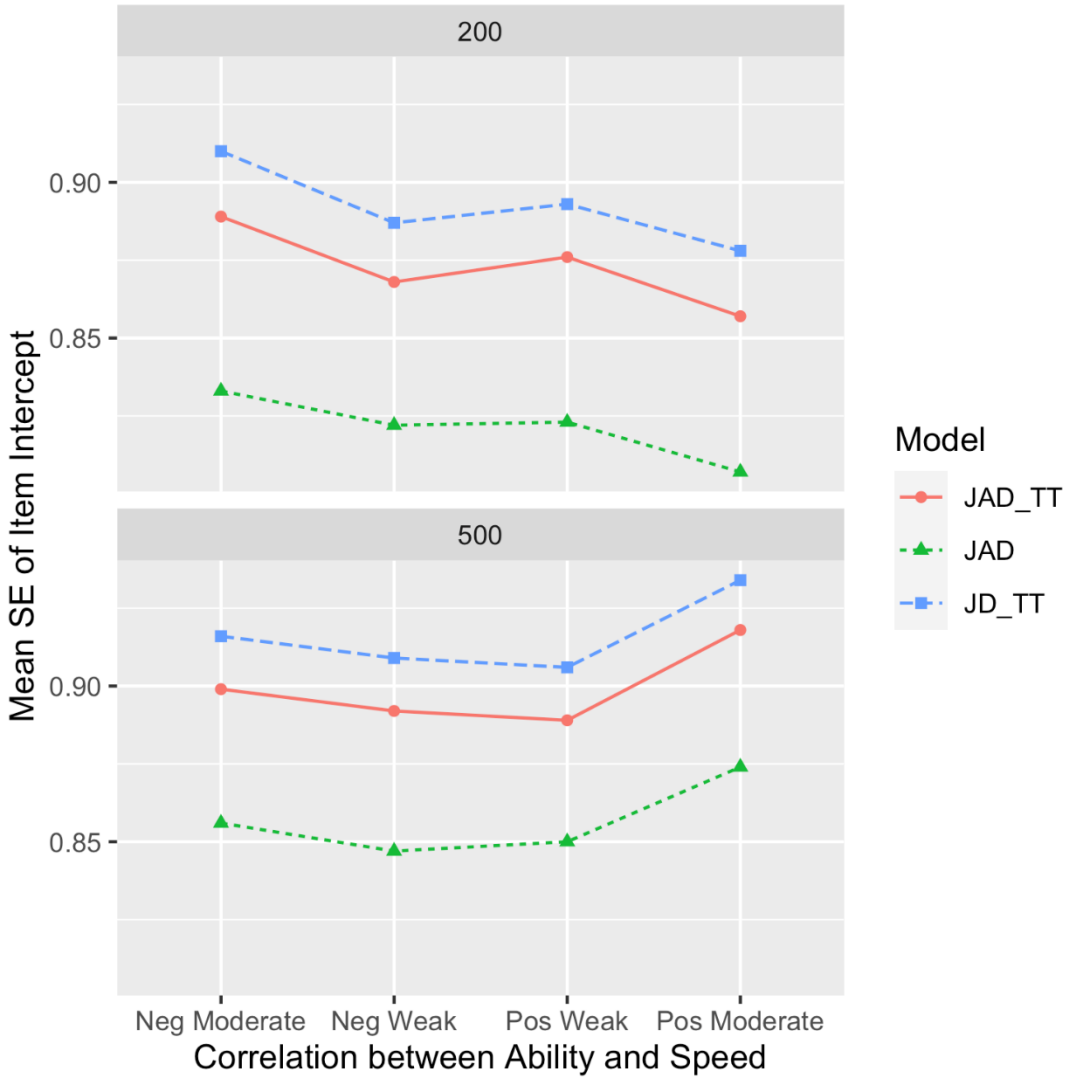

Figure 24 Significant three-way interaction of the correlation between θ_i and τ_i and model type on the mean SE for β_j for the sample size of 200 and 500.

Tables 24 and 25 are visual representations of the interaction effects of model type and $\rho_{\theta\tau}$ at two sample sizes and of the interaction of model type and testlet variance at four levels of $\rho_{\theta\tau}$ on the marginal mean SE of β_j. Consistent across the levels, marginal mean SE for β_j as estimated by the proposed model and the Joint Testllet-DINA model are larger than by the JRT-DINA-R/RT/AC model. Further, compared to the Joint Testlet-DINA model, the proposed model yielded smaller marginal mean SEs for β_j. Variation by $\rho_{\theta\tau}$ is inconsistent across the two sample

sizes. At the sample size level of 500 and across the models, marginal mean SEs for β_j are smaller for weak $\rho_{\theta\tau}$ than for moderate $\rho_{\theta\tau}$. At the sample size level of 200, marginal mean SEs for β_j are the smallest for positive moderate $\rho_{\theta\tau}$ and are comparable for positive weak and negative weak $\rho_{\theta\tau}$. Further, as is shown in Figure 25, as testlet variance increases, across the correlation levels and the models, the marginal mean SEs of β_j decrease and are the lowest at the large testlet variance level.

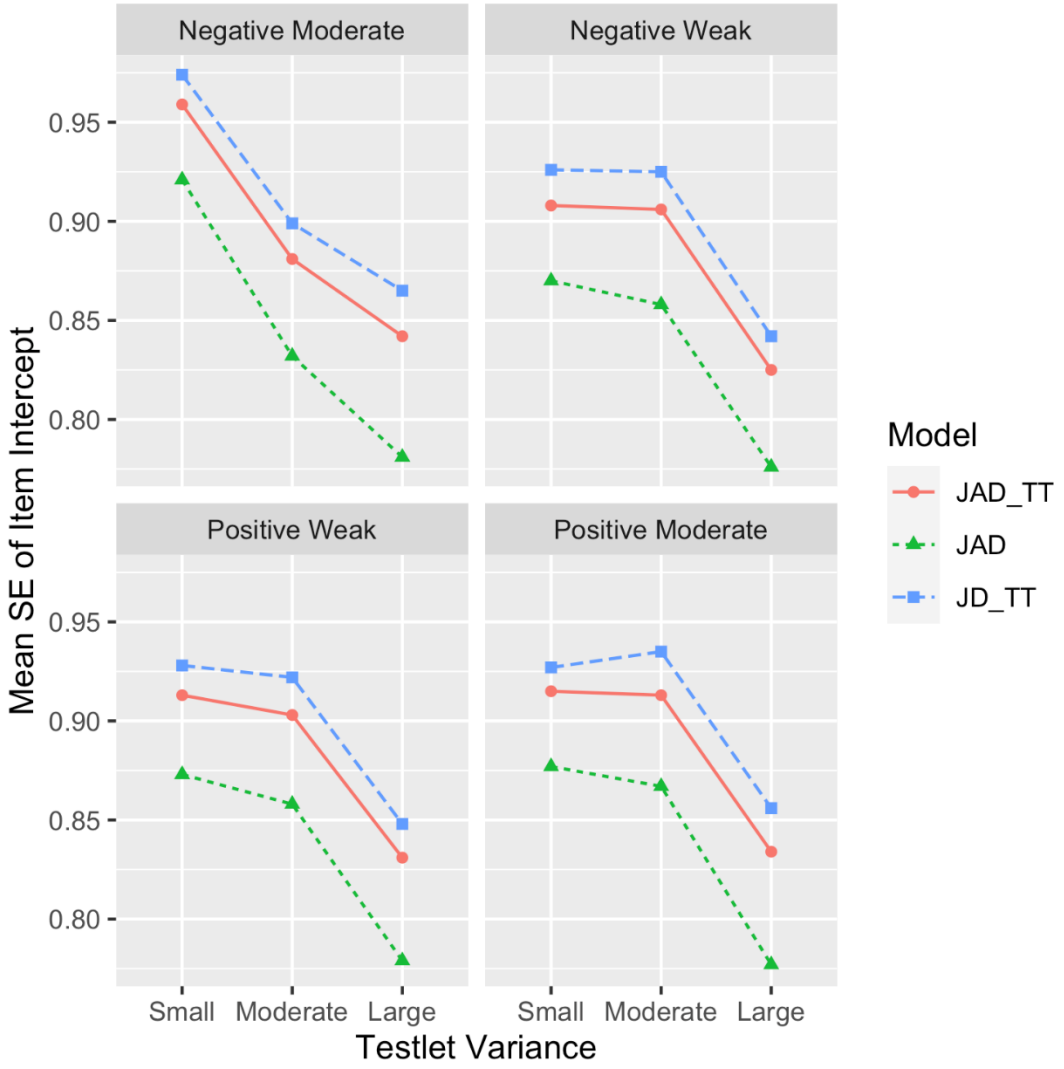

Figure 25 Significant three-way interaction of testlet variance and model type on the mean SE for β_j for four levels in the correlation between θ_i and τ_i.

Item Interaction Parameter. Table 16 summarizes the significant main and interaction effects on the recovery of δ_j identified in the mixed-effect ANOVA. Similar to significant effects on β_j, model, sample size, and testlet variance all have a significant main effect on the bias and SE of δ_j. Significant two-way interaction effects on the bias and SE of δ_j stem from the interaction of model and sample size and of model and testlet variance. Additionally, the interaction of sample size and testlet variance, and of $\rho_{\theta\tau}$ and testlet variance have a significant effect on the SE of δ_j. Three three-way interactions of model, sample size, and testlet variance, of model, $\rho_{\theta\tau}$, and testlet variance, and of sample size, $\rho_{\theta\tau}$, and testlet variance significantly affect the bias and SE of δ_j. The highest-order interaction of all three manipulated factors and model type significantly affects the SE of δ_j.

Table 16
Effect Sizes in the Mixed-Effect ANOVA Results of the Bias and SE of the Item Interaction Estimates

Source	Bias of δ_j			SE of δ_j		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment						
Model	35602.089	0.000	0.984	4658.176	0.000	0.890
Model*Sample Size	287.736	<0.001	0.333	57.602	<0.001	0.091
Model*Testlet Variance	228.442	<0.001	0.442	62.449	<0.001	0.178
Model*Sample Size*Testlet Variance	3.187	0.027	0.011	3.073	0.027	0.011
Model*Correlation*Testlet Variance	3.038	0.002	0.031	3.593	<0.001	0.036
Model*Sample Size*Correlation*Testlet Variance				2.396	0.011	0.024
Between-Subject Effects						
Sample Size	32.221	<0.001	0.053	17.386	<0.001	0.029
Testlet Variance	366.352	<0.001	0.560	85.913	<0.001	0.230
Sample Size*Testlet Variance				3.653	0.027	0.013
Correlation*Testlet Variance				2.623	0.016	0.027
Sample Size*Correlation*Testlet Variance	2.465	0.023	0.025	2.207	0.041	0.022

Note: Effect Size is classified as follows: Small (0.01≤partial η^2 <0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

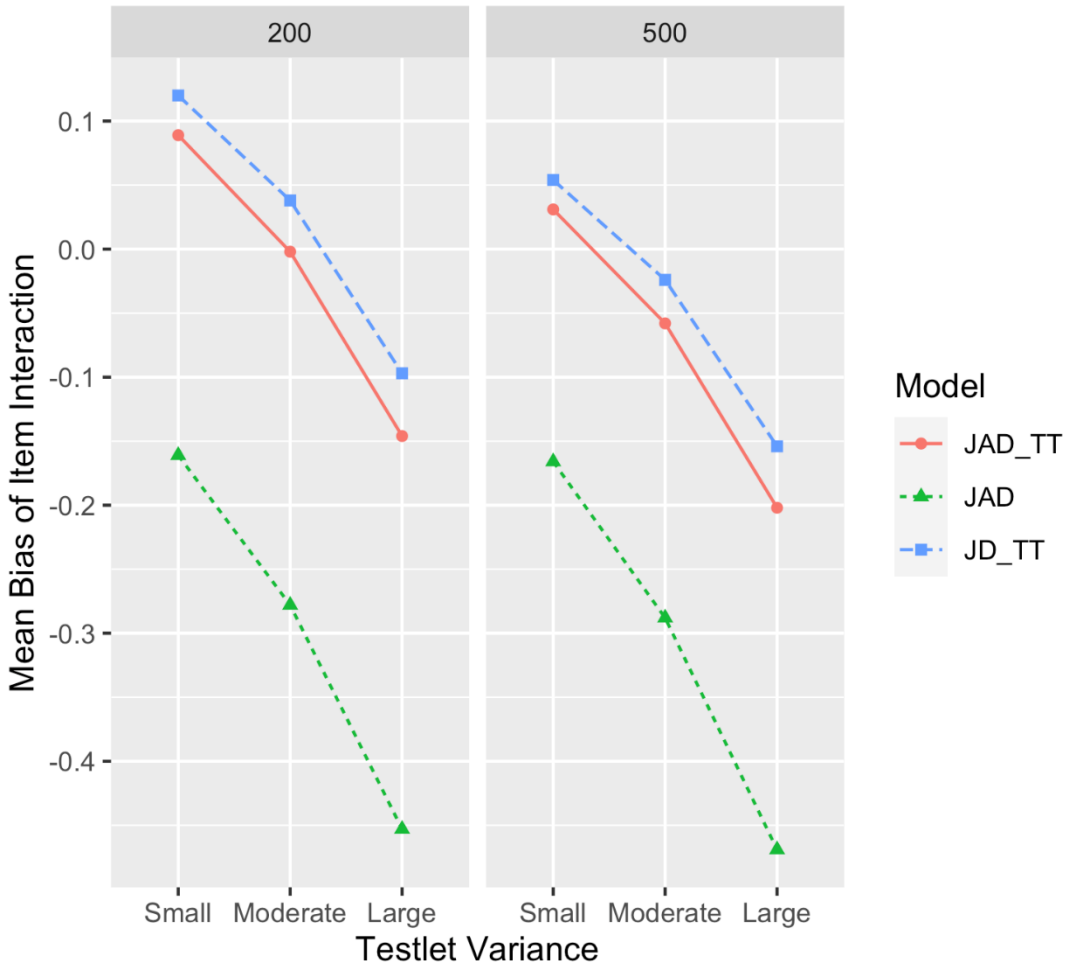

Figure 26 Significant three-way interaction of testlet variance and model type on the mean bias for δ_j for the sample size of 200 and 500.

Figures 26 and 27 display the marginal mean bias of δ_j at the three testlet variance levels for the two sample sizes and four $\rho_{\theta\tau}$ levels. Across the sample sizes and $\rho_{\theta\tau}$ levels, at small testlet variance, the absolute value of the marginal mean biases of δ_j as estimated by the proposed model are the lowest compared to the other two models. Further, across the levels of the manipulated factors, estimation by the proposed model and the Joint Testlet-DINA model yields smaller absolute value of the marginal mean biases of δ_j than the JRT-DINA-R/RT/AC model. The effect of

testlet variance is consistent across the models, sample sizes, and levels of $\rho_{\theta\tau}$: the absolute value of the marginal mean biases of δ_j decreases as testlet variance increases from small to medium and is the highest when testlet variance is large at 1.

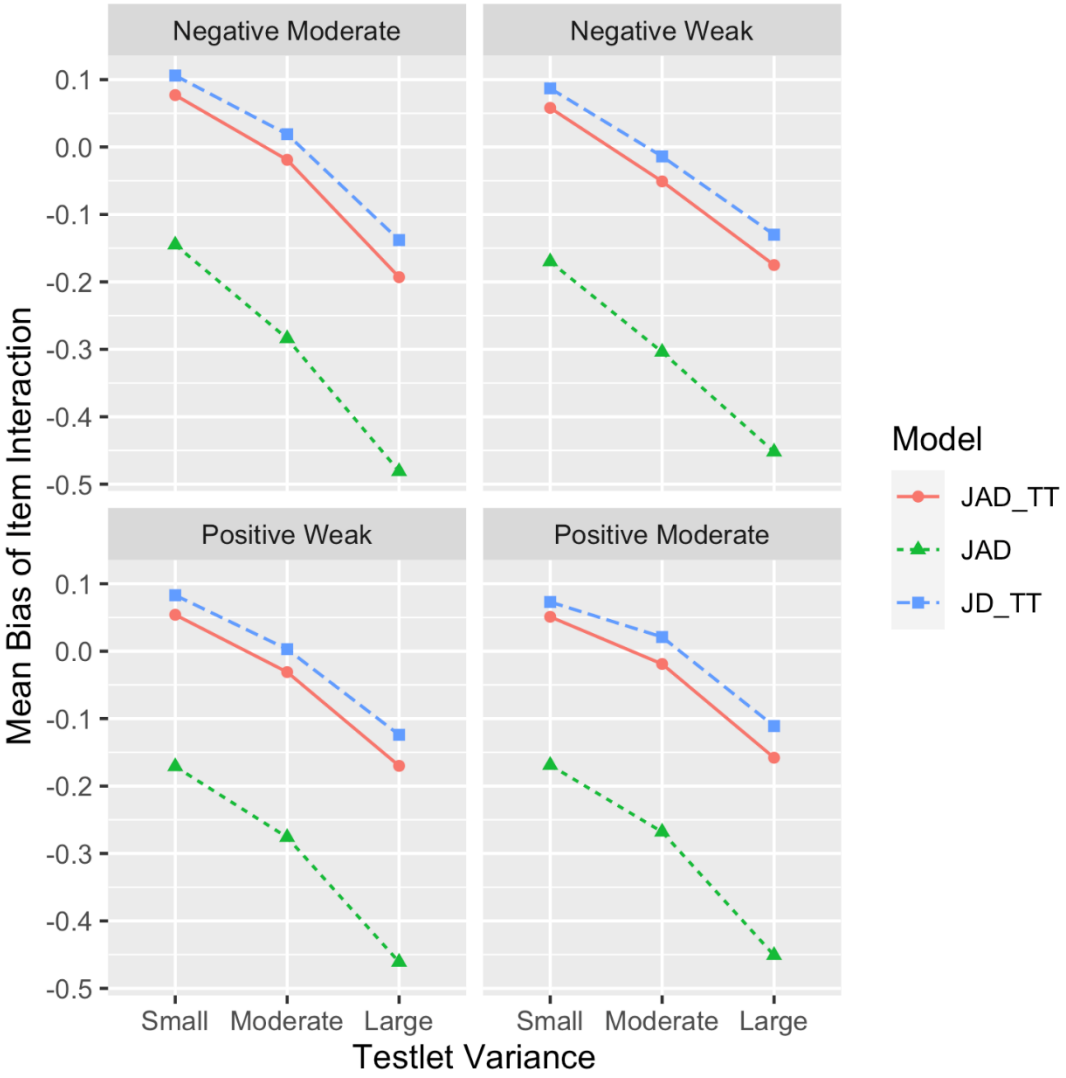

Figure 27 Significant three-way interaction of testlet variance and model type on the mean bias for δ for four levels in the correlation between θ_i and τ_i.

Figure 28 displays the highest-order interaction effect of $\rho_{\theta\tau}$ and testlet variance on marginal mean bias for δ_j as estimated by the proposed model at the two sample sizes. At the sample size of 200, the absolute value of the marginal mean bias is the lowest at the moderate testlet variance level, increases at the level of small

testlet variance, and is the highest when testlet variance is large. At 500, they are the lowest at the small testlet variance level, increases at the level of moderate testlet variance, and are the largest at the level of large testlet variance. Variation by $\rho_{\theta\tau}$ is specific to the testlet variance level and the sample size. At the sample size of 200, as an example, the lowest mean bias for the levels of small and large testlet variance is at positive weak $\rho_{\theta\tau}$. At 500, the lowest mean bias for the moderate testlet variance level is at negative moderate $\rho_{\theta\tau}$.

Figure 29 visualizes the highest-order interaction of model type and the three manipulated factors on the marginal mean SE for δ_j. Across the sample sizes and levels of $\rho_{\theta\tau}$, mean SE for δ_j as estimated by the proposed model is larger than estimation by the JRT-DINA-R/RT/AC model, but smaller than estimation by the Joint Testlet-DINA model. Variation by levels of the testlet variance is not consistent across the levels of $\rho_{\theta\tau}$ and sample sizes. At negative moderate $\rho_{\theta\tau}$ and across the sample sizes, an increase in testlet variance is associated with a decrease in the marginal mean SE of δ_j. Similar variation is observed for positive $\rho_{\theta\tau}$ at the sample size of 500 and for negative weak $\rho_{\theta\tau}$ at the sample size of 200. At all other levels of the interaction of the two manipulated factors, the marginal mean SE of δ_j is the smallest at large testlet variance and the largest at moderate testlet variance.

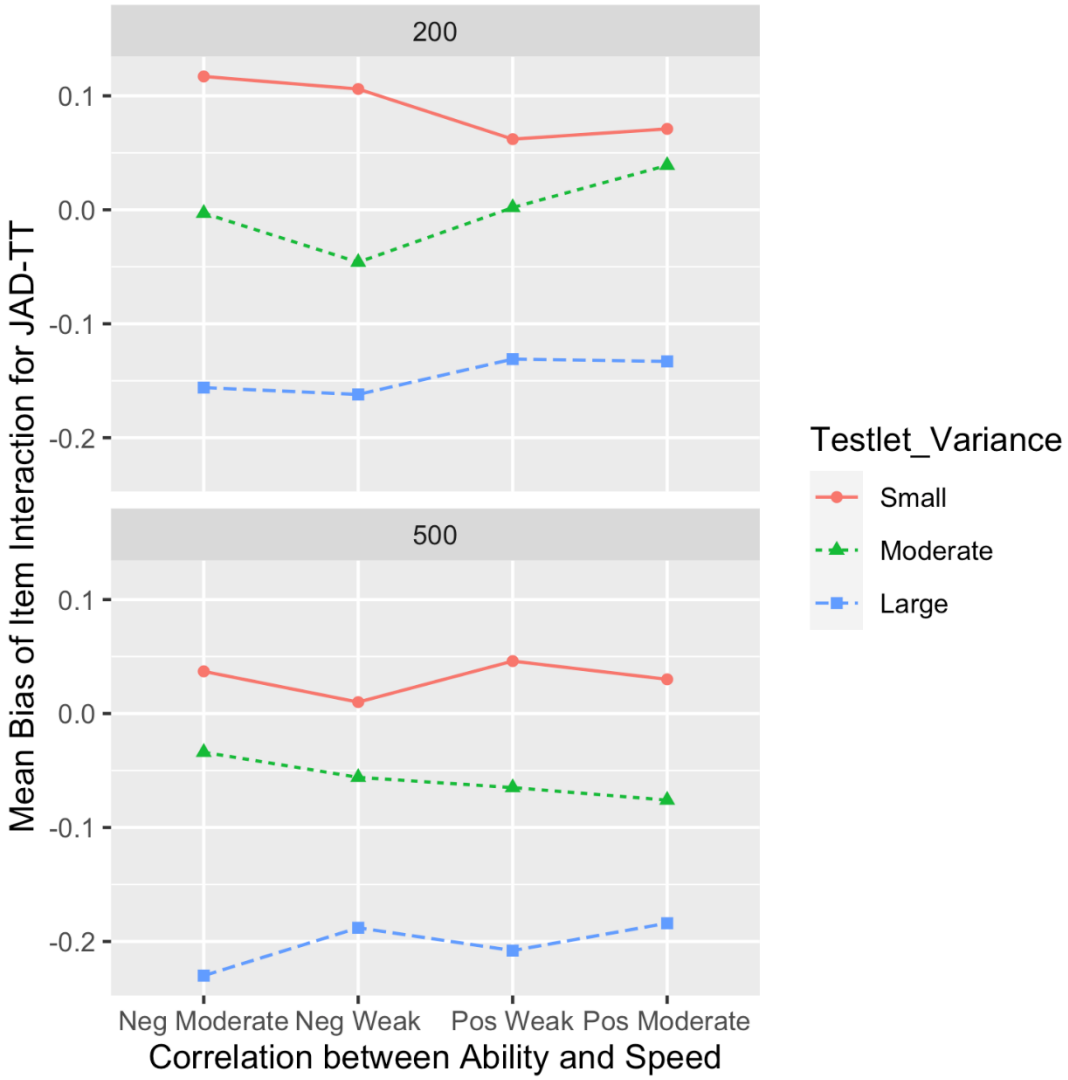

Figure 28 Significant three-way interaction of sample size, correlation, and testlet variance on the mean bias for δ_j of the proposed model.

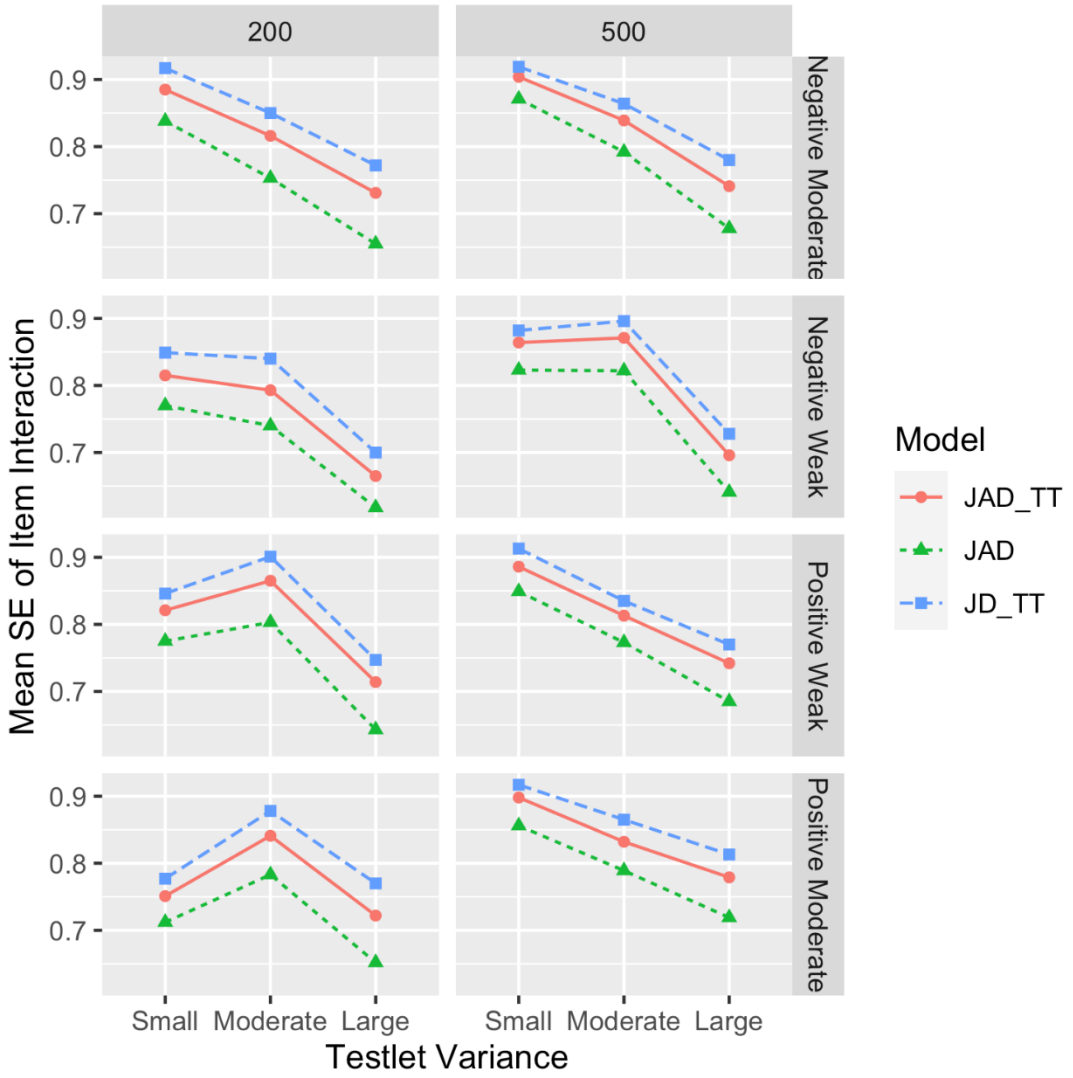

Figure 29 Significant four-way interaction of model type, sample size, correlation, and testlet variance on the mean SE for δ_j.

Item Time Intensity Parameter. Table 17 presents the significant Effects in the mixed-effect ANOVA results of the bias and SE of the item time intensity parameter, ζ_j. As is seen in the table, main effects of model, sample size, testlet variance, and $\rho_{\theta\tau}$ are all significant. Further, all two-way, three-way, and four-way interaction of model type and the manipulated factors are significant. The effect sizes for the bias of ζ_j are mostly large whereas for the SE of ζ_j, they are of all three magnitudes.

Table 17

Effect Sizes in the Mixed-Effect ANOVA Results of the Bias and SE of the Item Time Intensity

Source	Bias of ζ_j			SE of ζ_j		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment						
Model	251.184	<0.001	0.304	89.292	<0.001	0.134
Model*Sample Size	121.017	<0.001	0.174	18.376	<0.001	0.031
Model*Correlation	703.304	0.000	0.786	54.786	<0.001	0.222
Model*Testlet Variance	272.654	<0.001	0.486	49.652	<0.001	0.147
Model*Sample Size*Correlation	390.621	<0.001	0.670	15.452	<0.001	0.074
Model*Sample Size*Testlet Variance	595.868	<0.001	0.674	148.568	<0.001	0.340
Model*Correlation*Testlet Variance	204.507	<0.001	0.681	23.242	<0.001	0.195
Model*Sample Size*Correlation*Testlet Variance	329.272	0.000	0.774	27.429	<0.001	0.222
Between-Subject Effects						
Sample Size	44010.067	0.000	0.987	14.634	<0.001	0.025
Correlation	12994.185	0.000	0.985	43.141	<0.001	0.183
Testlet Variance	23132.192	0.000	0.988	1241.174	<0.001	0.812
Sample Size*Correlation	5190.496	0.000	0.964	13.867	<0.001	0.067
Sample Size*Testlet Variance	15392.066	0.000	0.982	6.799	0.001	0.023
Correlation*Testlet Variance	26337.376	0.000	0.996	44.476	<0.001	0.317
Sample Size*Correlation*Testlet Variance	15525.668	0.000	0.994	5.765	<0.001	0.057

Note: Effect Size is classified as follows: Small (0.01≤partial η^2<0.06), Medium (0.06≤partial η^2<0.14), Large (partial η^2≥0.14)

Figures 30 and 31 visualizes the highest-order interaction of model type, sample size, $\rho_{\theta\tau}$, and testlet variance on the marginal mean bias and SE of ζ_j at the two sample sizes and for the four levels in $\rho_{\theta\tau}$. Across the levels of the manipulated factors, the marginal mean biase and SE of ζ_j estimated by the three models are close to identical, suggesting the model specification has little impact on the recovery of this parameter. Variation in the marginal mean bias of ζ_j by testlet variance is specific to the levels of sample size and $\rho_{\theta\tau}$, whereas variation in the marginal mean SE of ζ_j by testlet variance is more consistent, with it being greater for larger testlet variance.

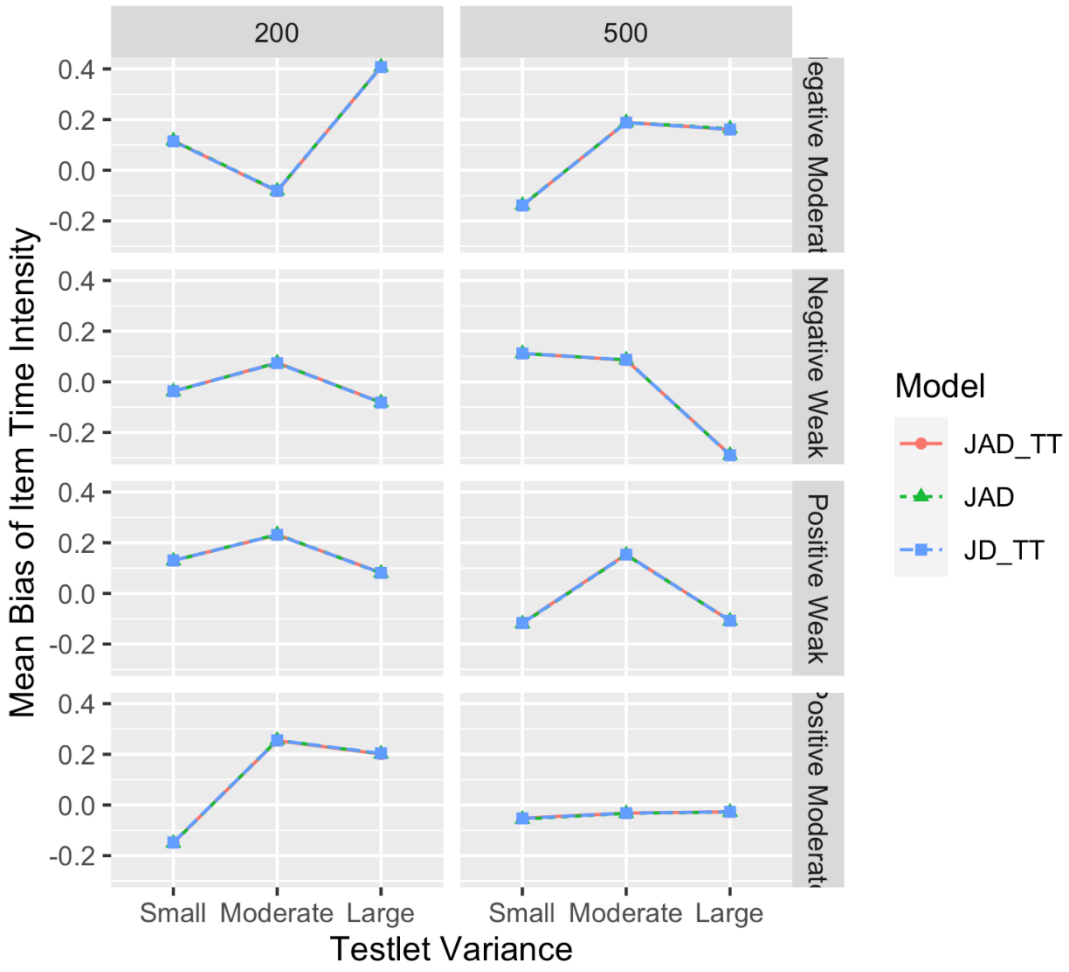

Figure 30. Four-way interaction of model type, sample size, correlation, and testlet variance on the mean bias of ζ_j.

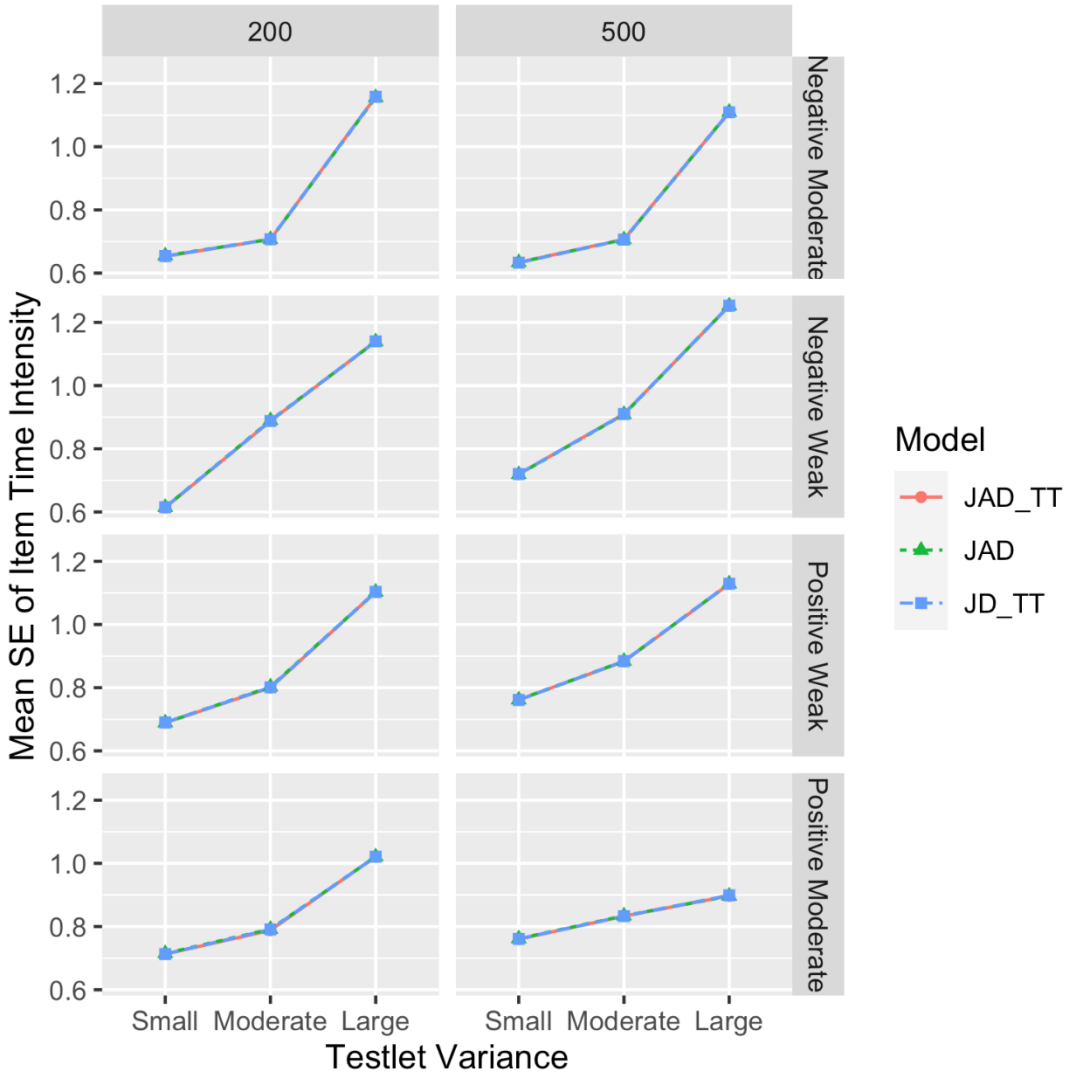

Figure 31. Significant four-way interaction of model type, sample size, correlation, and testlet variance on the mean SE of ζ_j.

4.1.6 Recovery of the Item Mean Vector and Item Variance/Covariance Matrix

Elements of the item mean vector evaluated in this study are μ_β, mean of the item intercept parameter, μ_δ, mean of the item interaction parameter, and μ_ζ, mean of the item time-intensity parameter. μ_b, mean of the item difficulty parameter is excluded from this study as it is constrained to 0. This section compares the bias and

SE for the three parameters for all three models under the 24 simulated conditions to examine and compare the extent to which the three models successfully recover the item mean vector. Table A.8 presents the bias for the item mean vector for the three models under each simulated condition. Table A.9 presents its SE. Across the simulated conditions, the absolute value of the bias for the three item mean parameters and the SE for μ_β and μ_δ is < 0.30. The SE for μ_ζ is< 1.2.

Item Mean Vector Figures 32 through 37 display the marginal mean biases and SEs of the μ_β, μ_δ, and μ_ζ as estimated by the three models and for all levels of the manipulated factors. As is shown in the figures, model specification has little or no effect on the marginal mean bias and SE of μ_ζ: they are close to identical for the three models. Marginal mean bias and SE of μ_β and μ_δ are comparable for estimates generated by the proposed model and the Joint Testlet-DINA model. Slightly better marginal mean bias for μ_β and μ_δ is observed for the proposed model at the small testlet variance level and is observed for μ_δ only for the sample size of 200. At all other levels of the three manipulated factors, the marginal mean biases of μ_β and μ_δ are slightly larger for the proposed model than for the Joint Testlet-DINA model. Across the levels, the marginal mean SEs are slightly smaller for the proposed model. Compared with the JRT-DINA-R/RT/AC model, the two testlet-based models yield smaller biases of μ_β and μ_δ and larger SEs.

Variation by levels of the manipulated factors in the two indices quanitfying the random and systematic error is specific to the factors and the elements being

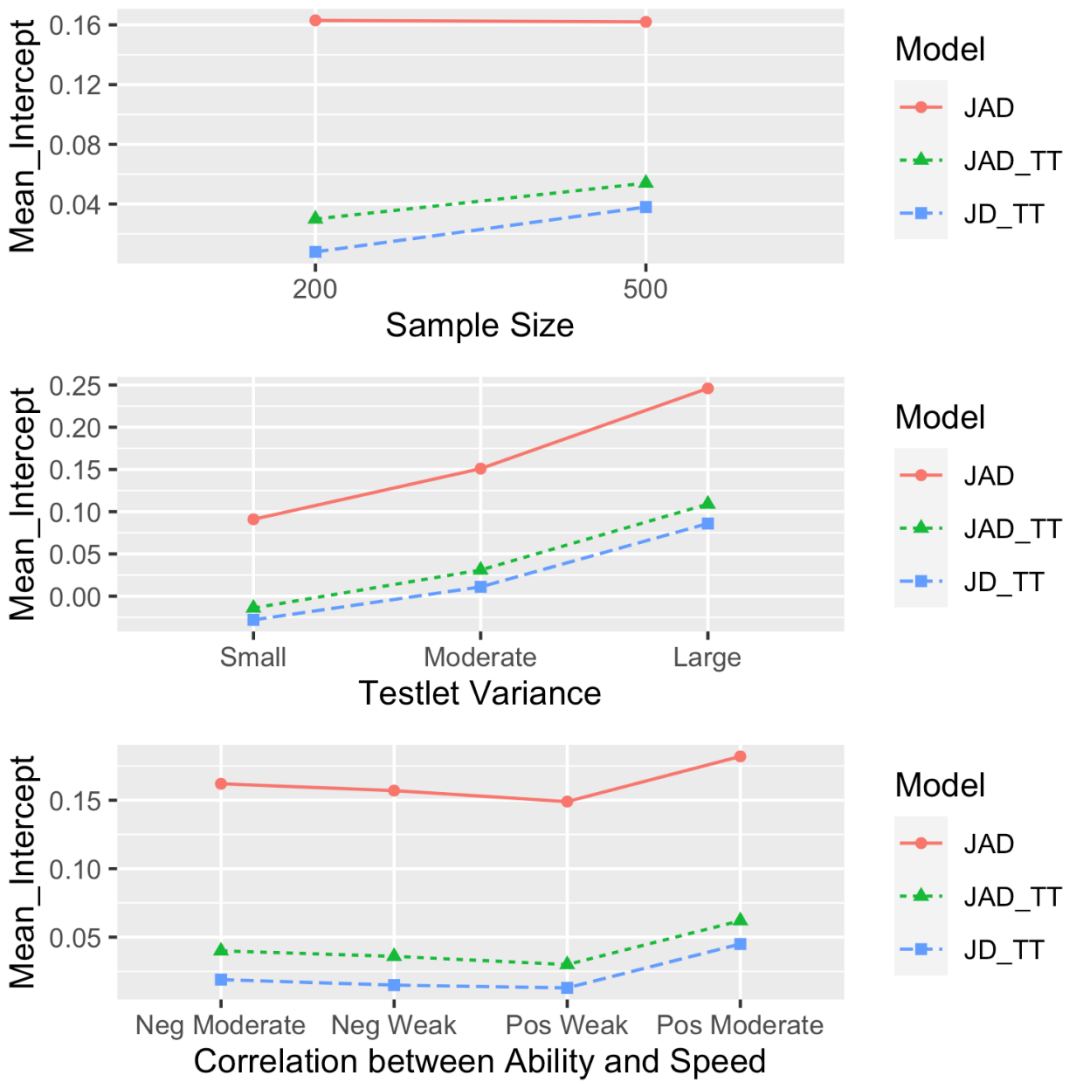

Figure 32 Marginal mean bias of the mean item intercept μ_β at all levels of the manipulated factors.

evaluated. Variation in the marginal mean bias by levels of $\rho_{\theta\tau}$ is consistent for μ_β and μ_δ: their marginal mean biases show little variation for the levels of negative and positive weak $\rho_{\theta\tau}$; their absolute value is the highest at positive moderate $\rho_{\theta\tau}$. The marginal mean SE of μ_β is the highest at negative weak $\rho_{\theta\tau}$ and are comparable at the two positive $\rho_{\theta\tau}$. For μ_δ they are the lowest at positive moderate $\rho_{\theta\tau}$. As sample size increases from 200 to 500, across the models the

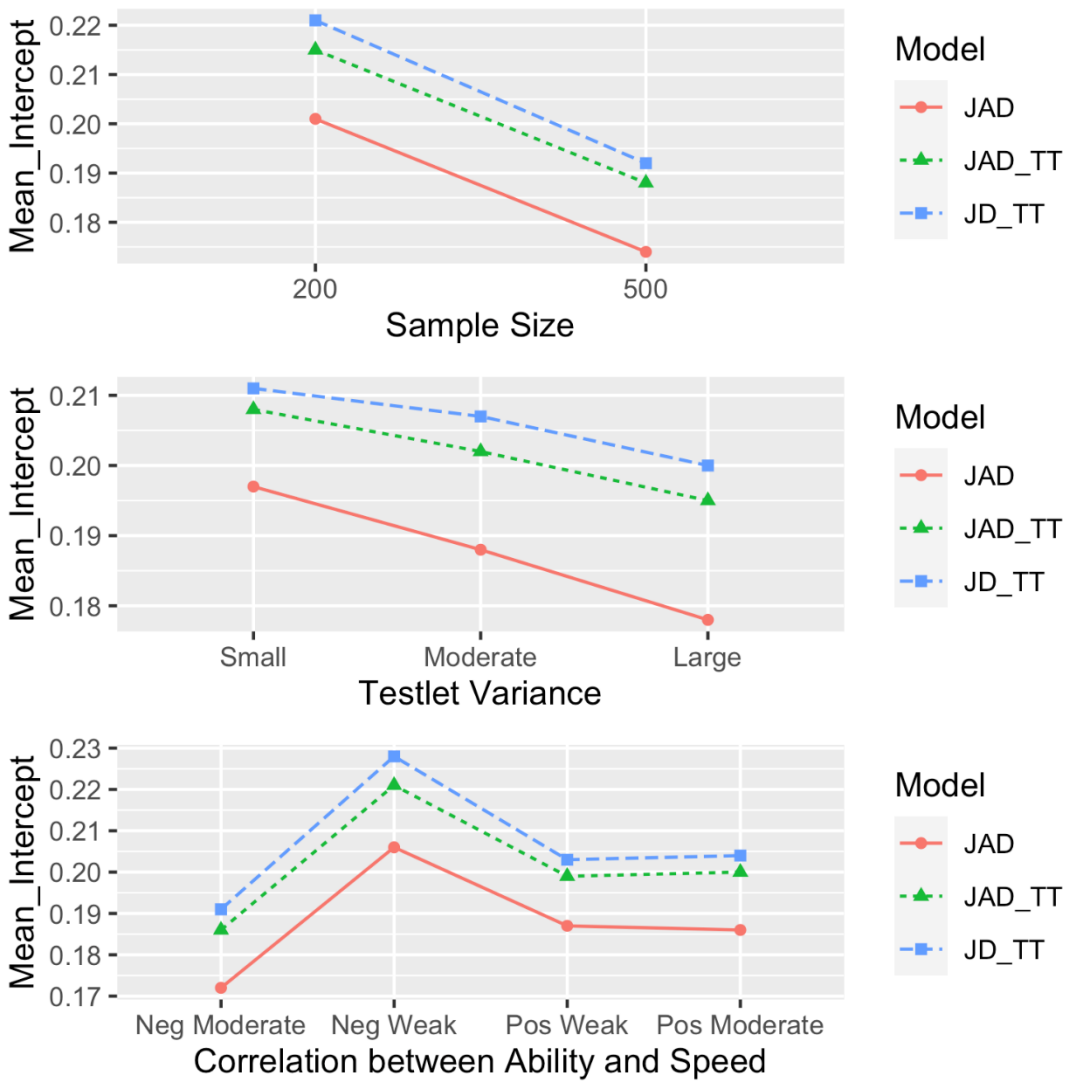

Figure 33 Marginal mean SE of the mean item intercept μ_β at all levels of the manipulated factors.

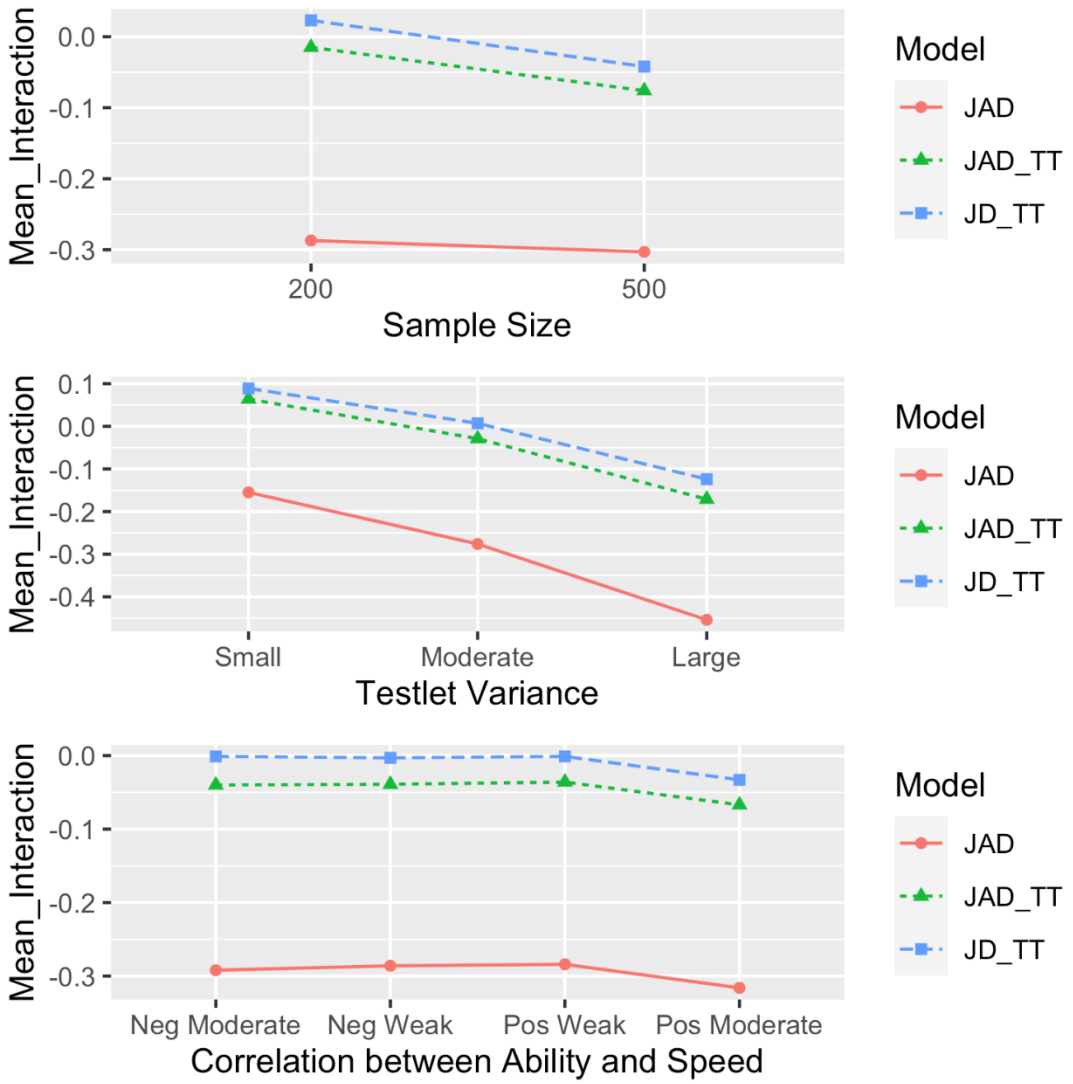

Figure 34. Marginal mean bias of the mean item interaction μ_δ at all levels of the manipulated factors.

marginal mean SEs of μ_β and μ_δ decrease. Their marginal mean biases show model specific variation: they show little variation for the JRT-DINA-R/RT/AC model and increases for the other two models. Similar contrasting pattern is shown for the variation in the marginal mean biases by testlet variance: as testlet variance increases, the marginal mean biases of μ_β increase whereas for μ_δ they decrease from small to moderate testlet variance and are the highest at large testlet variance. The marginal SE of μ_β decreases with increased testlet variance whereas for μ_δ, they are the largest

at small testlet variance and the smallest at moderate testlet variance. Variation in the marginal mean bias and SE for μ_ζ is similarly specific to the factors being manipulated as is shown in Figures 37 and 38.

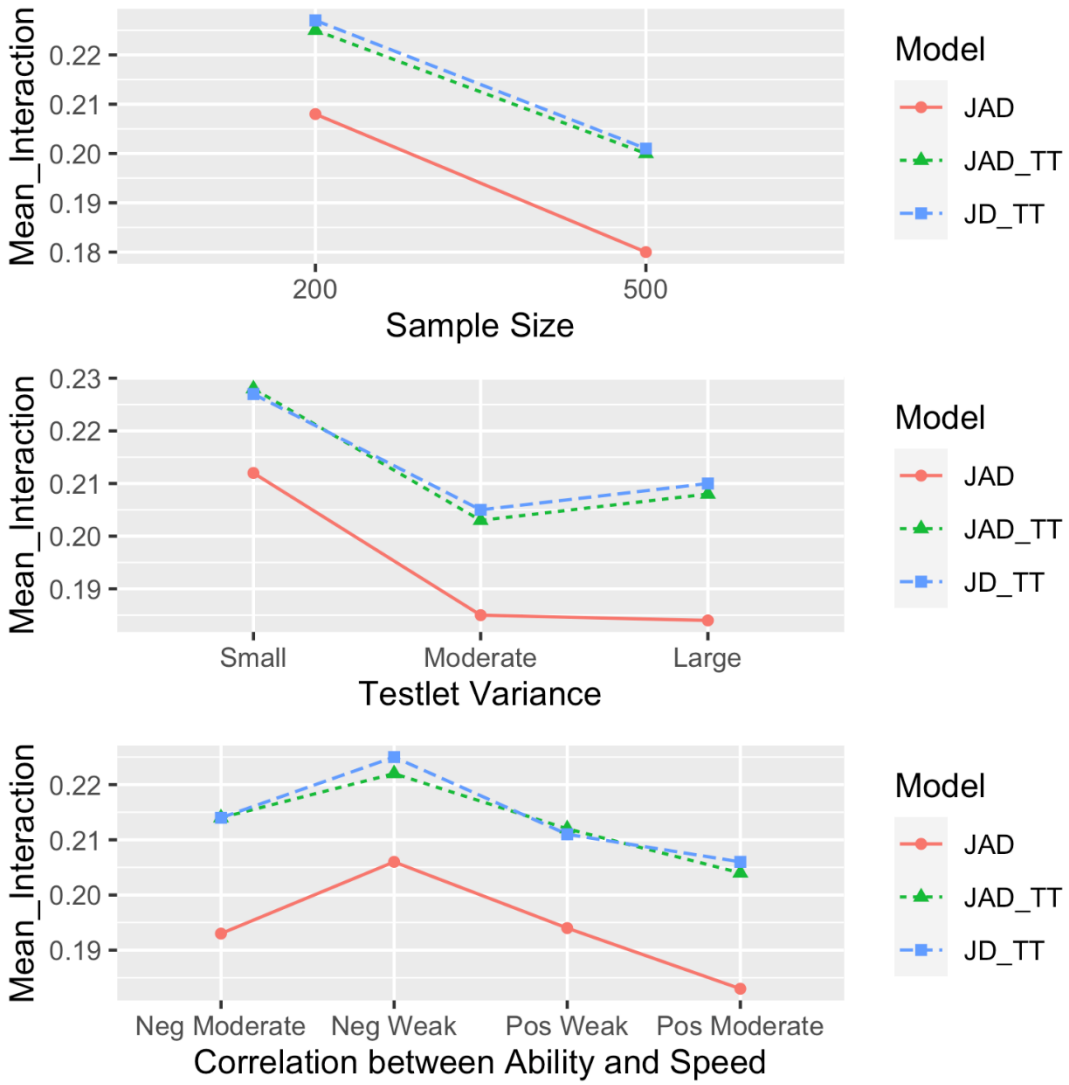

Figure 35 Marginal mean SE of the mean item interaction μ_δ at all levels of the manipulated factors.

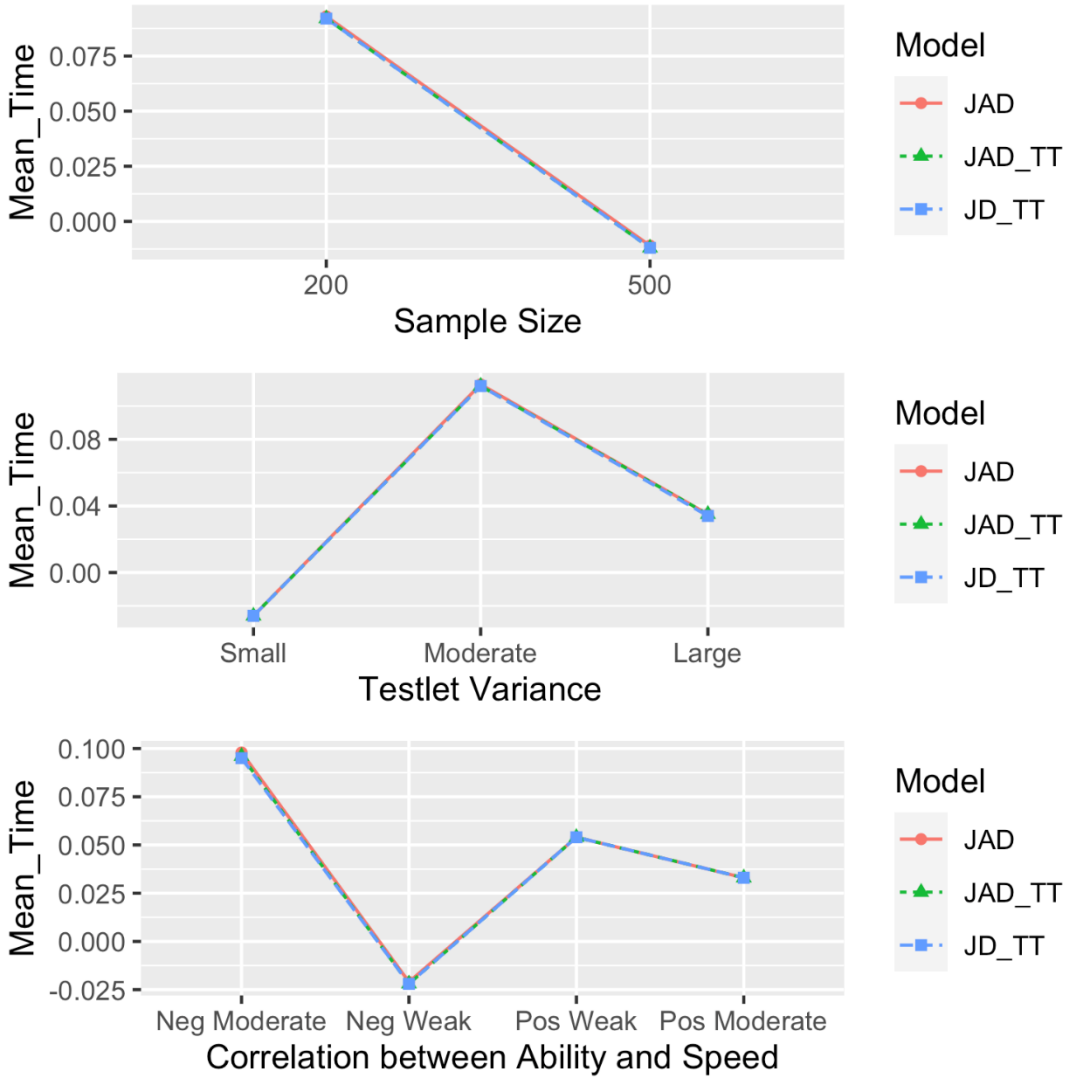

Figure 36. Marginal mean bias of mean item time intensity μ_ζ at all levels of the manipulated factors.

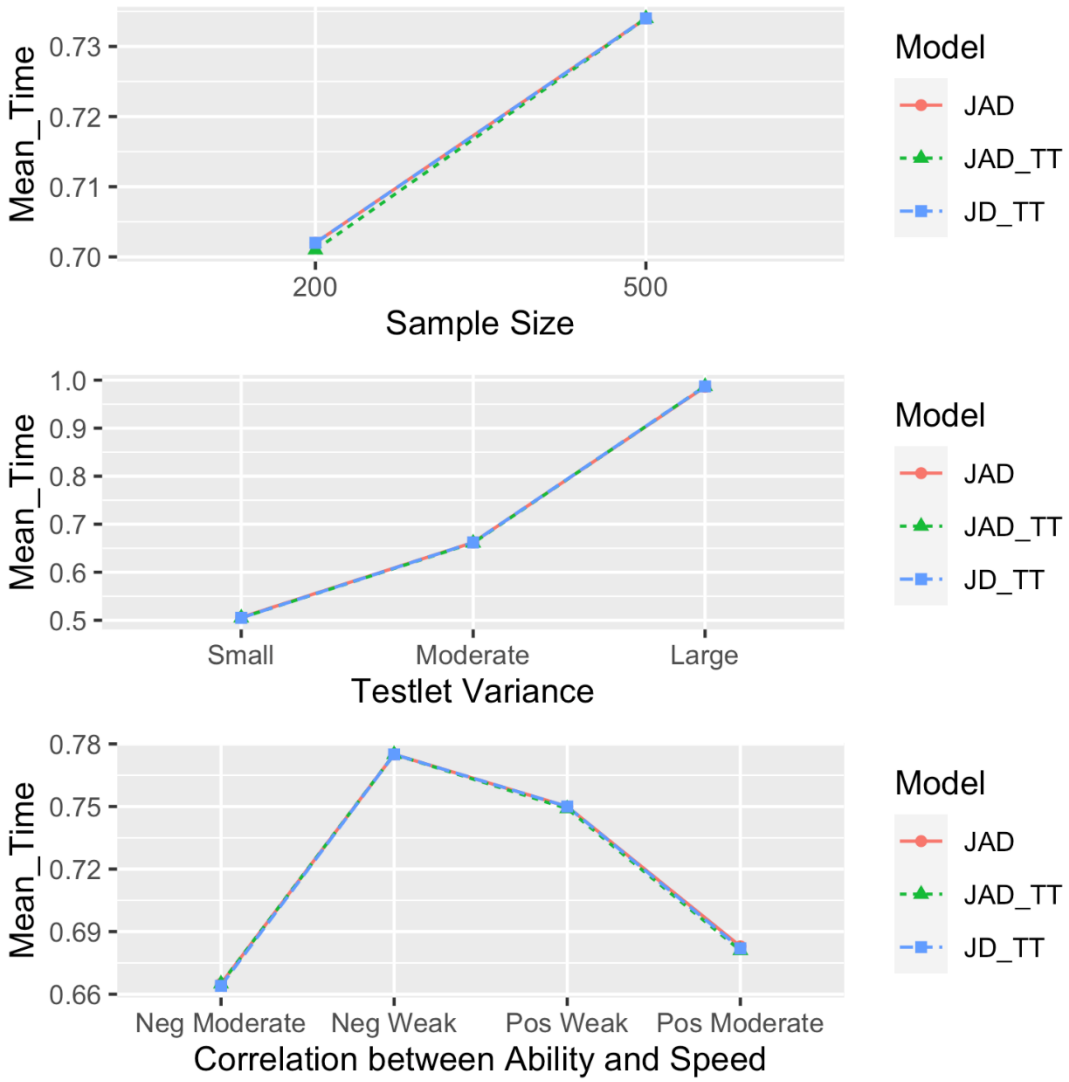

Figure 37. Marginal mean SE of mean item time intensity μ_ζ at all levels of the manipulated factors.

Item Variance and Covariance Matrix. Tables A.10.1-4 and A.11.1-4 present the biases and SEs of the elements of the item variance and covariance matrix for the 24 simulated conditions. Figures 38 through 57 are visual presentations of the marginal mean biases and SEs of the estimates of these elements for all three models and at all levels of the manipulated factors. As the Joint Testlet-DINA model does not have the item difficulty parameter, specific elements of the item variance/covariance matrix, $\sigma_{\beta b}$, $\sigma_{\delta b}$, $\sigma_{\zeta b}$, and σ_b^2 are estimated by the other two models that include the item difficulty parameter in their model specification. For these elements, comparisons are drawn between the proposed model and the JRT-DINA-R/RT/AC model.

As is seen in figures 53 through 58, model specification has little or no impact on the marginal mean biases and SEs of the elements that relate to item time intensity parameter ζ_j and item difficulty parameter b_j: for σ_b^2, σ_ζ^2, and $\sigma_{\zeta b}$ they are identical or close to identical. A and $\sigma_{\delta \mathrm{b}}$, with less systematic but higher random error than the JRT-DINA-R/RT/AC model as their marginal mean biases are lower and marginal mean SEs are higher.

Of the other elements of the item variance/covariance matrix, the proposed model recovers σ_β^2 with least systematic error across the two sample sizes, at the level of small testlet variance, and at the three levels of $\rho_{\theta\tau}$: negative moderate and positive $\rho_{\theta\tau}$. At all other levels of the manipulated factors, estimation of σ_β^2 by the Joint

Testlet-DINA model yields lower marginal mean biases than by the proposed model. Similarly, the marginal mean bias for σ_δ^2 as estimated by the proposed model is the lowest at the small and moderate testlet variance levels. When the testlet variance is large, the proposed model yields higher marginal mean bias for σ_δ^2 than the Joint Testlet-DINA model. Compared with the proposed model and the Joint Testlet-DINA model, estimation by the JRT-DINA-R/RT/AC model yields highest marginal

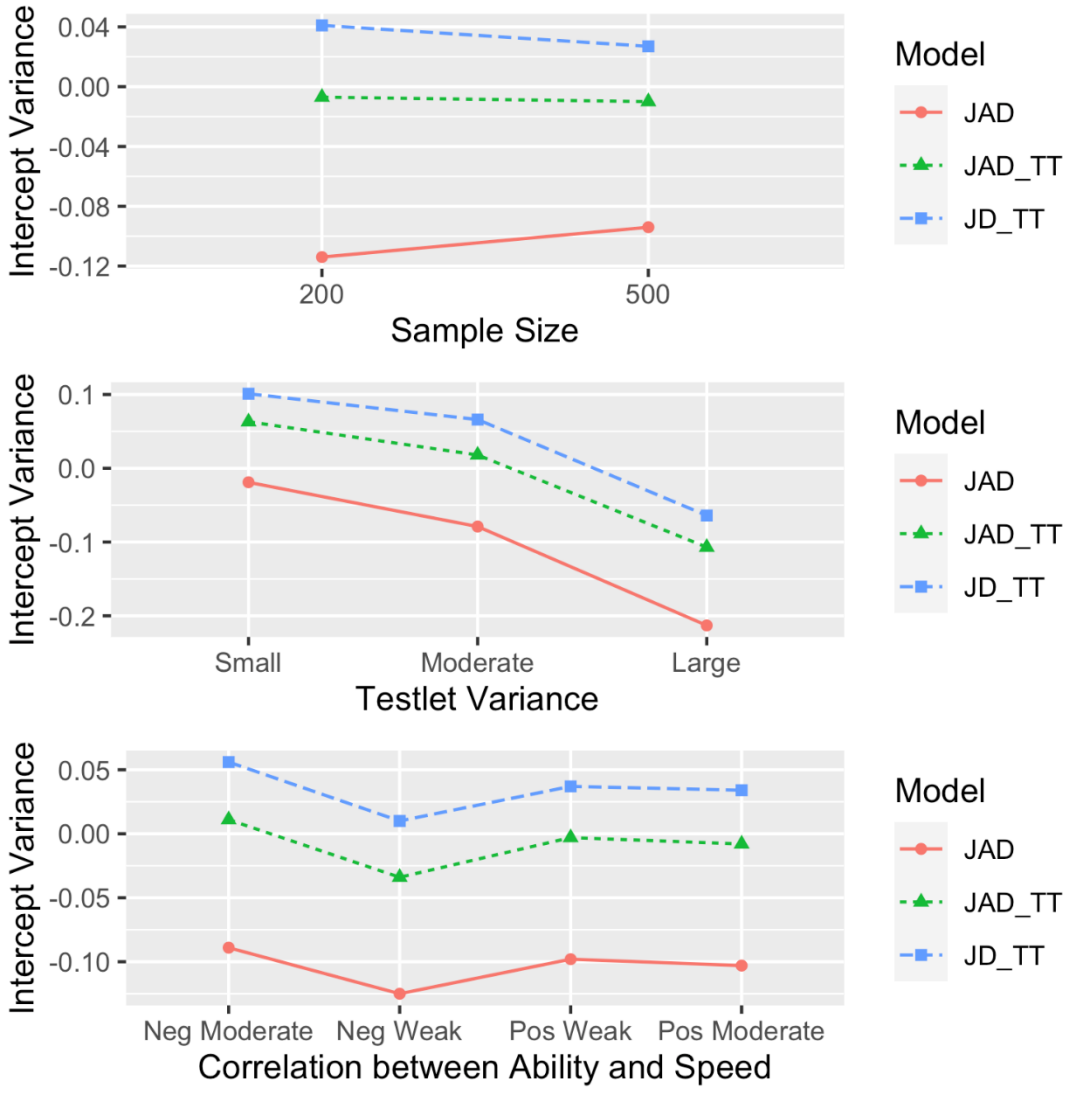

Figure 38. Marginal mean bias of the estimates of item intercept variance σ_β^2 at all levels of the manipulated factors.

mean biases for σ_β^2 and σ_δ^2. Consistent across all levels of the manipulated factors, the Joint Testlet-DINA model yields the lowest marginal mean biases for marginal mean biases for $\sigma_{\beta\delta}$, $\sigma_{\beta\zeta}$ and $\sigma_{\delta\zeta}$ and the JRT-DINA-R/RT/AC model yields the highest biases.

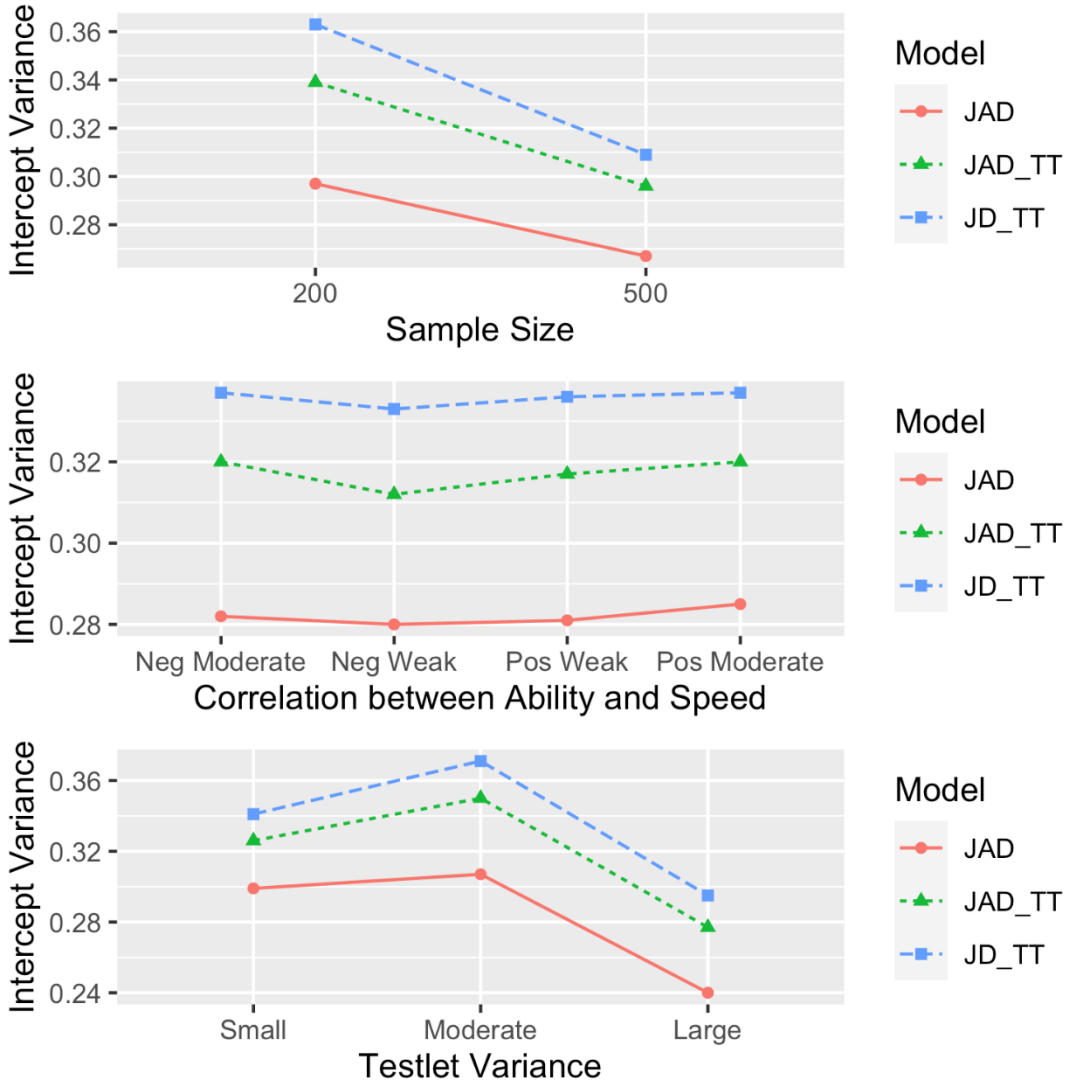

Figure 39 Marginal mean SE of the estimates of item intercept variance σ_β^2 at all levels of the manipulated factors.

Discrepancy in the marginal mean SEs of the elements in the item variance/covariance matrix attributable to model specification follows the same

uniform pattern: across the levels of the manipulated factors, they are the highest as estimated by the Joint Testlet-DINA model, and the lowest as estimated by the JRT-DINA-R/RT/AC model. This suggests that the proposed model recovers elements of the item variance/covariance matrix with less random error than the Joint Testlet-DINA model, but more random error than the JRT-DINA-R/RT/AC model.

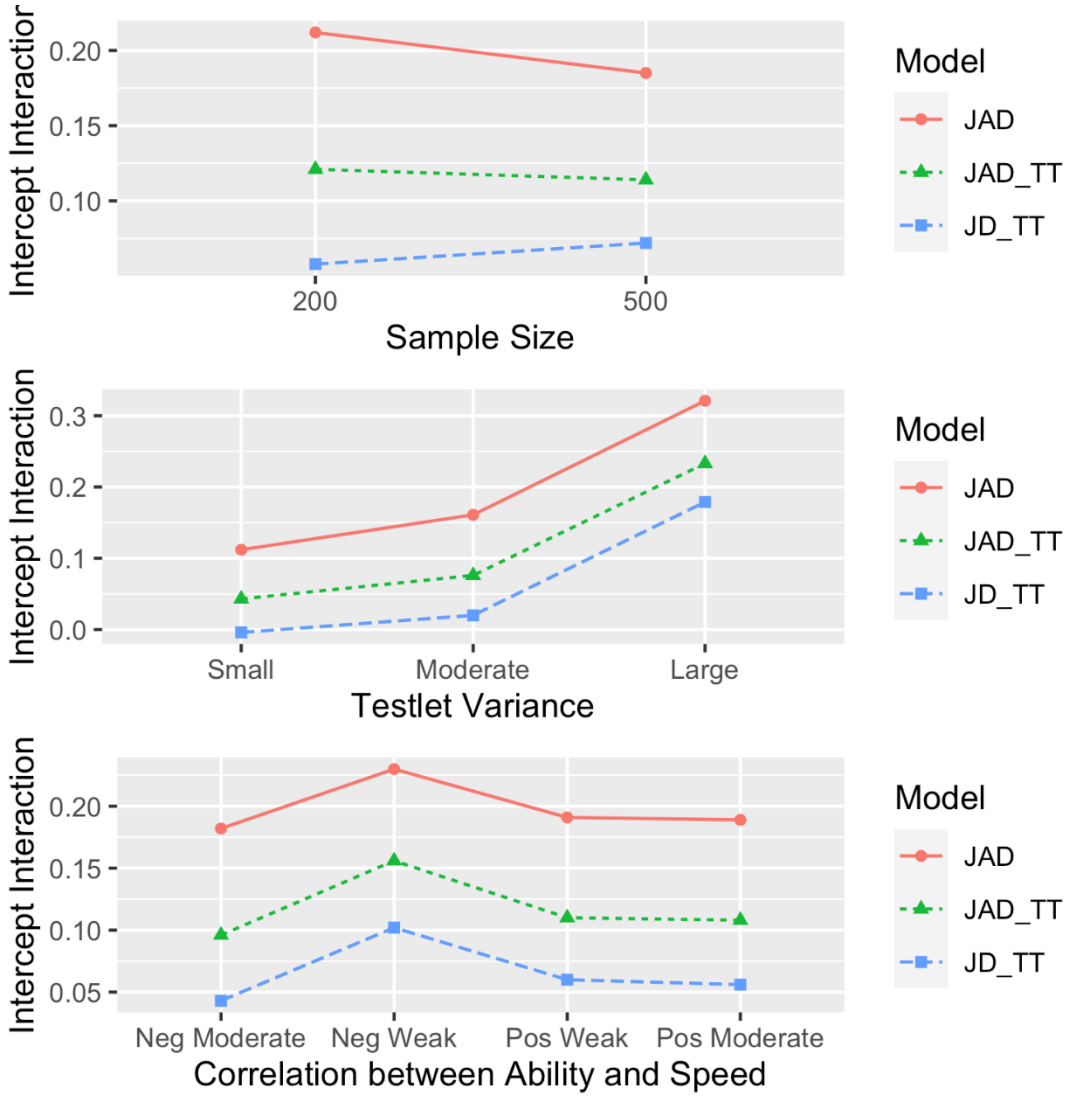

Figure 40. Marginal mean bias of the estimates of covariance of item intercept and item interaction $\sigma_{\beta\delta}$ at all levels of the manipulated factors.

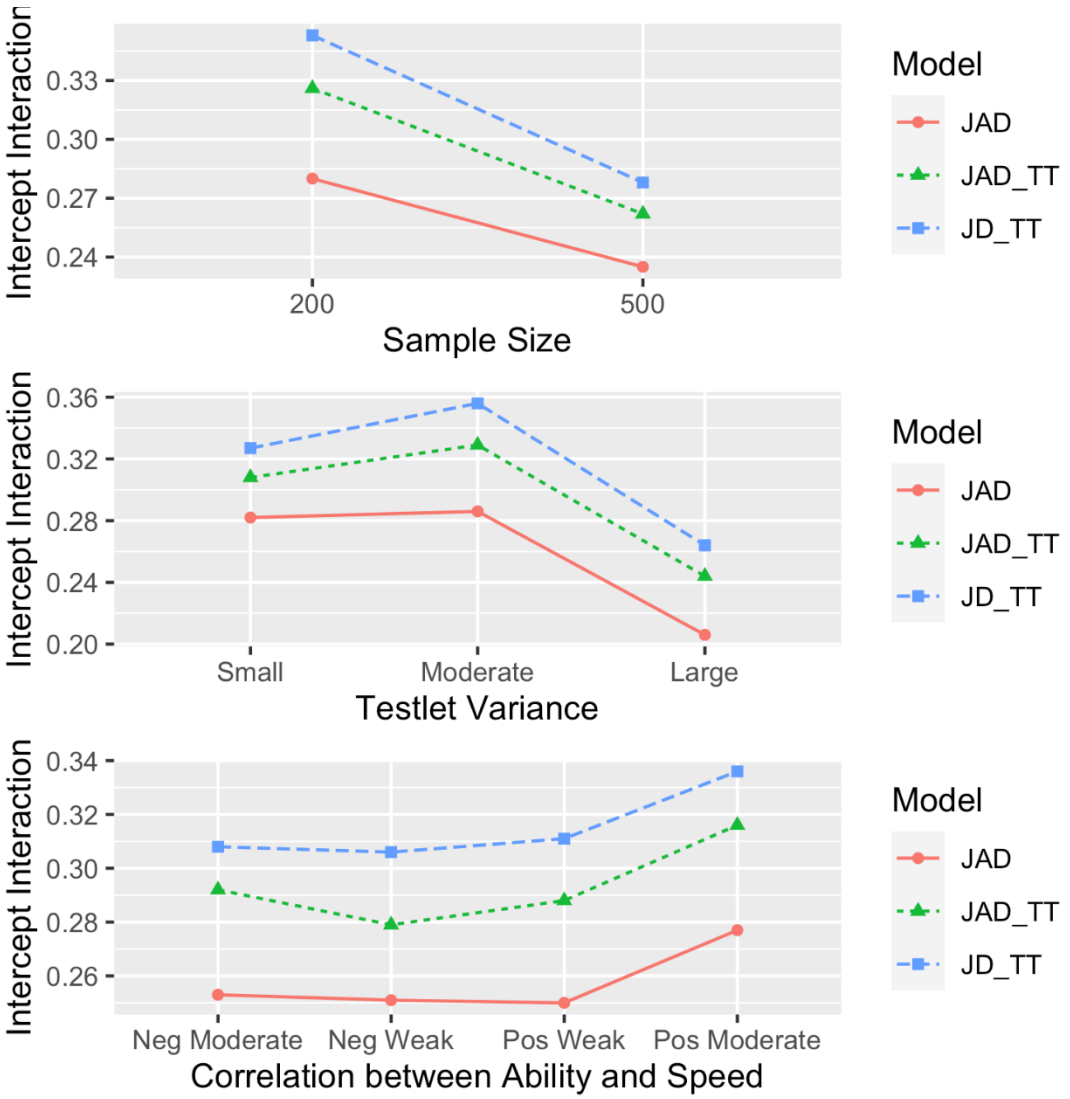

Figure 41 Marginal mean SE of the estimates of covariance of item intercept and item interaction $\sigma_{\beta\delta}$ at all levels of the manipulated factors.

Variation in the marginal mean biases and SEs of the elements in the item variance and covariance matrix by levels of $\rho_{\theta\tau}$ and testlet variance is specific to the elements being estimated. As an example, the absolute value of the marginal mean biases of σ_{β}^2, σ_{δ}^2, $\sigma_{\beta\zeta}$, $\sigma_{\delta\zeta}$, $\sigma_{\beta\delta}$, and $\sigma_{\beta\zeta}$ increase as testlet variance increases. For $\sigma_{\delta b}$ it decreases as the testlet variance level changes from small to moderate and is the highest at large testlet variance. $\sigma_{\delta b}$, $\sigma_{\delta\zeta}$, and σ_{δ}^2 see highest bias at negative weak $\rho_{\theta\tau}$ whereas for $\sigma_{\beta\zeta}$ its marginal mean bias is the highest at positive weak $\rho_{\theta\tau}$.

Variation by levels of the sample size is similarly specific to the elements being evaluated: across the models, the absolute marginal mean biases and SEs are higher at 200 than at 500 for elements such as σ_{β}^2, $\sigma_{\beta\delta}$, and $\sigma_{\beta\zeta}$; yet for $\sigma_{\beta b}$ and σ_{ζ}^2, their marginal mean bias is lower at 200 than at 500.

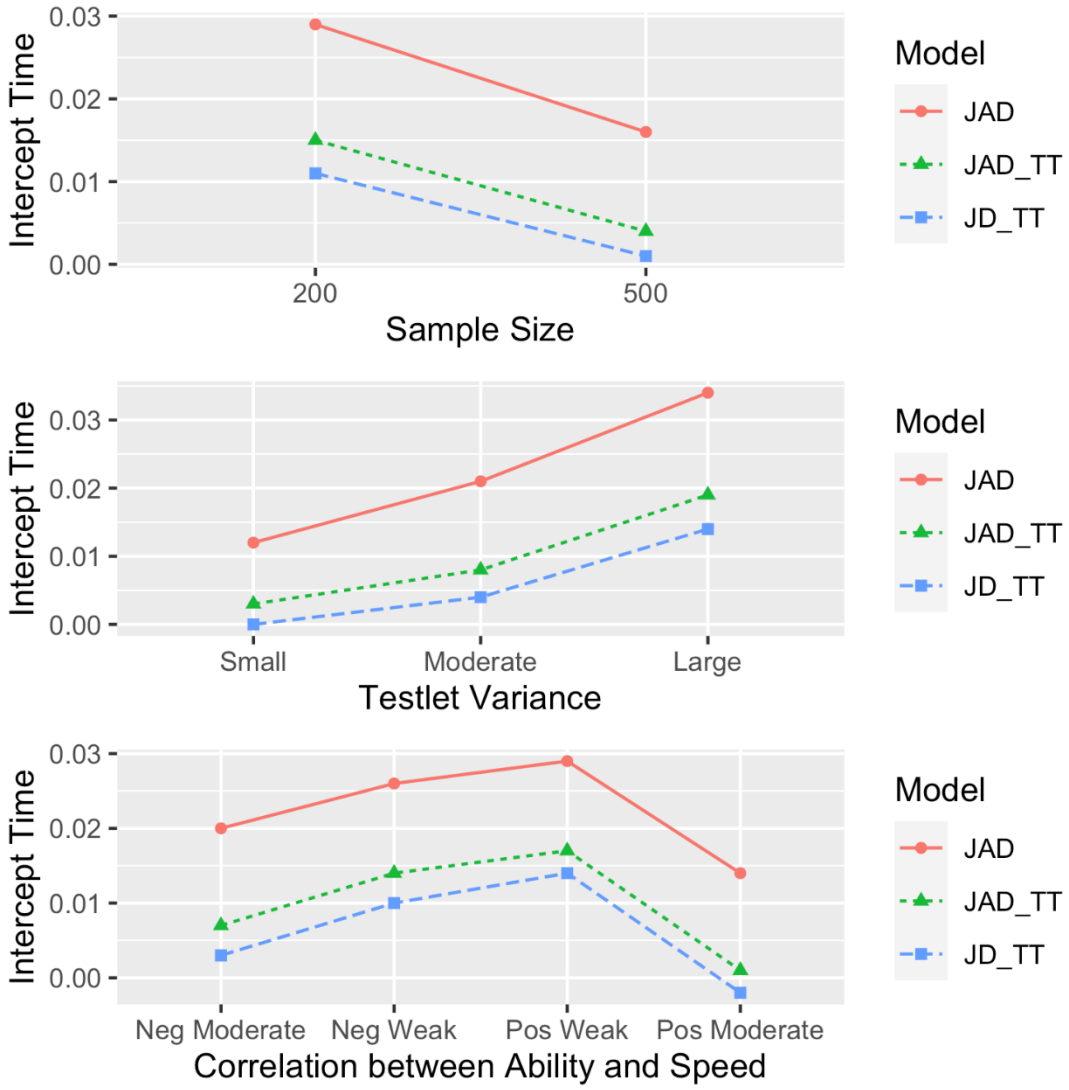

Figure 42 Marginal mean bias of the estimates of covariance of item intercept and item time intensity $\sigma_{\beta\zeta}$ at all levels of the manipulated factors.

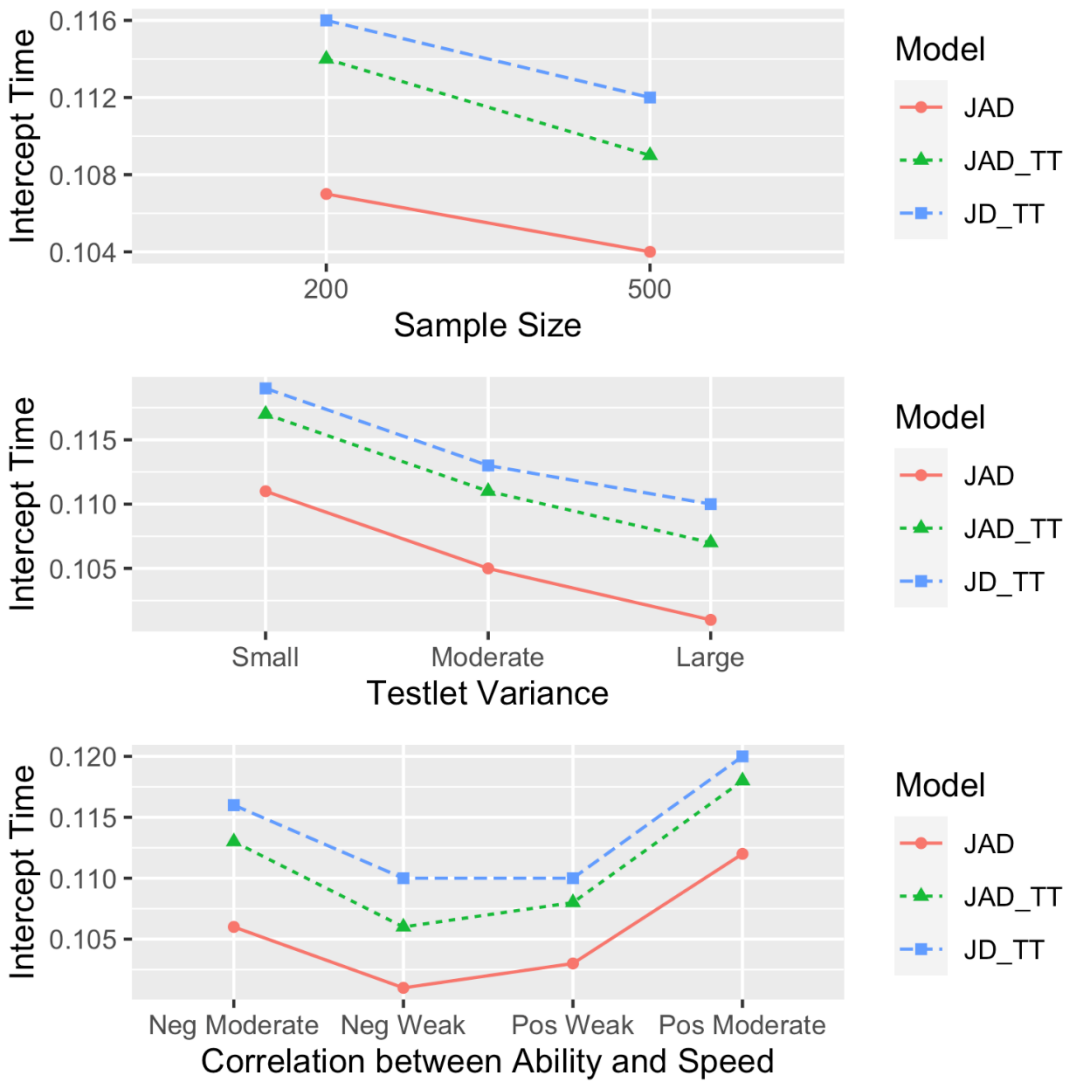

Figure 43. Marginal mean SE of the estimates of covariance of item intercept and item time intensity $\sigma_{\beta\zeta}$ at all levels of the manipulated factors.

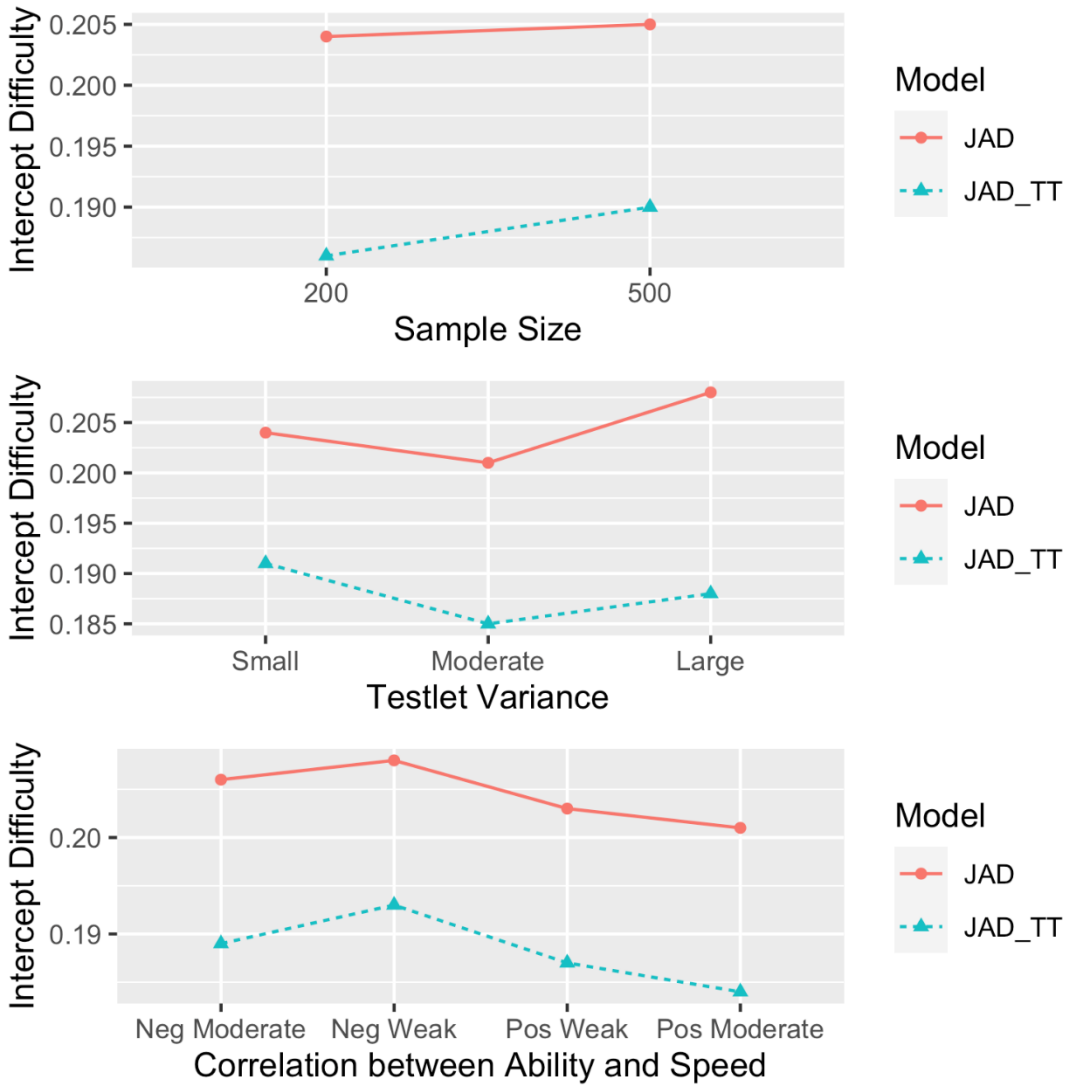

Figure 44 Marginal mean bias of the estimates of covariance of item intercept and item difficulty $\sigma_{\beta b}$ at all levels of the manipulated factors.

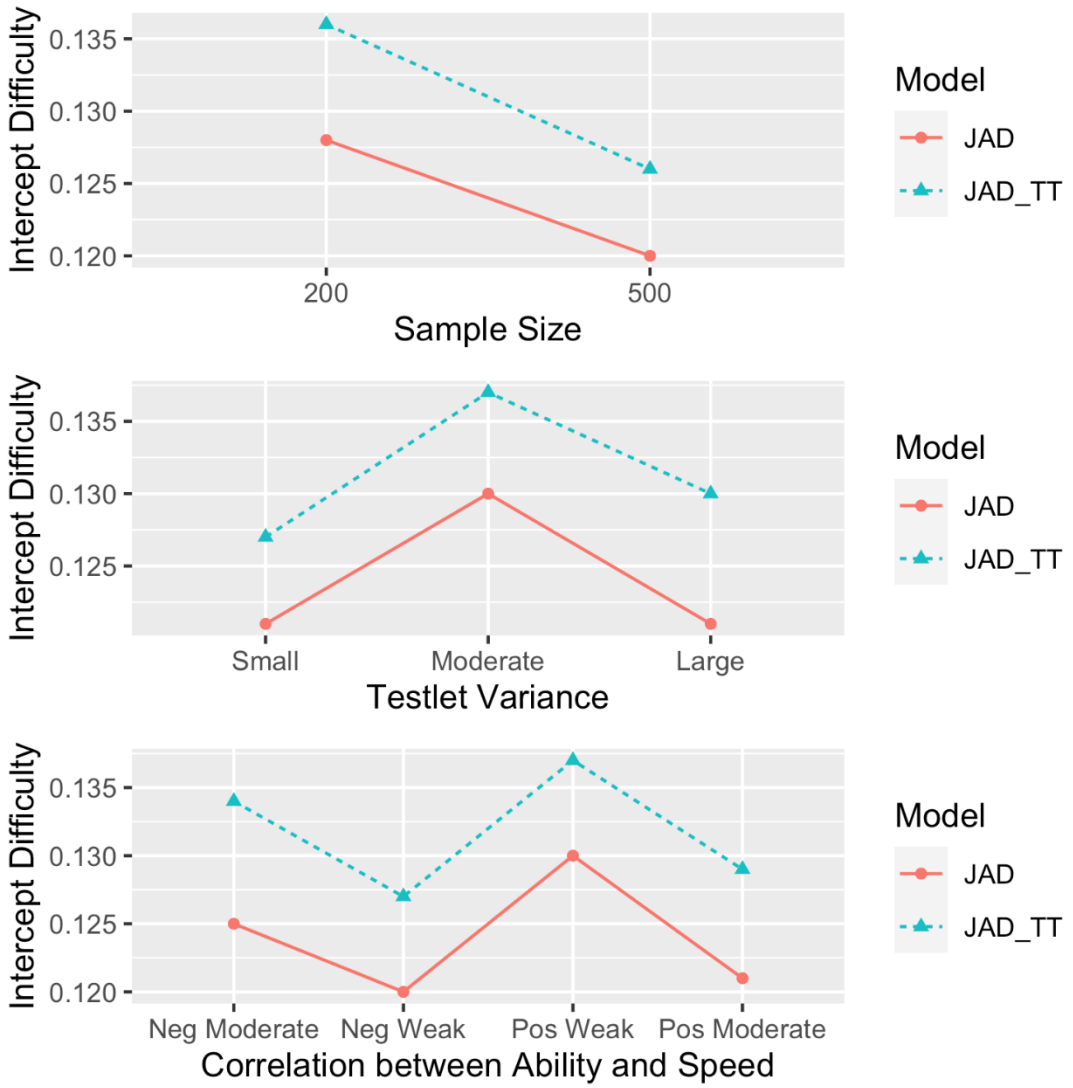

Figure 45. Marginal mean SE of the estimates of covariance of item intercept and item difficulty $\sigma_{\beta b}$ at all levels of the manipulated factors.

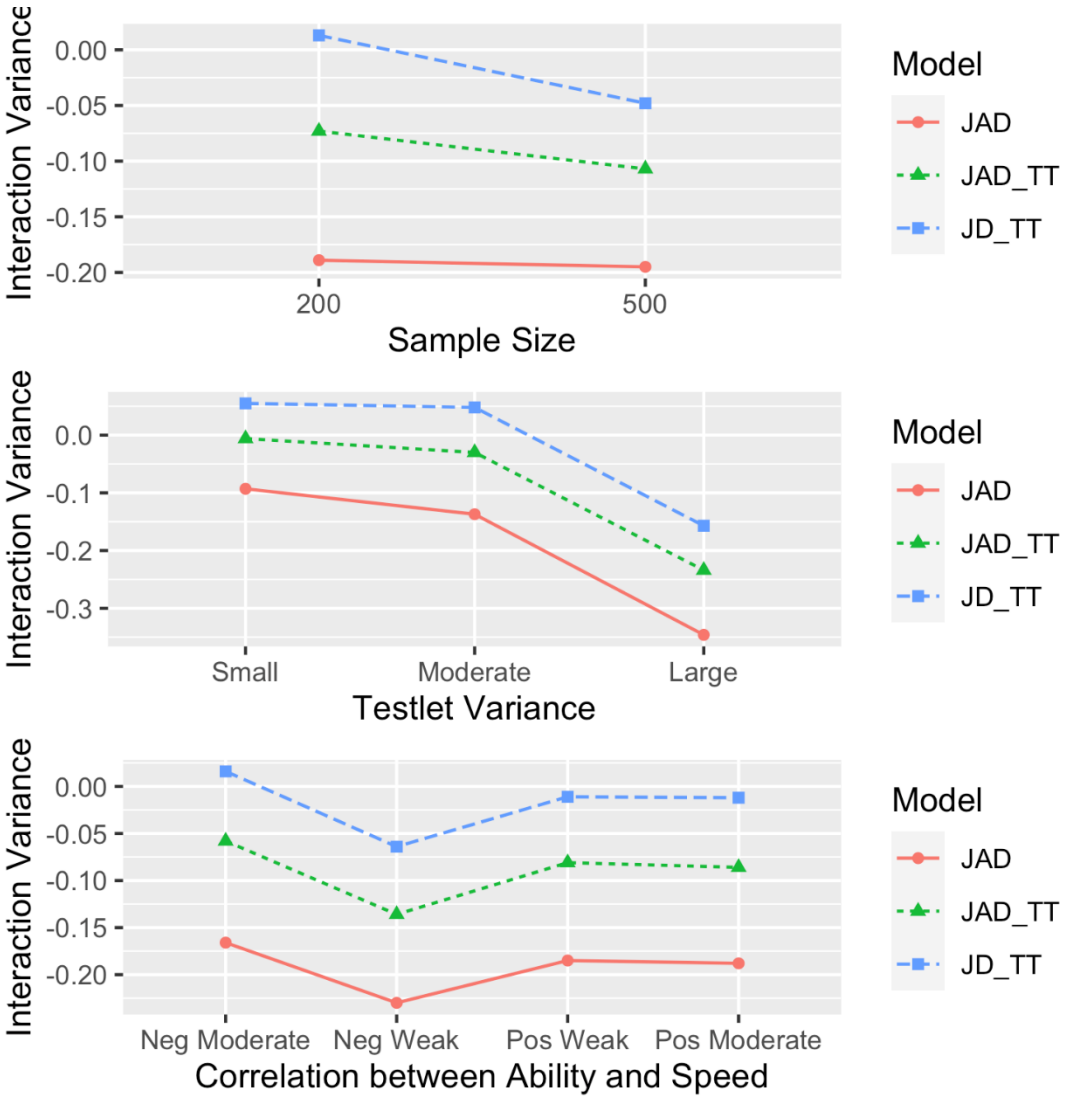

Figure 46. Marginal mean bias of the estimates of the variance of item interaction σ_δ^2 at all levels of the manipulated factors.

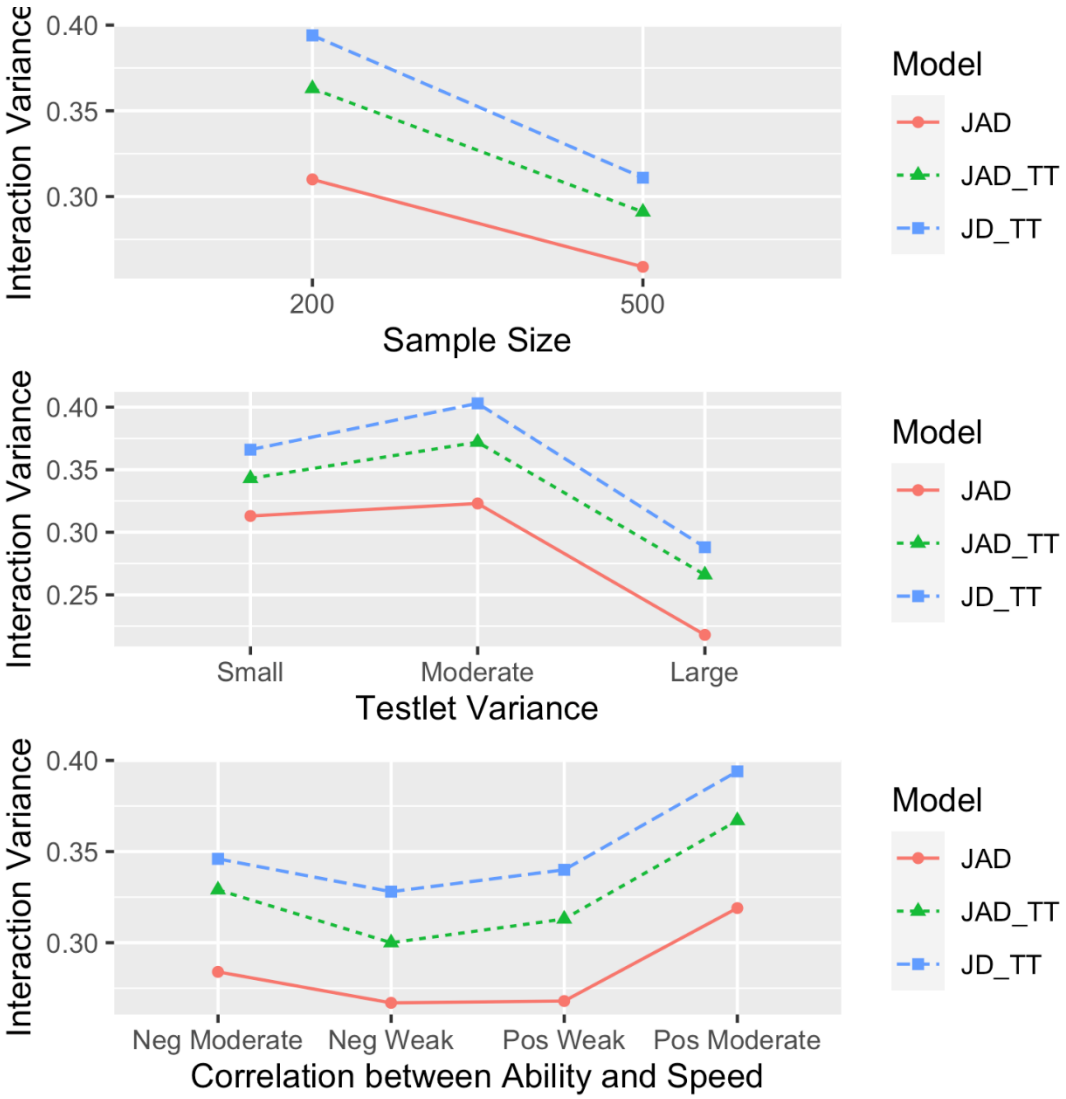

Figure 47 Marginal mean SE of the estimates of the variance of item interaction σ_{δ}^{2} at all levels of the manipulated factors.

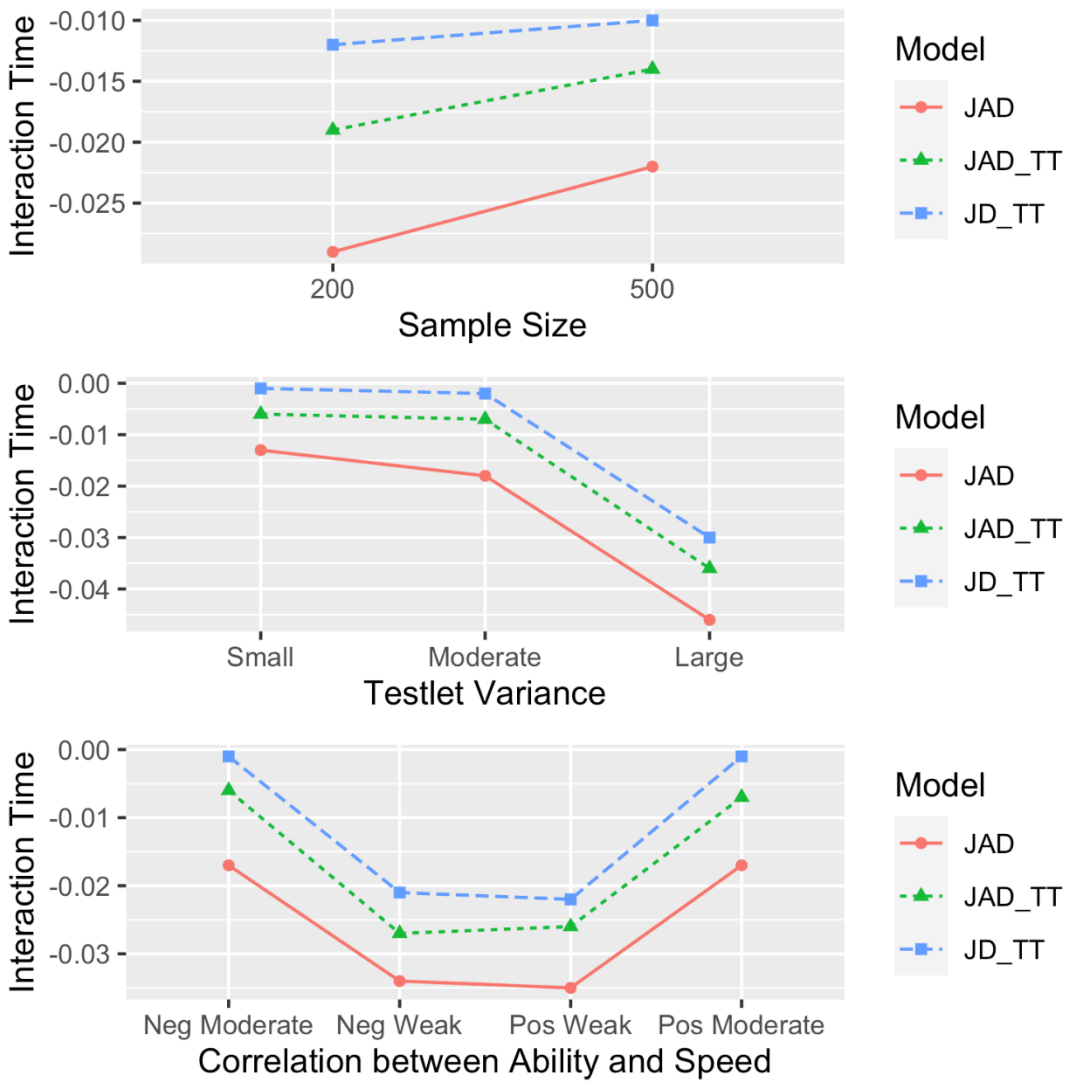

Figure 48. Marginal mean bias of the estimates of covariance of item interaction and item time intensity $\sigma_{\delta\zeta}$ at all levels of the manipulated factors.

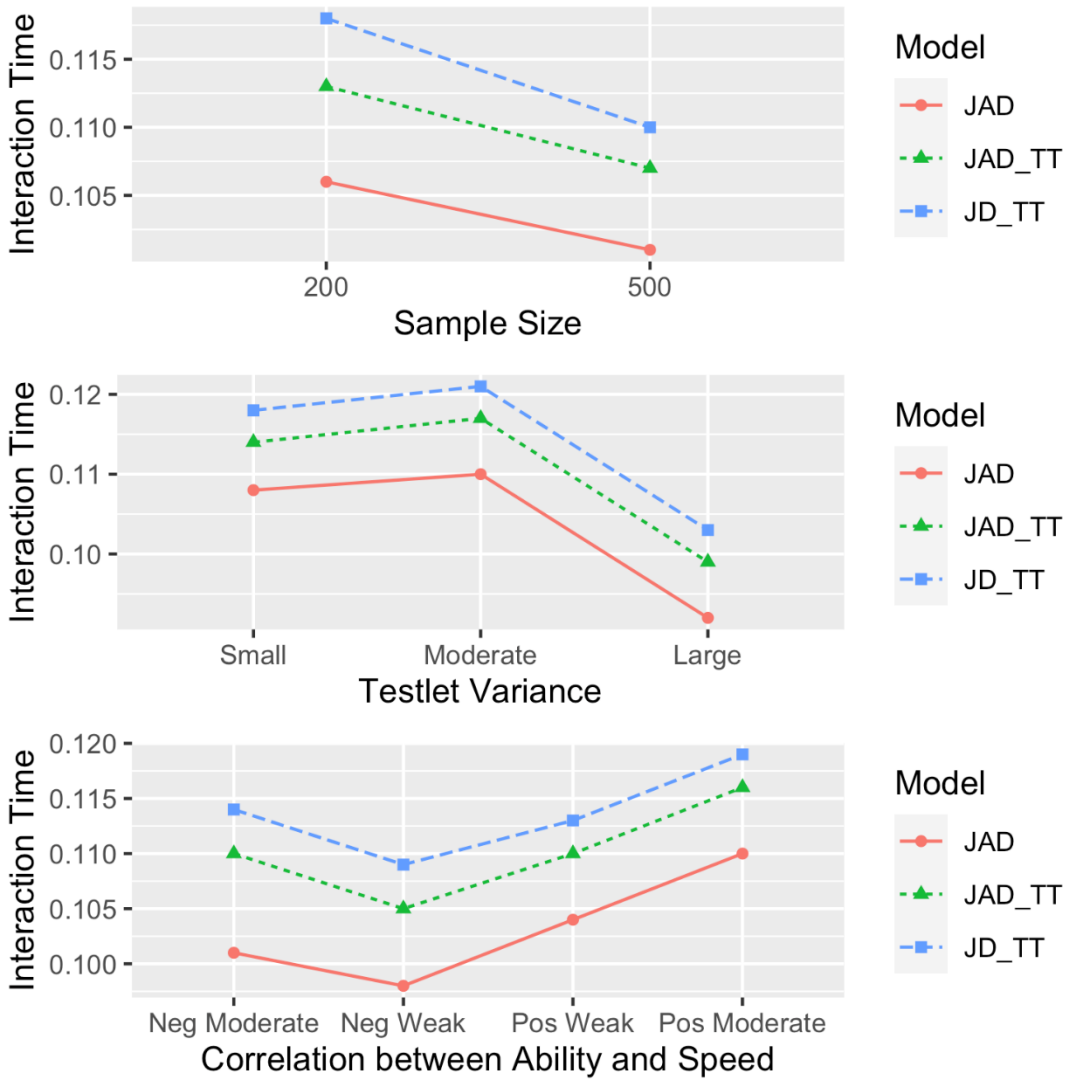

Figure 49. Marginal mean SE of the estimates of covariance of item interaction and item time intensity $\sigma_{\delta\zeta}$ at all levels of the manipulated factors.

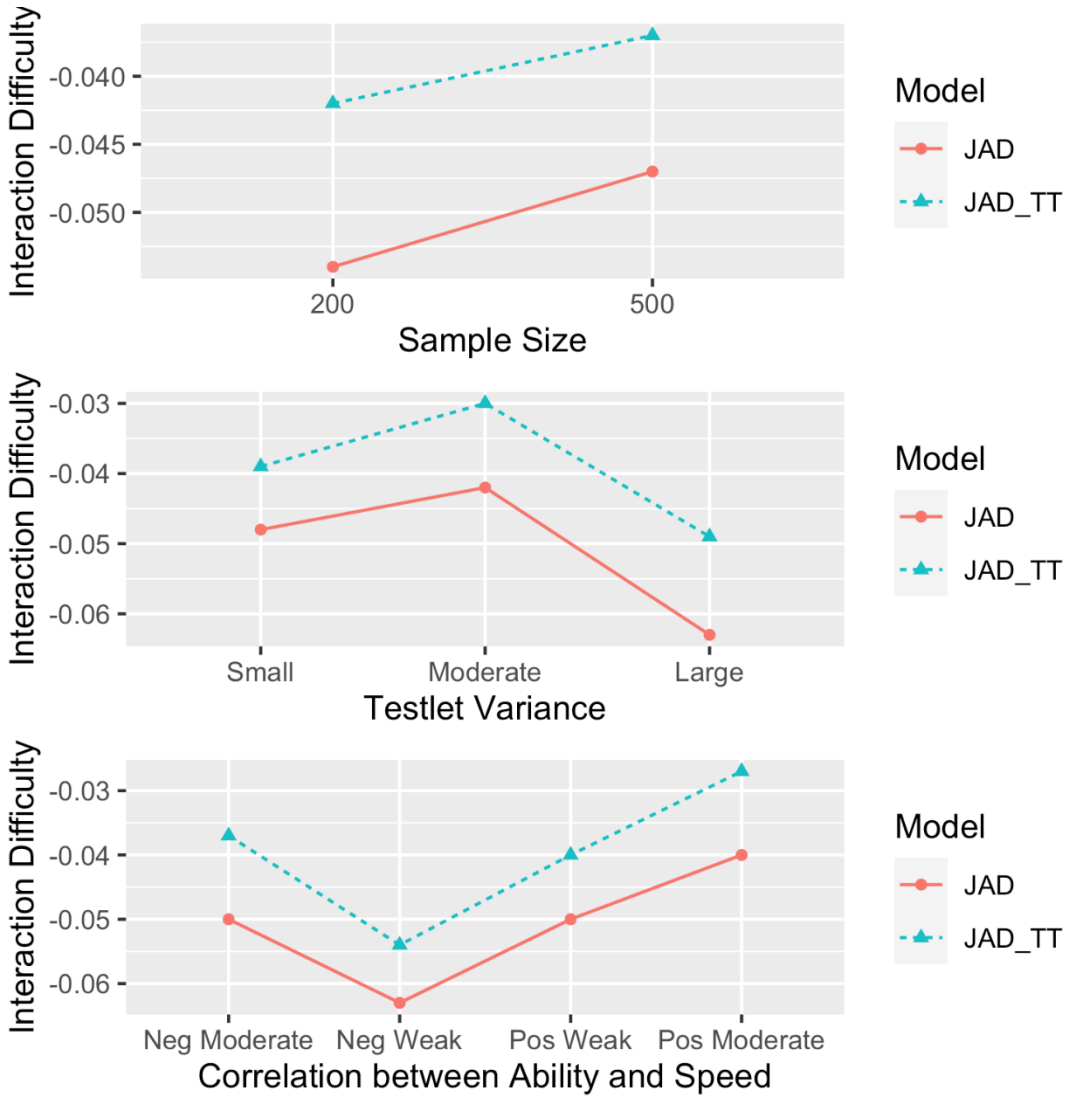

Figure 50. Marginal mean bias of the estimates of covariance of item interaction and item difficulty $\sigma_{\delta b}$ at all levels of the manipulated factors.

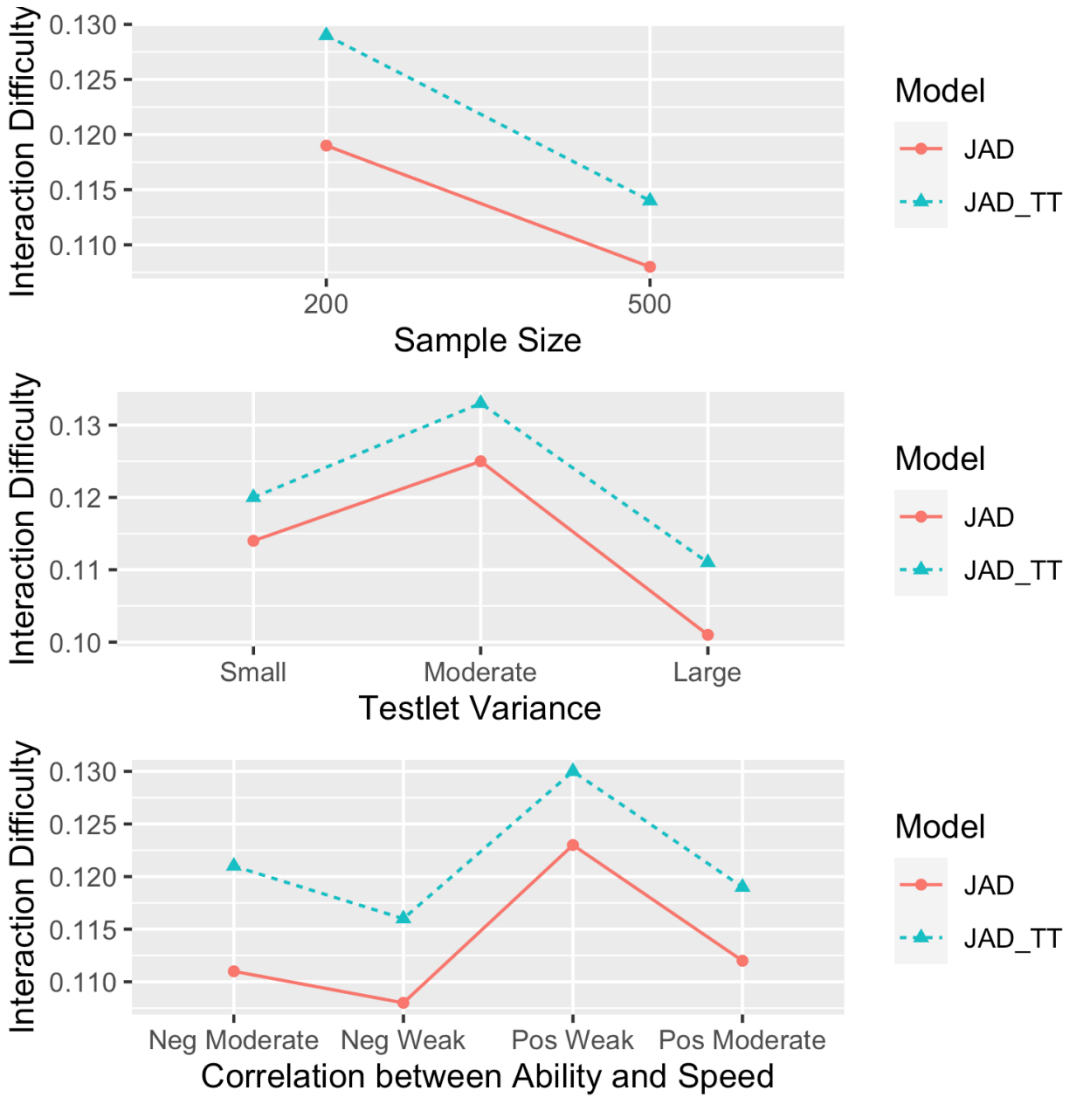

Figure 51. Marginal mean SE of the estimates of covariance of item interaction and item difficulty $\sigma_{\delta b}$ at all levels of the manipulated factors.

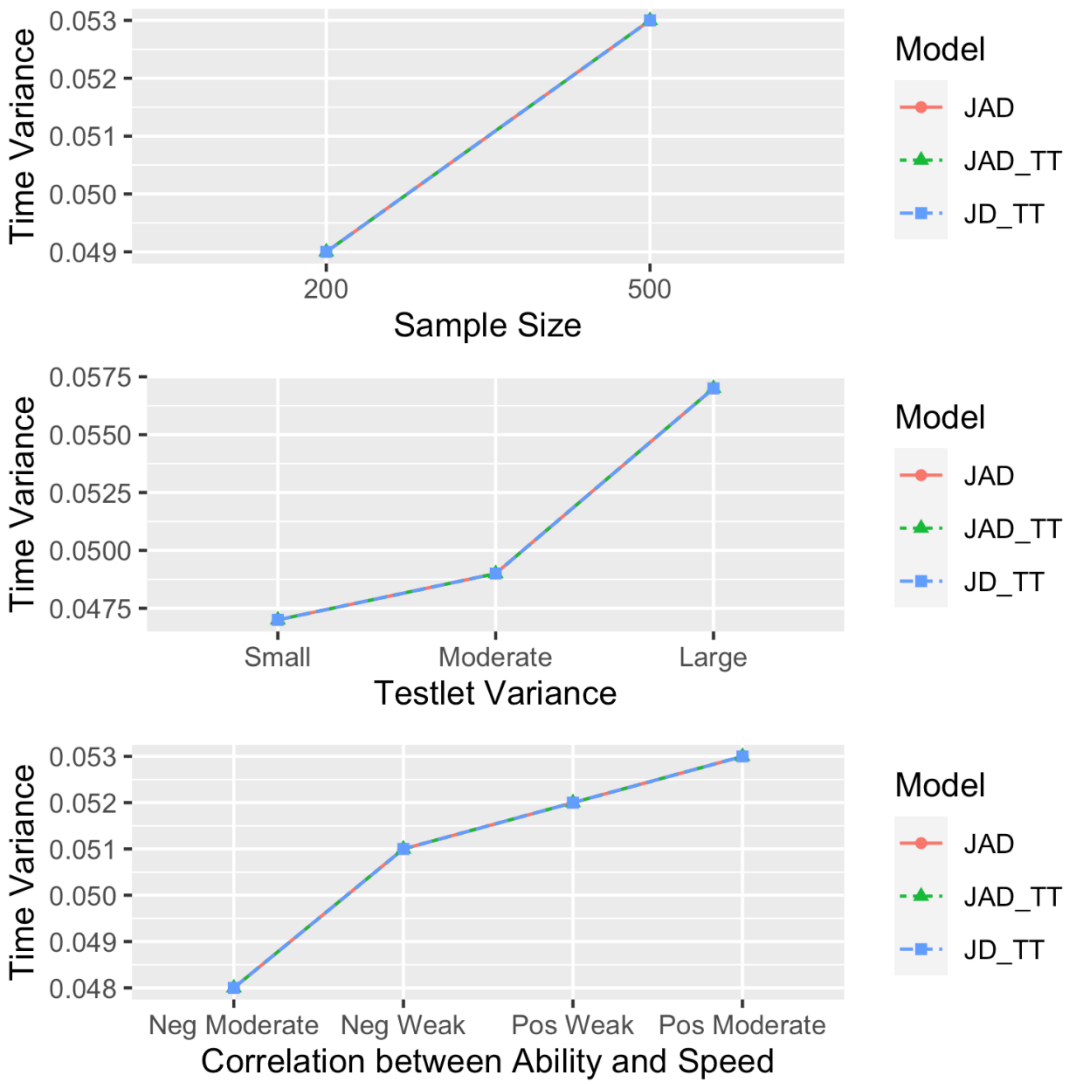

Figure 52. Marginal mean bias of the estimates of the variance of item time intensity σ_ζ^2 at all levels of the manipulated factors.

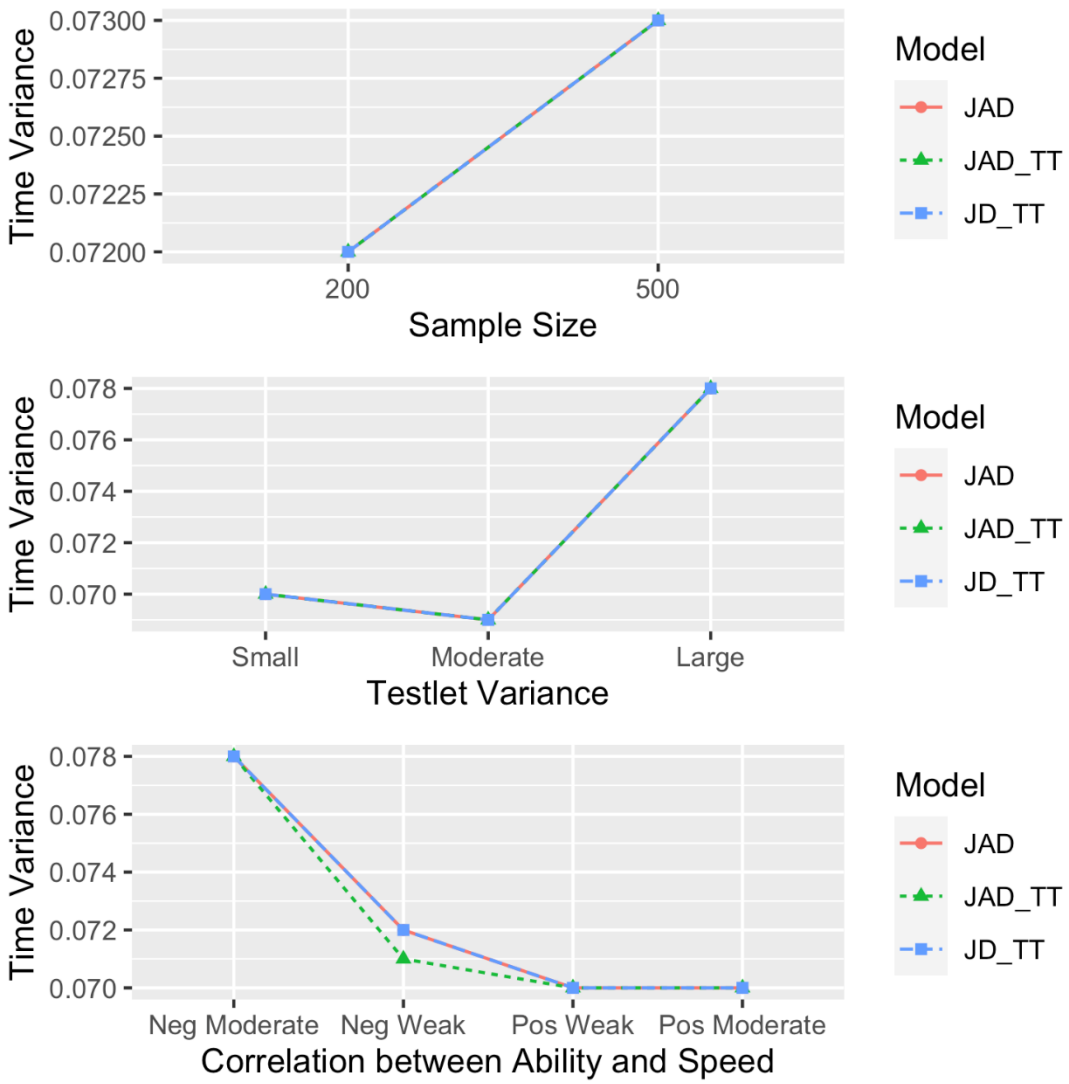

Figure 53. Marginal mean SE of the estimates of the variance of item time intensity σ_ζ^2 at all levels of the manipulated factors.

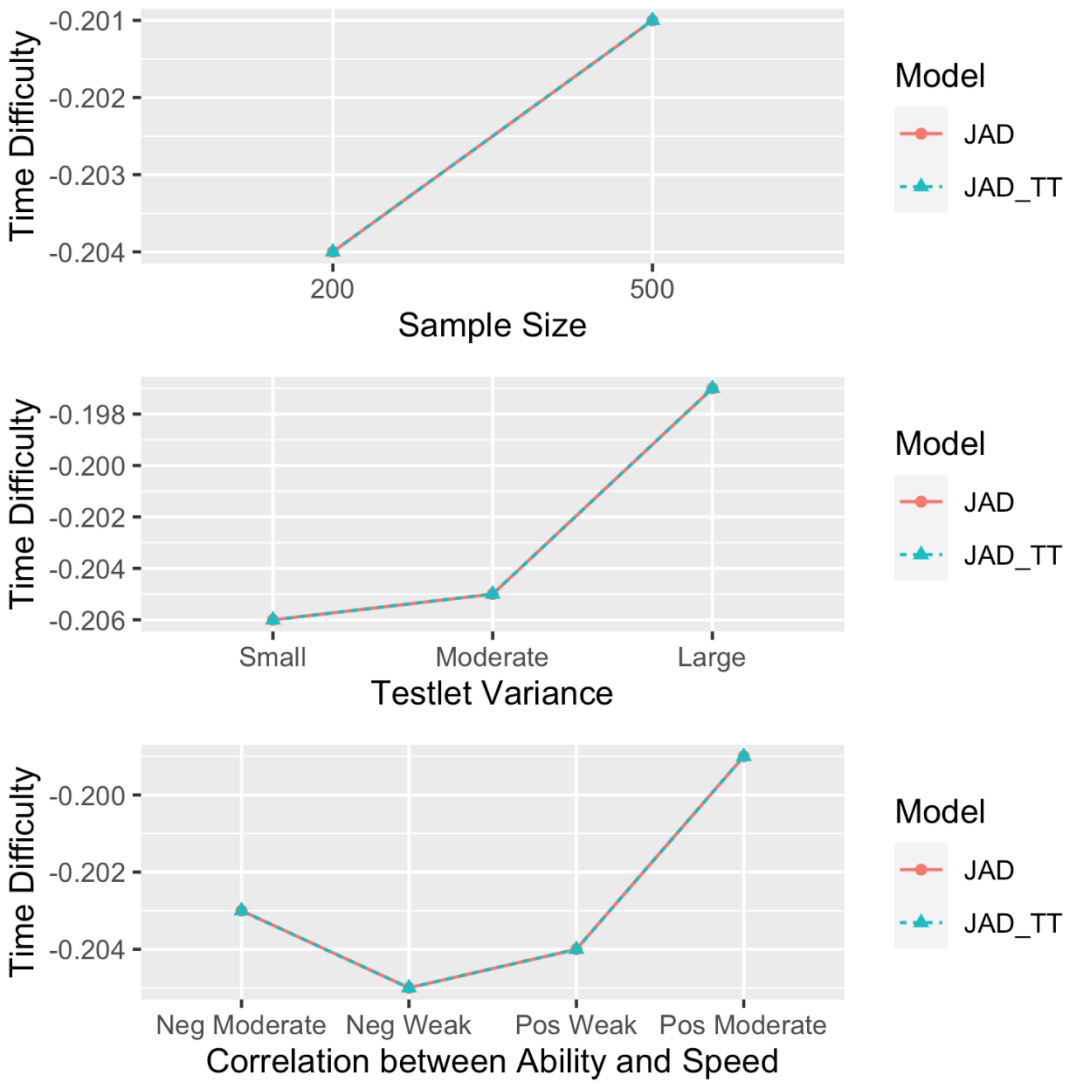

Figure 54. Marginal mean bias of the estimates of covariance of item time intensity and item difficulty $\sigma_{\zeta b}$ at all levels of the manipulated factors.

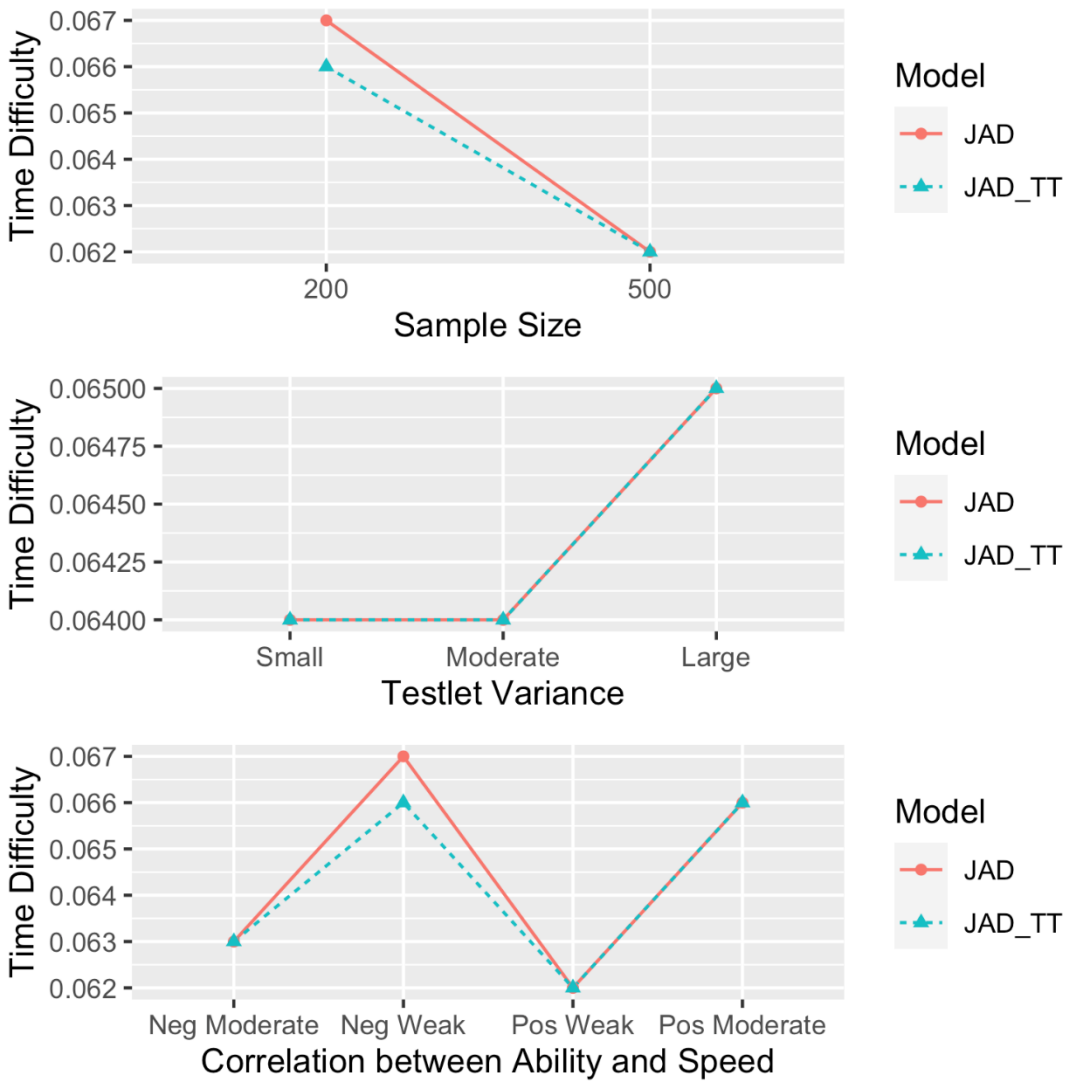

Figure 55 Marginal mean SE of the estimates of covariance of item time intensity and item difficulty $\sigma_{\zeta b}$ at all levels of the manipulated factors.

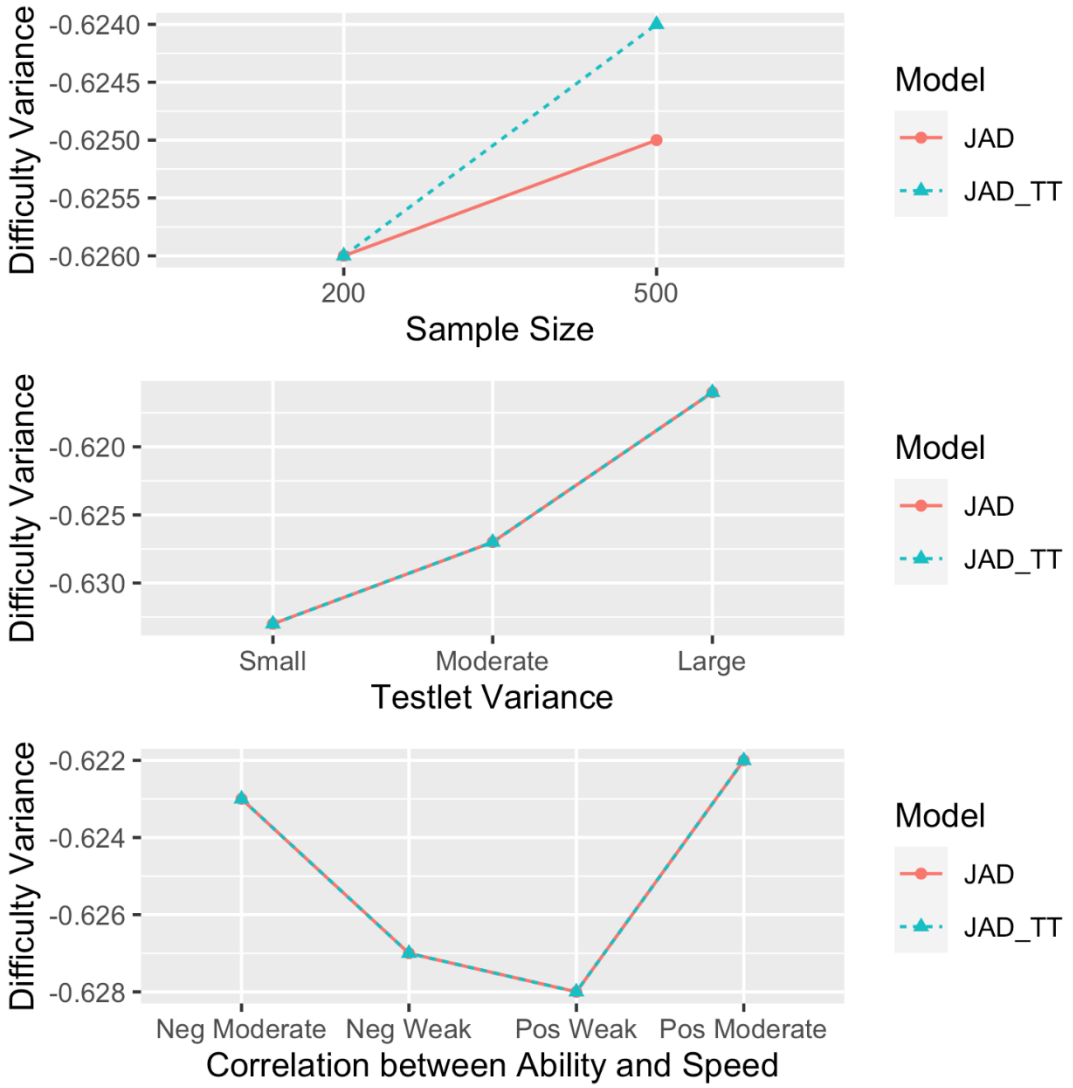

Figure 56. Marginal mean bias of the estimates of the variance of item difficulty σ_b^2 at all levels of the manipulated factors.

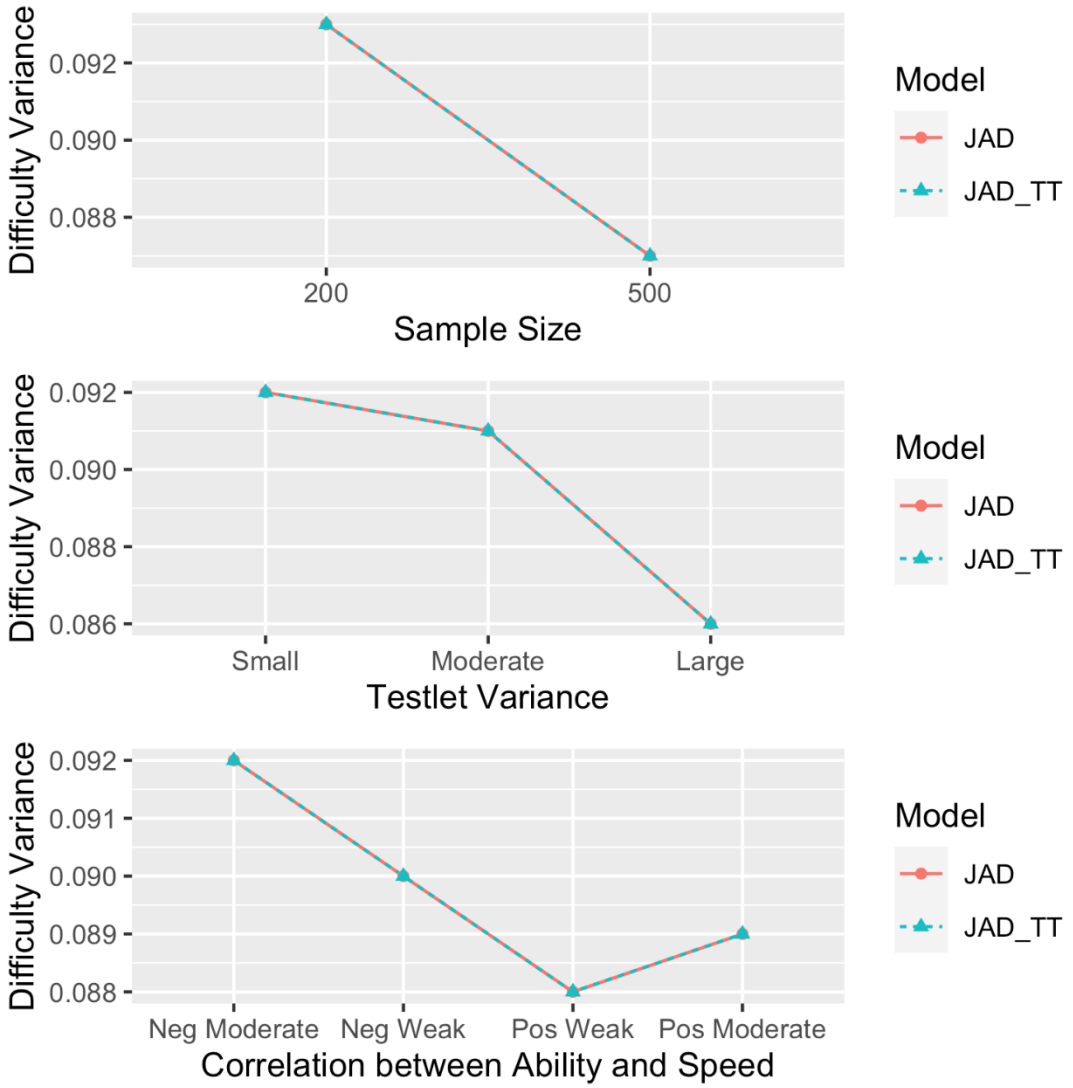

Figure 57. Marginal mean SE of the estimates of the variance of item difficulty σ_b^2 at all levels of the manipulated factors.

4.1.7 Recovery of the Testlet Variance/Covariance Matrix

As is discussed in Chapter 3, the response and RT testlet parameters are generated from a uniform bivariate normal distribution with their means fixed as zero and their correlation set as -0.5. Testlet variance is a manipulated factor having three levels: 0.25 representing small testlet effects, 0.5 indicating moderate testlet effects, and 1 representing large testlet effects. Figures 59 through 70 display the marginal mean bias and SE for estimates of the variance of the response testlet parameter σ_γ^2, of the variance of the RT testlet parameter σ_λ^2, and of their covariance $\sigma_{\gamma\lambda}$ generated by the two models specifying the testlet parameters at all levels of the manipulated factors, an examination of which would reveal how they vary by model type and by levels of the factors manipulated in this study. Tables A.12.1-3 and A.13.1-3 present their descriptive statistics under the 24 simulated conditions.

Variance of the Response Testlet Parameter. Figures 58 through 63 are visual representations of the marginal mean biases and SEs of the estimates of the variance of the response testlet parameter σ_γ^2. As is shown in the figures, consistent across levels of the manipulated factors, the marginal mean biases of the estimates generated by the proposed model are placed lower than those generated by the JRT-DINA-R/RT/AC model. Thus when the marginal mean bias is negative, the absolute value of the marginal mean bias of estimation by the proposed model is higher than the JRT-DINA-R/RT/AC model. When they are positive, as when the testlet variance is small, estimation by the proposed model recovers σ_γ^2 with less systematic

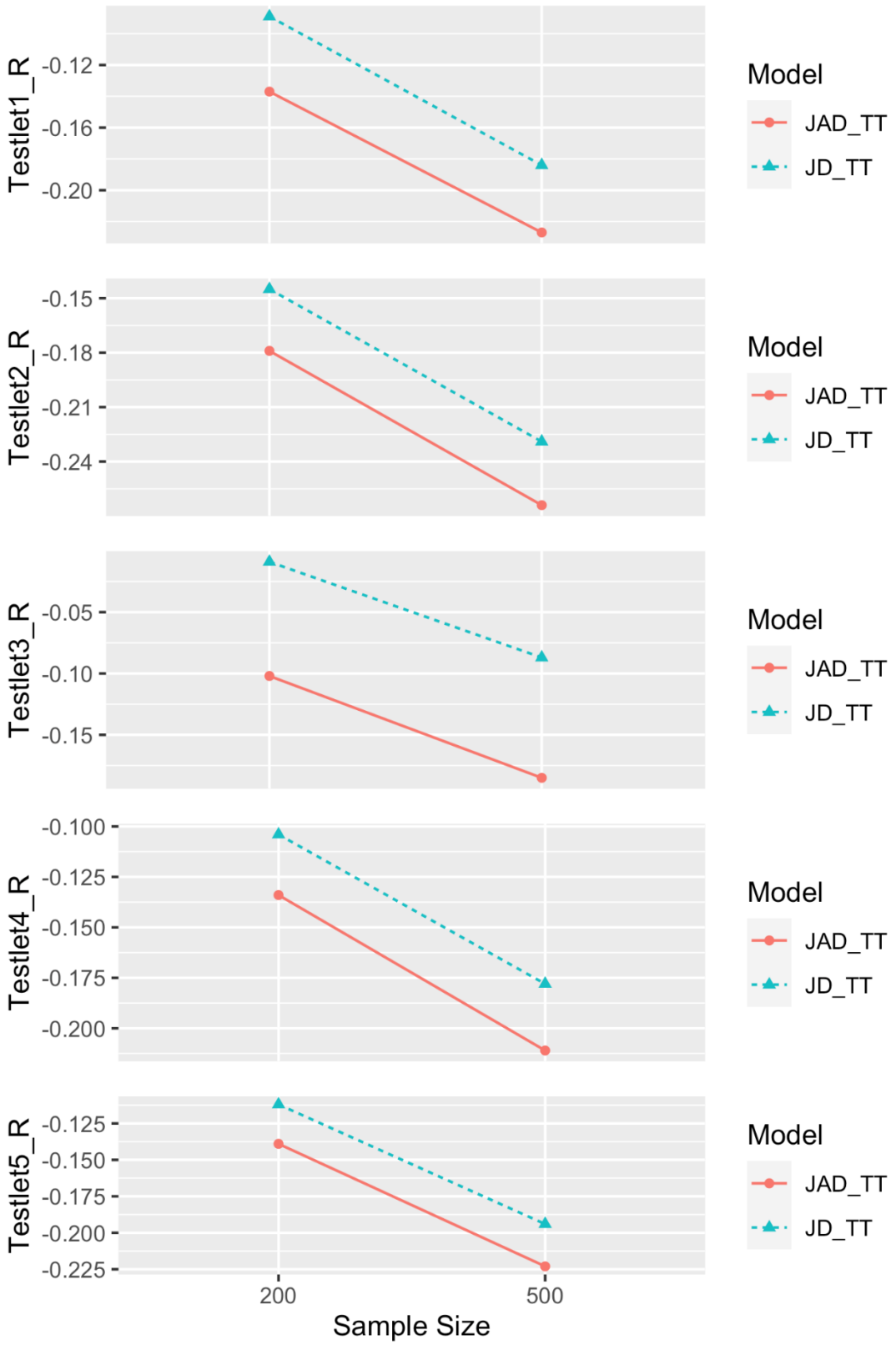

Figure 58 Marginal mean bias of the estimates of the variance of the testlet effects for responses for each of the five testlets at two sample size levels.

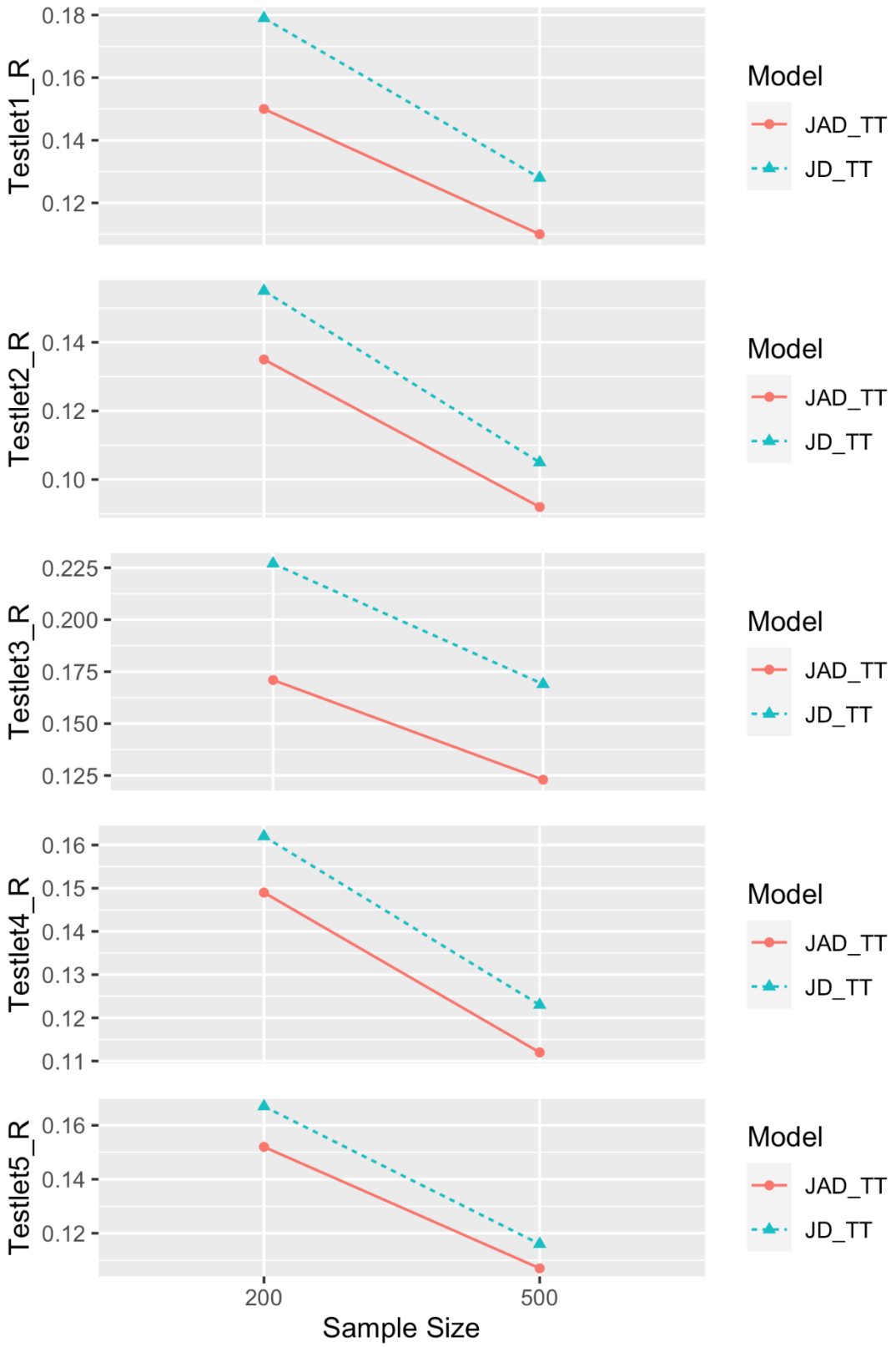

Figure 59. Marginal mean SE of the estimates of the variance of the testlet effects for responses for each of the five testlets at two sample size levels.

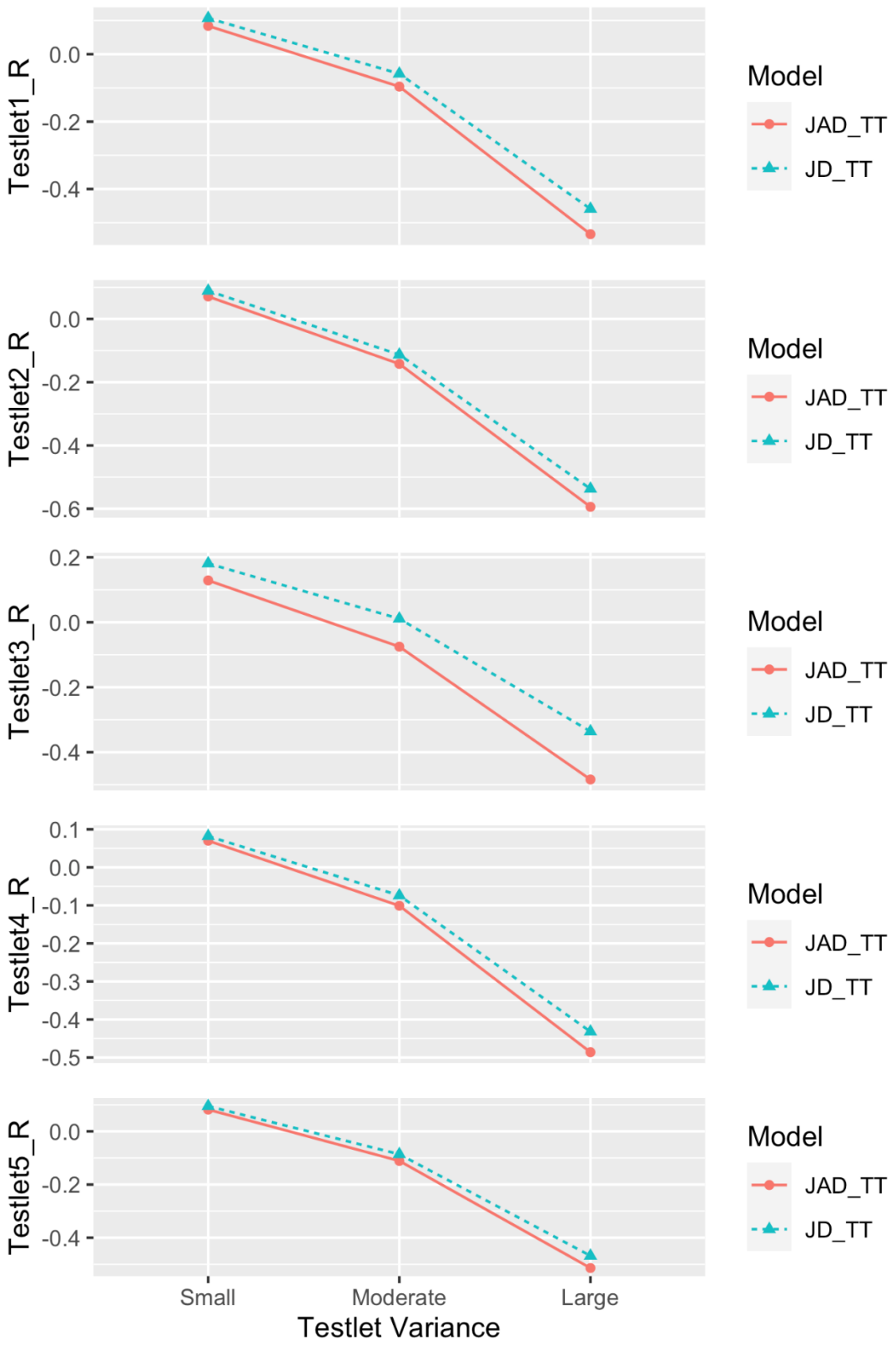

Figure 60 Marginal mean bias of the estimates of the variance of the testlet effects for responses for each of the five testlets at three testlet variance levels.

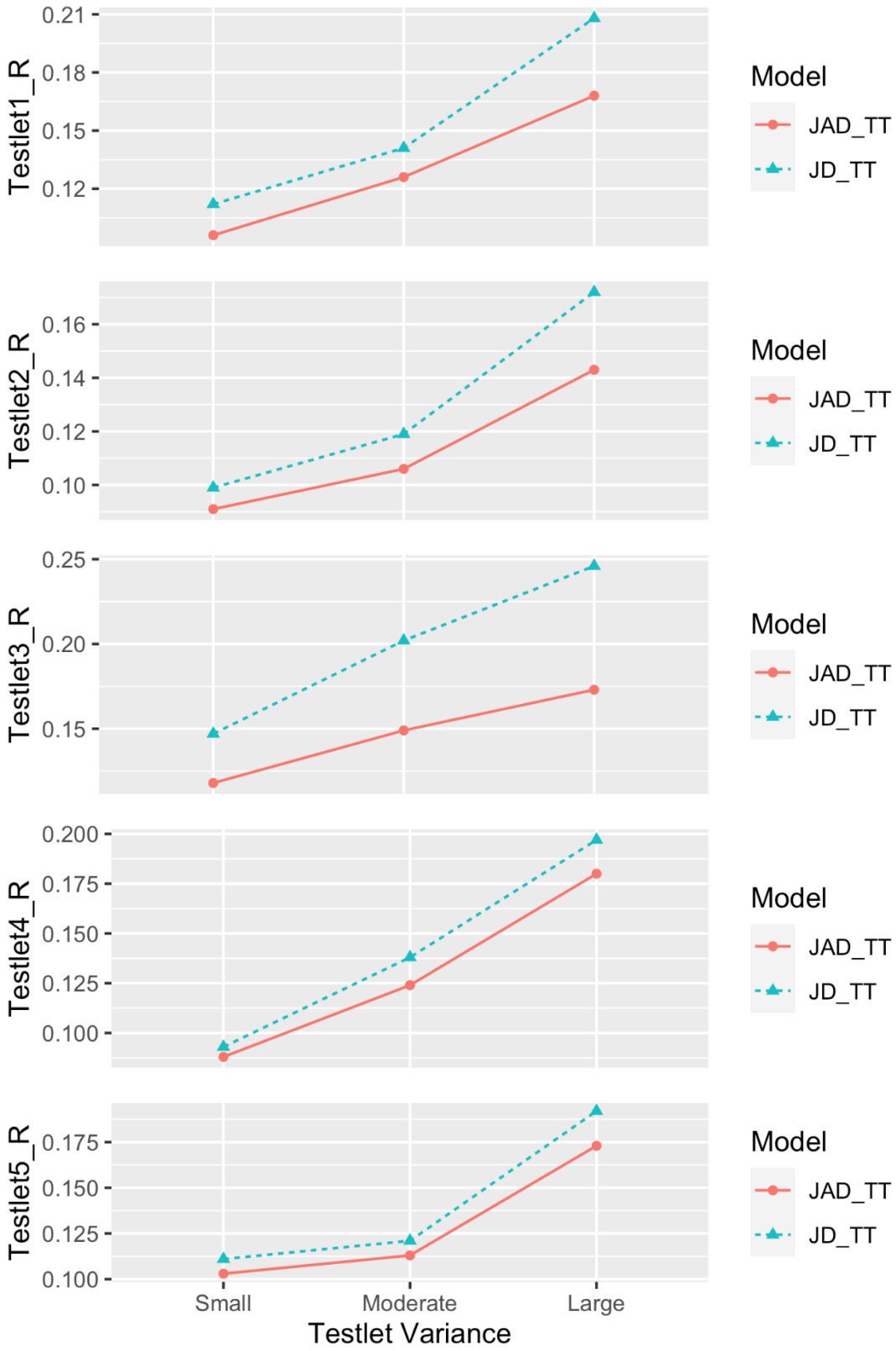

Figure 61. Marginal mean SE of the estimates of the variance of the testlet effects for responses for each of the five testlets at three testlet variance levels.

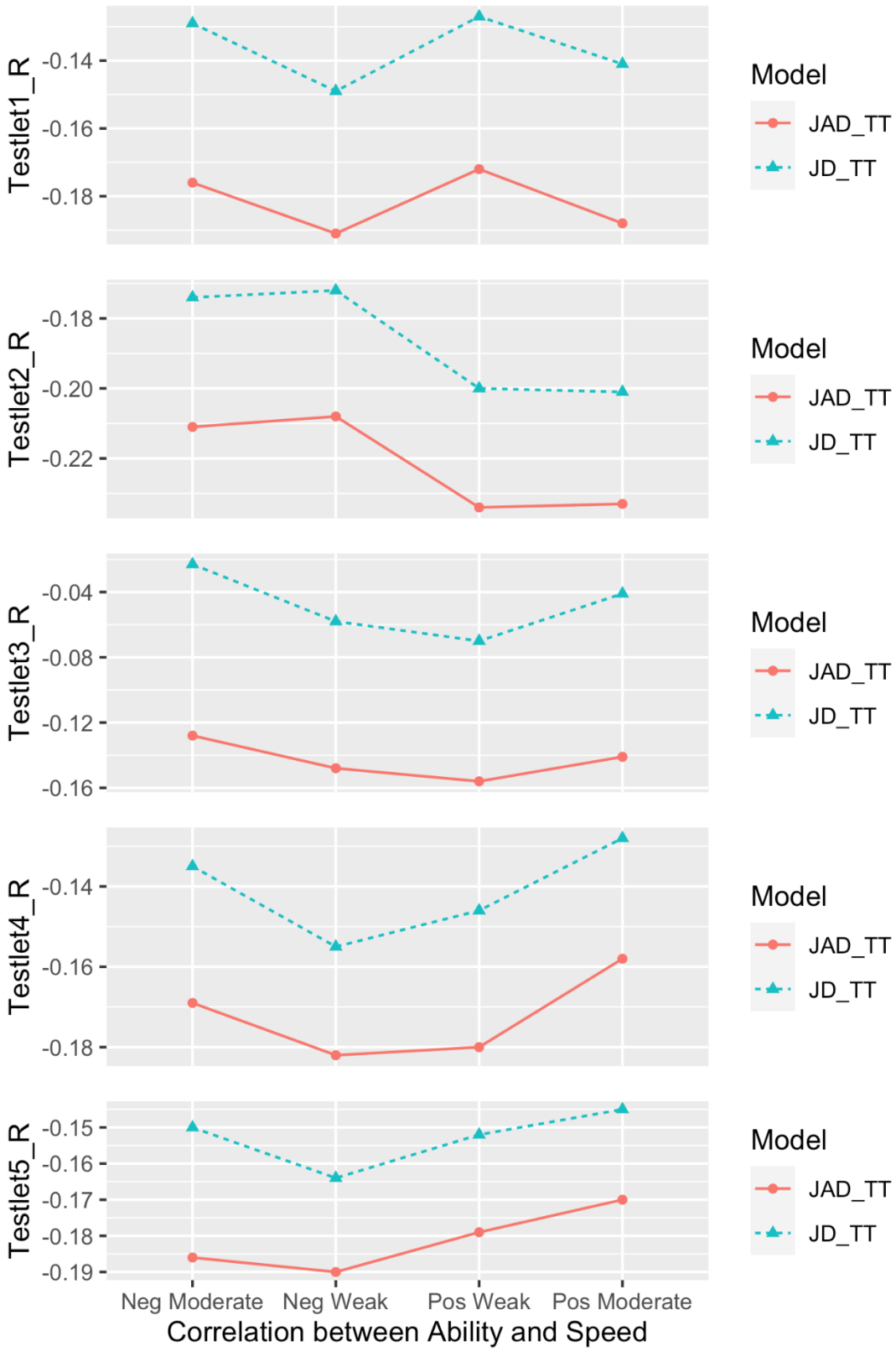

Figure 62. Marginal mean bias of the estimates of the variance of the testlet effects for responses for each of the five testlets at four correlation levels.

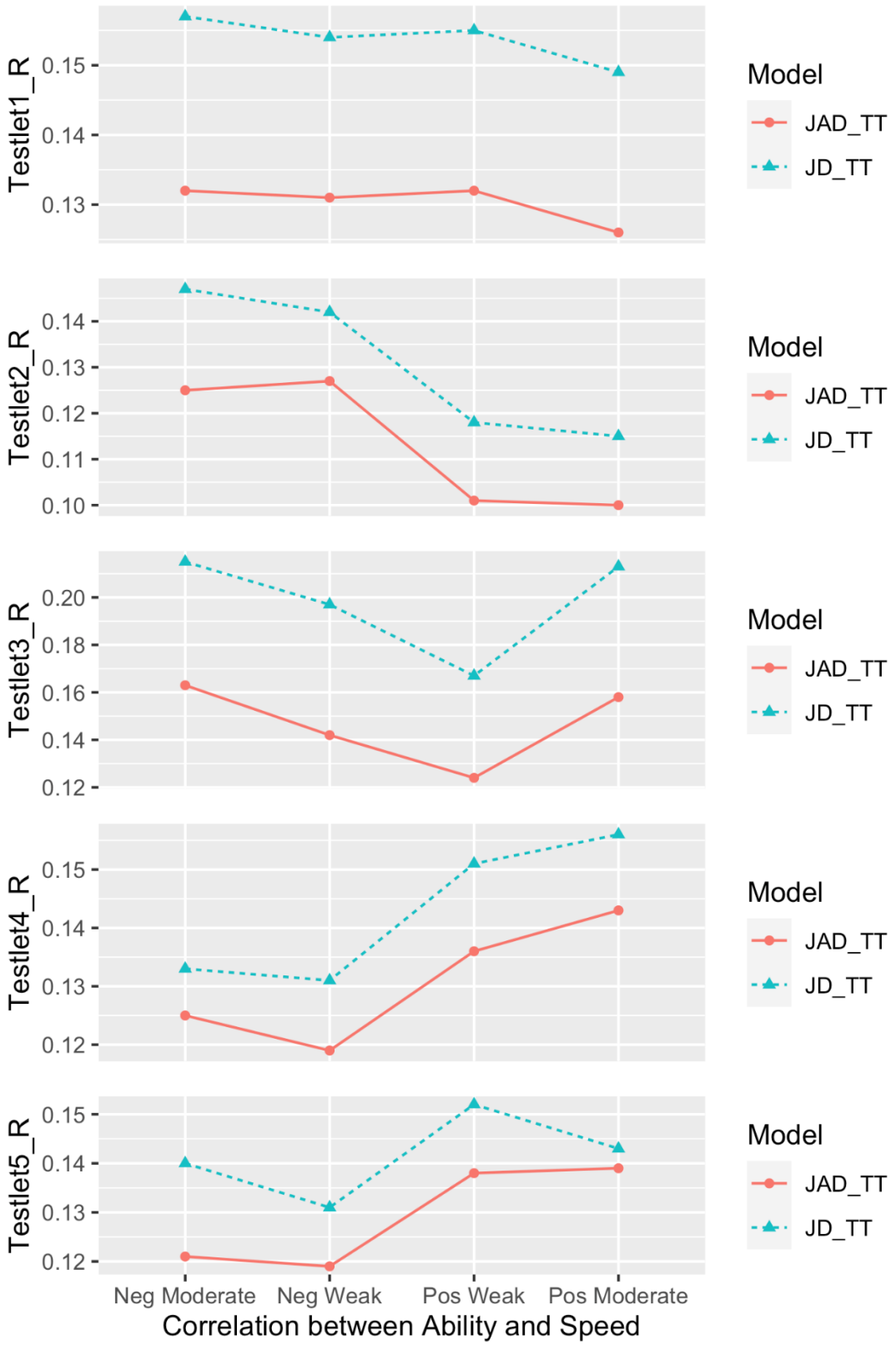

Figure 63. Marginal mean SE of the estimates of the variance of the testlet effects for responses for each of the five testlets at four correlation levels.

error than the JRT-DINA-R/RT/AC model. Regardless of the levels of the manipulated factors, the marginal mean SE for estimates generated by the proposed model is smaller than those generated by the JRT-DINA-R/RT/AC model, suggesting that the proposed model recovers σ_γ^2 with less random error than the JRT-DINA-R/RT/AC model.

Variation by levels of the manipulated factors is consistent across the five testlets. As the sample size increases from 200 to 500, the marginal mean biases and SEs of the estimates of σ_γ^2 drop, suggesting increased systematic error for marginal mean biases that are negative, and decreased random error. Variation by testlet variance shows similarly consistent pattern: as testlet variance increases from small to large, marginal mean biases drop, which suggests increased systematic error for values that are negative, whereas the marginal mean SEs increase as the testlet variance increases, indicating reduced random error associated with a drop in the testlet variance level. Variation by levels of $\rho_{\theta\tau}$ is specific to the testlet being evaluated yet consistent across the models. For instance, for testlet 1, the marginal mean biases at the levels of negative moderate $\rho_{\theta\tau}$ and positive weak $\rho_{\theta\tau}$ are similar and lower than those at the other two levels. Yet for testlet 2, they are similar for the two levels of negative $\rho_{\theta\tau}$ and for the two positive $\rho_{\theta\tau}$ levels, with the negative levels showing less systematic error than the positive levels. Marginal SEs show similar testlet-specific variation that is consistent across the two models.

Variance of the RT Testlet Parameters. Figures 65 through 70 are visual representations of marginal mean biases and SEs for estimates of σ_λ^2 by the two models for the five testlets and at all levels of the manipulated factors. As is shown in

the figures, model specification has little or no impact on the systematic and random error for estimating σ_λ^2 quantified by the marginal mean biases and SEs: consistent across the levels of the manipulated factors they are identical or close to identical for the two models. Variation the marginal mean biases by levels of the sample size and testlet variance shows patterns that are similar to those for the estimates of σ_γ^2: they drop as the sample size increases from 200 to 500 and as testlet variance increases from small to large, suggesting increased systematic error associated with a larger sample size and larger testlet variance. The marginal mean SEs similarly drop as sample size increases from 200 to 500, suggesting a reduction in random error associated with a larger sample size. However, they show little or no variation by levels testlet variance: across the two model, the marginal mean SEs of the estimates are identical or close across the levels of the testlet variance. Variation in the marginal mean bias and SE of the estimates generated by the two models by levels of $\rho_{\theta\tau}$ is minimal as they are identical or close to identical across the levels of $\rho_{\theta\tau}$.

Covariance of the Response and RT Testlet Parameters. Figures 71 through 76 display the marginal mean biases and SEs of the estimates of $\sigma_{\lambda\gamma}$ generated by the two models for the five testlets and at all levels of the manipulated factors. As is shown in the figures, model specification has little or no impact on the marginal mean biases and SEs of the estimates for this parameter: across the two sample sizes and levels of the testlet variane, they are identical or close to identical for the two models being compared, suggesting little or no impact.

Figure 74 shows that at specific levels of $\rho_{\theta\tau}$ the proposed model recovers $\sigma_{\lambda\gamma}$ with less systematic error than the JRT-DINA-R/RT/AC model: the marginal mean

bias for estimates of $\sigma_{\lambda\gamma}$ generated by the proposed model is lower at positive moderate $\rho_{\theta\tau}$ across the five testlets and at positive weak $\rho_{\theta\tau}$ for testlets 1 through 4. At other levels of $\rho_{\theta\tau}$ marginal mean biases for the JRT-DINA-R/RT/AC model are less than or equal to those for the proposed model. Discrepancy in the marginal mean SEs of the $\sigma_{\lambda\gamma}$ estimates generated by the two models is also evident if examined by testlet variance: across the five testlets and levels of the testlet variance, they are either lower for the proposed model or identical for the two models. The discrepancy, however, is negligible.

Of the three manipulated fators, variation by sample size is evident in the marginal mean SEs of the estimates: consistent across the five testlets and the two models, as sample size increases from 200 to 500, their marginal mean SE drops, suggesting an increase in the sample size results in reduced random error for estimating this parameter. Variation by levels of the testlet variance shows similarly uniform pattern: as testlet variance increases from small to large, marginal mean SE also increases, suggesting that increasing testlet variance leads to increased random error in estimating $\sigma_{\lambda\gamma}$.Variation by levels of $\rho_{\theta\tau}$ is specific to the model and the testlet and does not form a uniform pattern, as is shown in Figures 75 and 76.

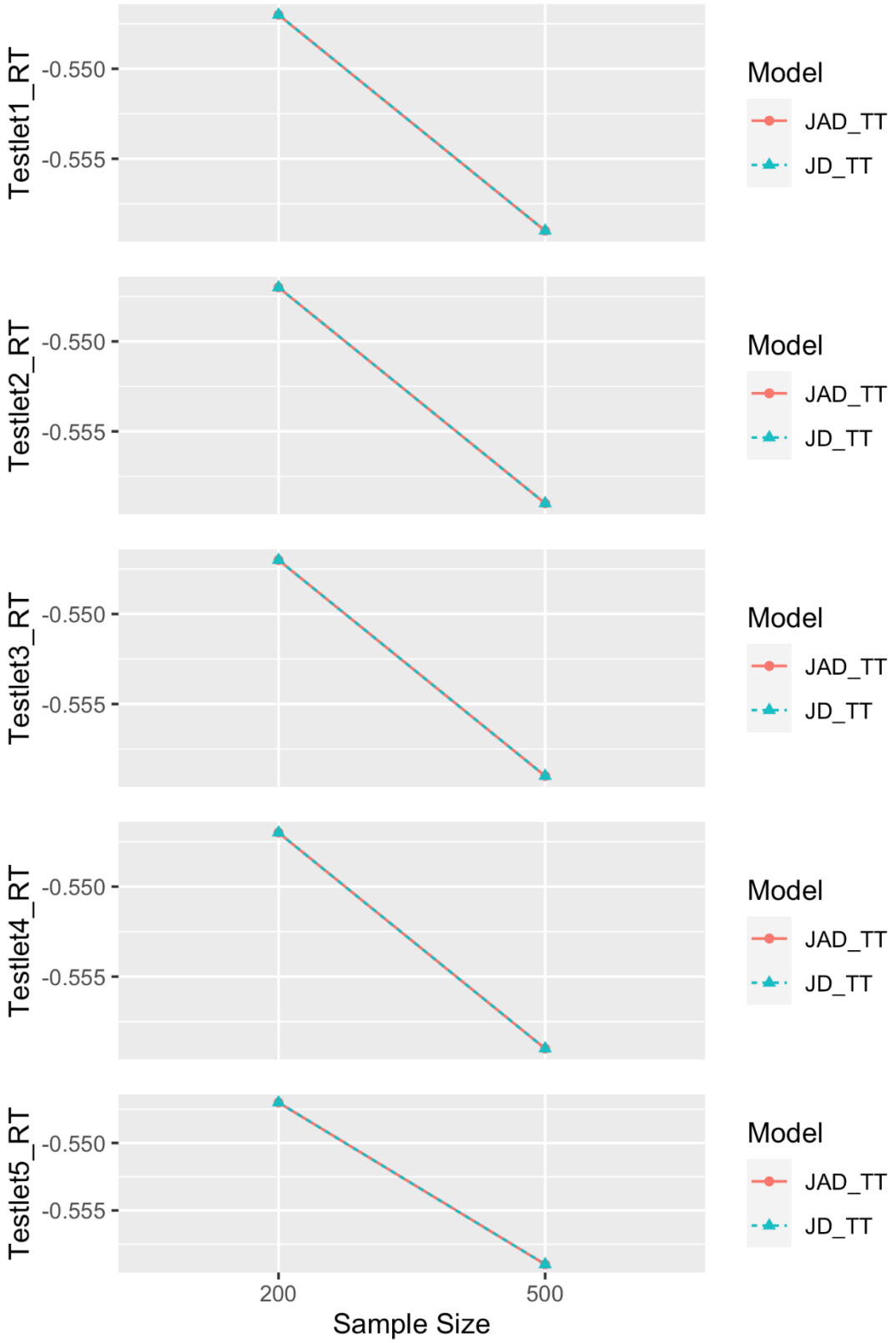

Figure 64 Marginal mean bias of the estimates of the variance of the testlet effects for response time for each of the five testlets at two sample size levels.

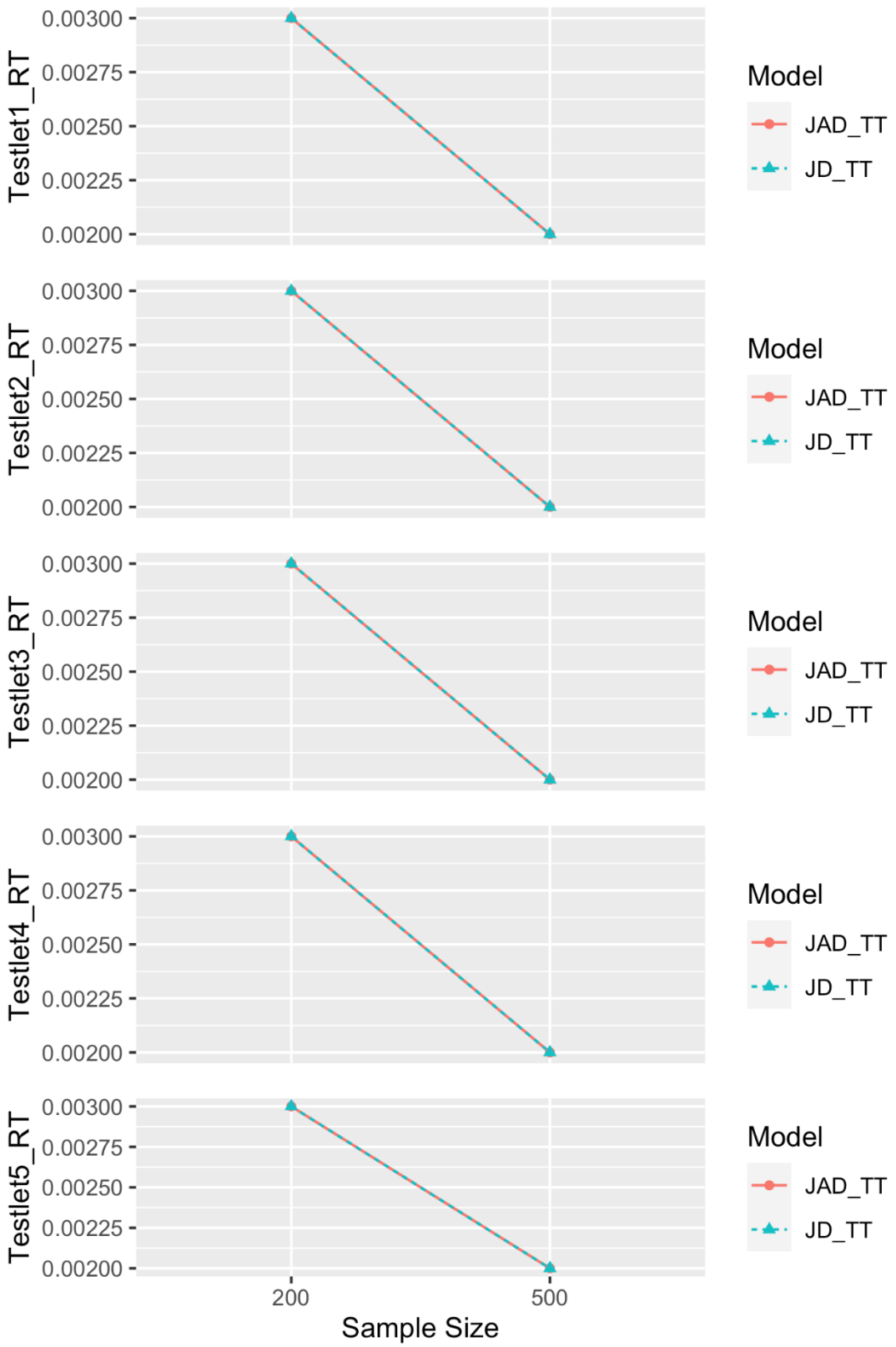

Figure 65. Marginal mean SE of the estimates of the variance of the testlet effects for response time for each of the five testlets at two sample size levels.

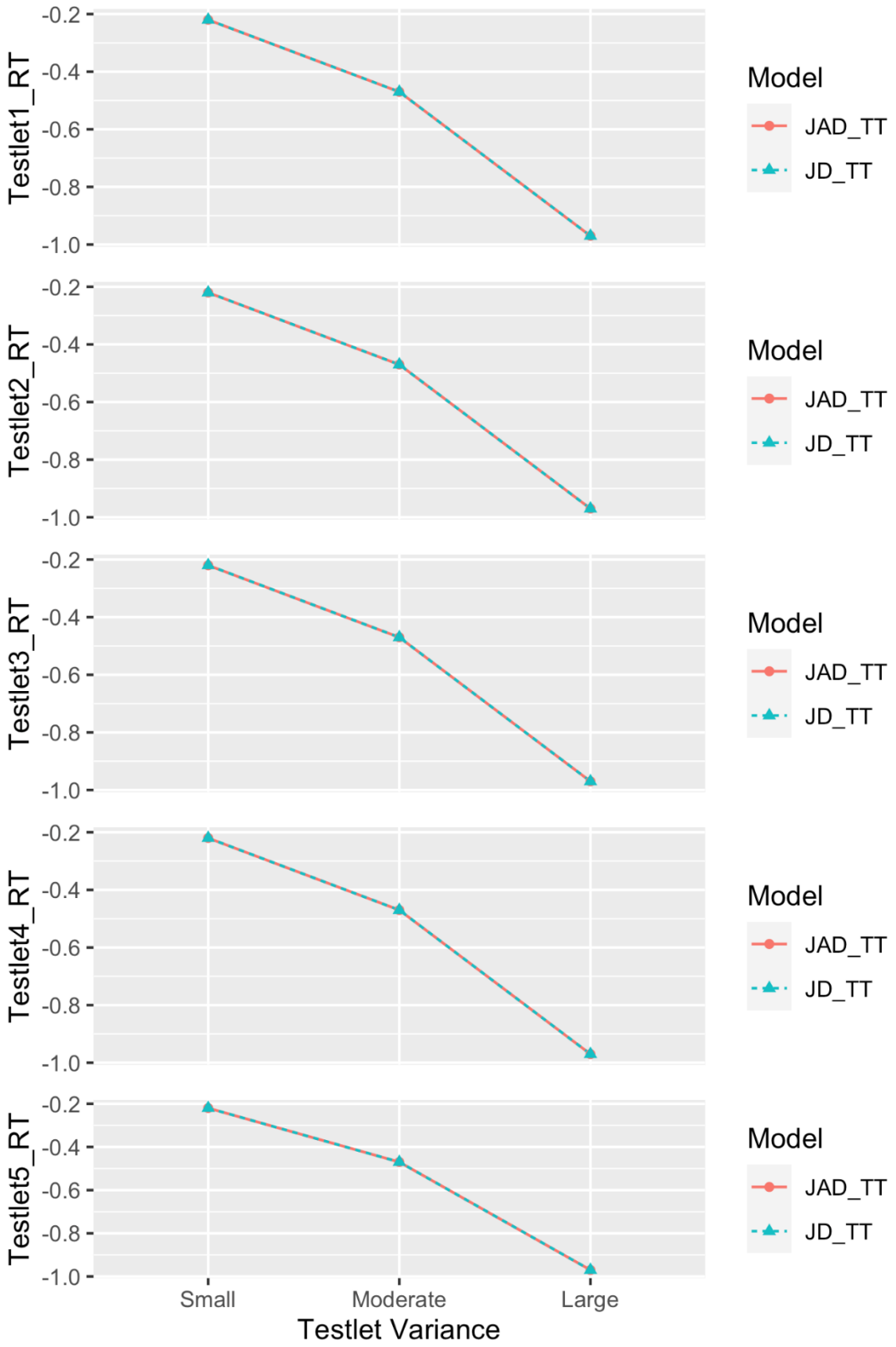

Figure 66 Marginal mean bias of the estimates of the variance of the testlet effects for response time for each of the five testlets at three testlet variance levels.

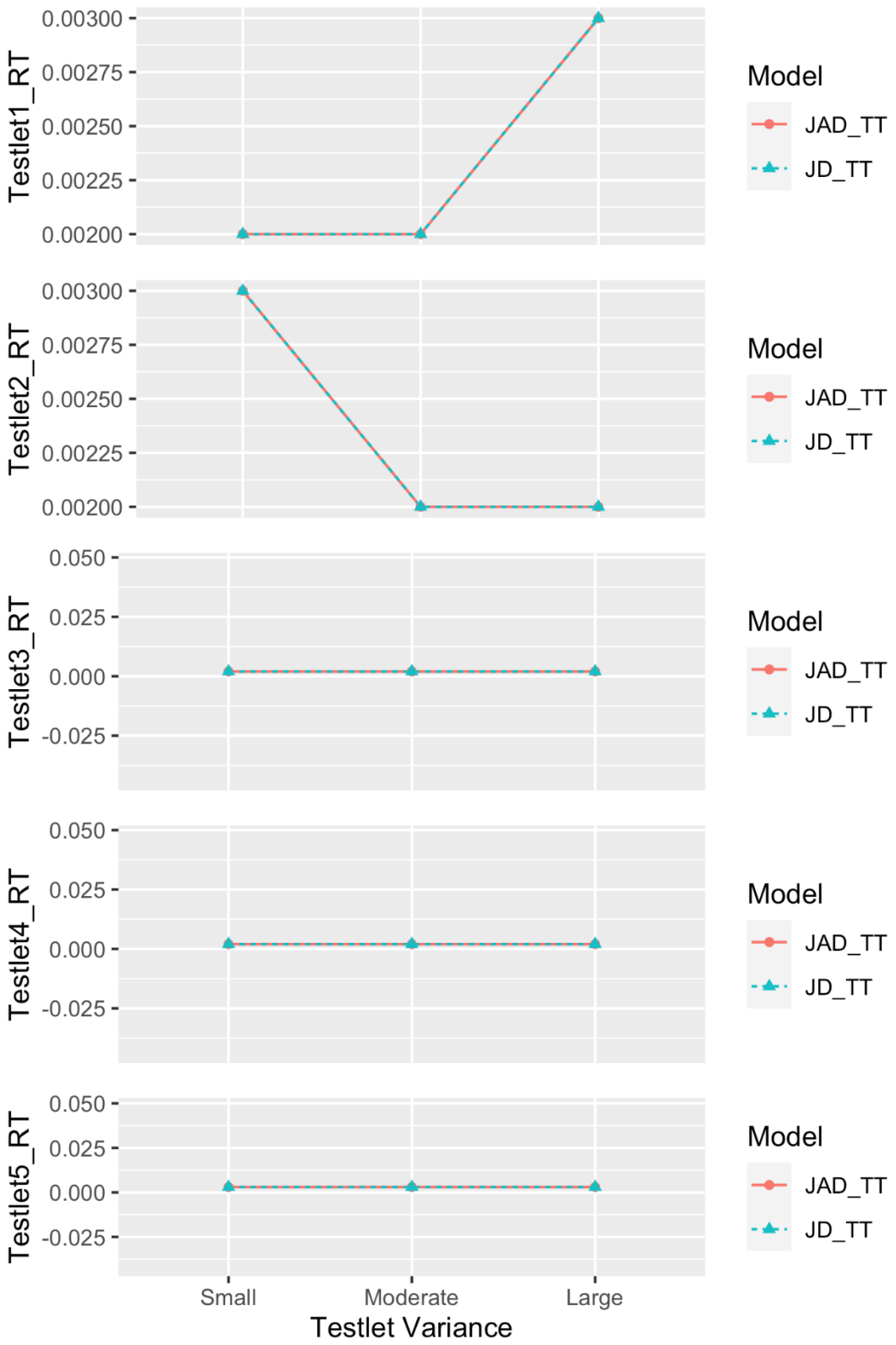

Figure 67. Marginal mean SE of the estimates of the variance of the testlet effects for response time for the five testlets at three testlet variance levels.

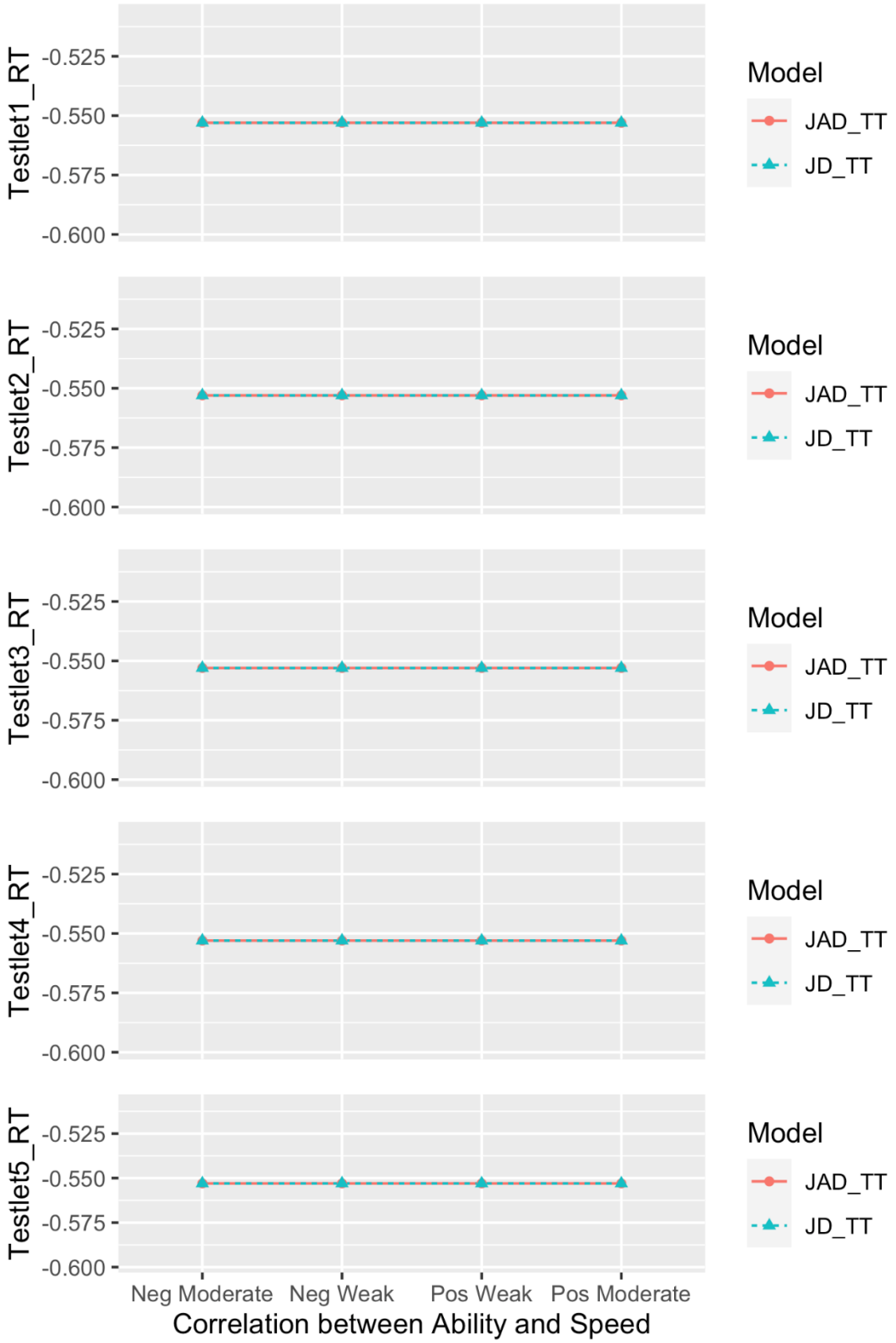

Figure 68 Marginal mean bias of the estimates of the variance of the testlet effects for response time for the five testlets at four correlation levels.

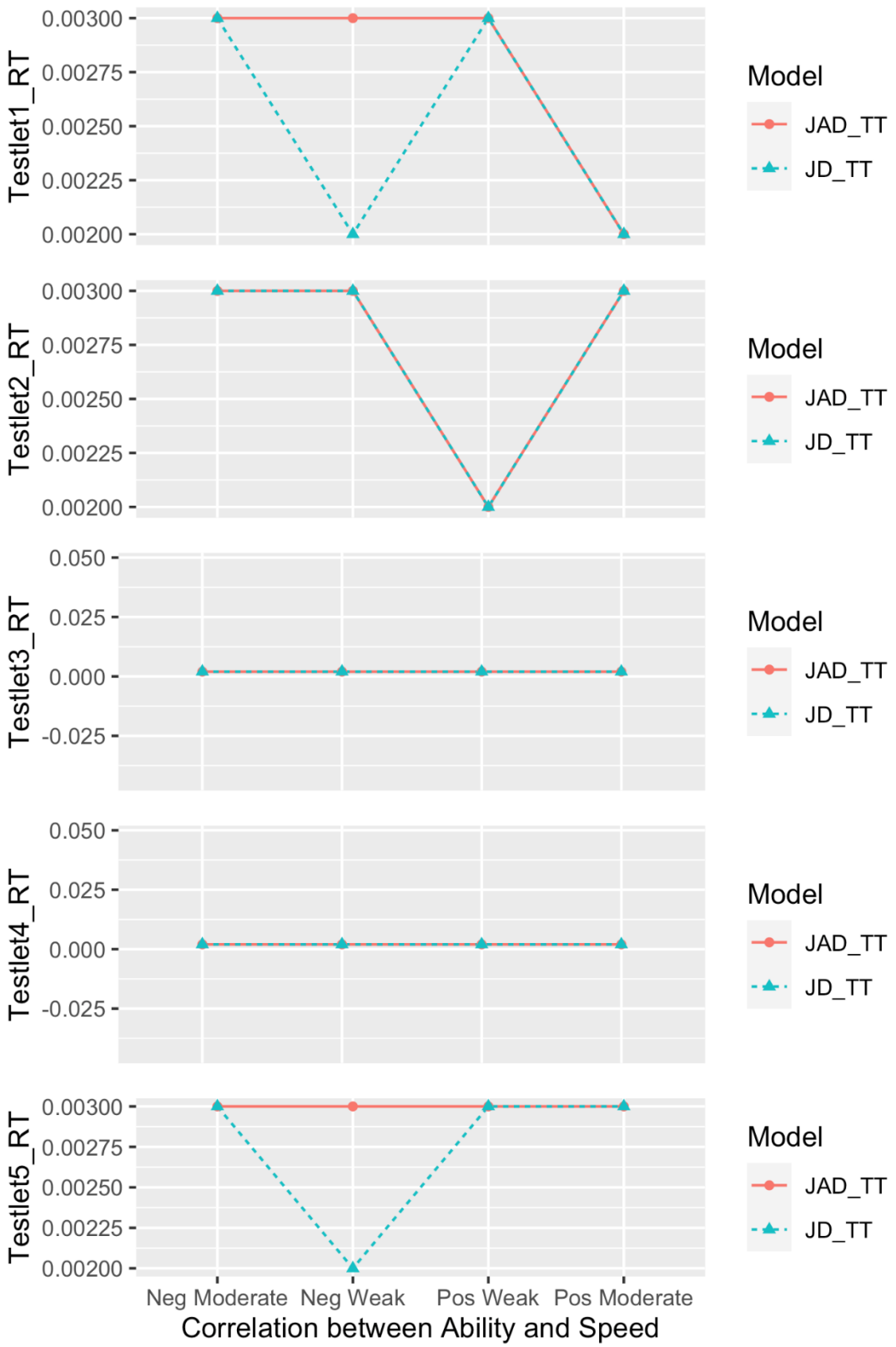

Figure 69 Marginal mean SE of the estimates of the variance of the testlet effects for response time for the five testlets at four correlation levels.

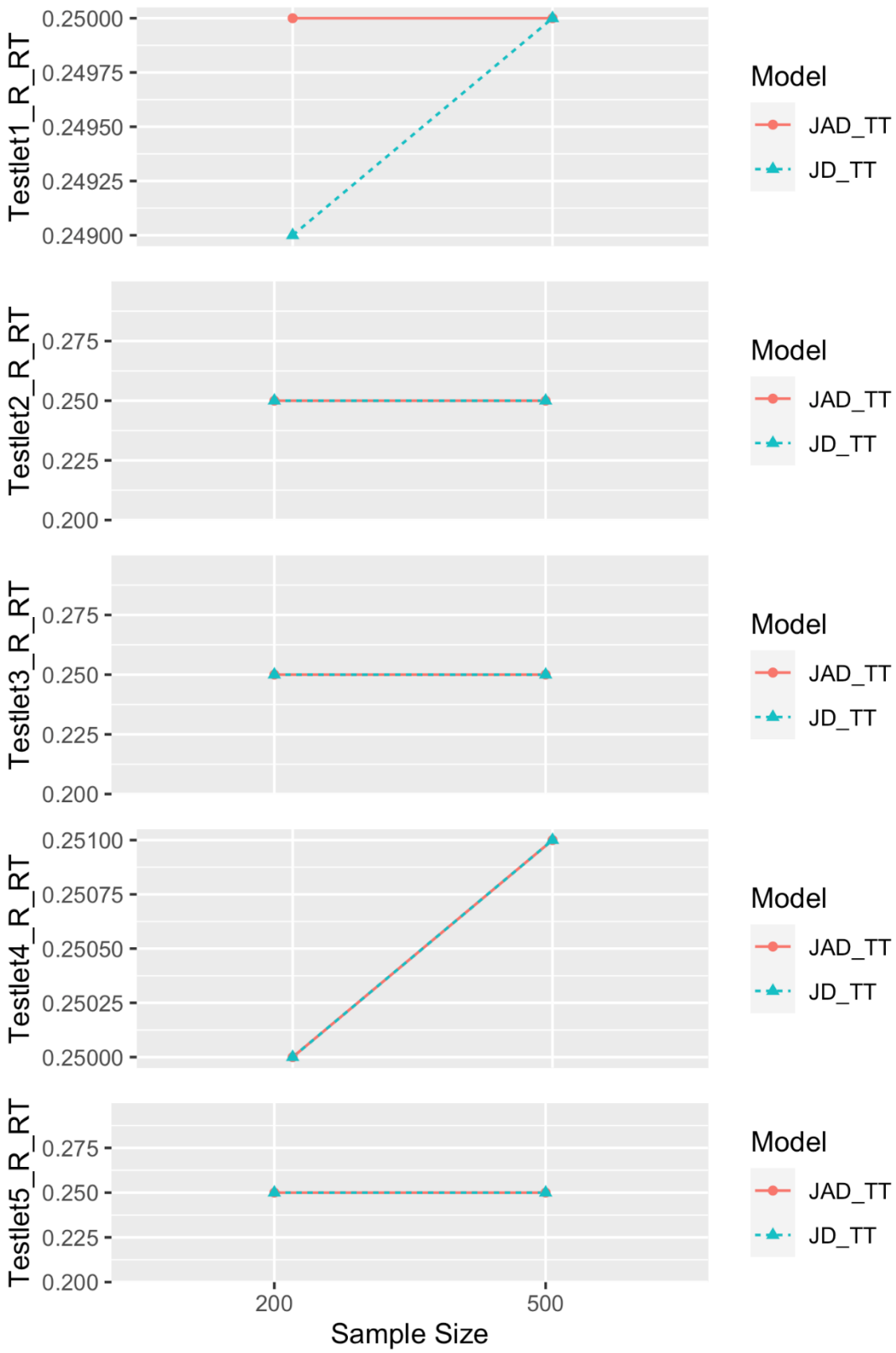

Figure 70 Marginal mean bias of the estimates of the covariance of the testlet response and response time effects for the five testlets at two sample size levels.

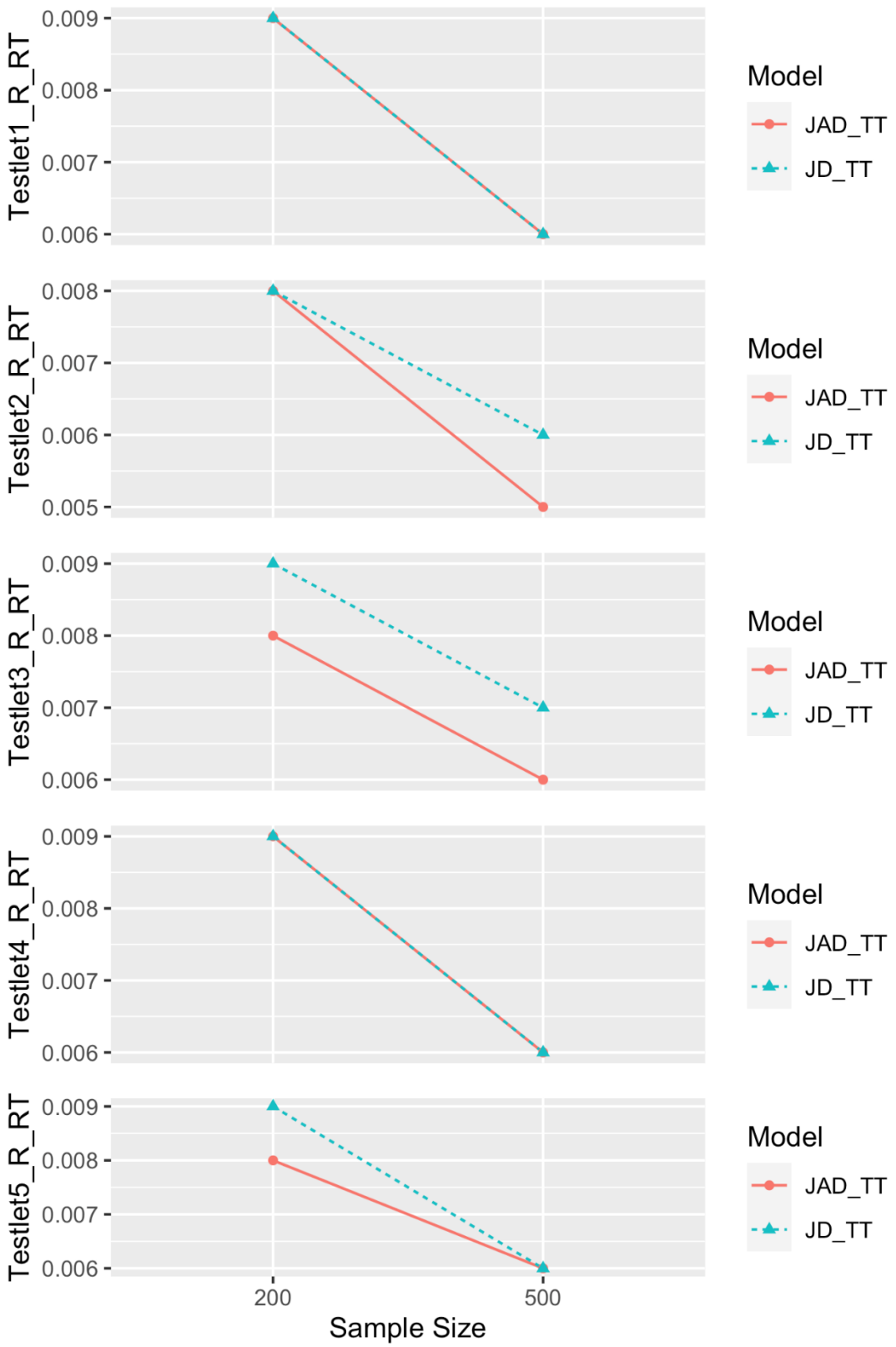

Figure 71 Marginal mean SE of the estimates of the covariance of the testlet response and response time effects for the five testlets at two sample size levels.

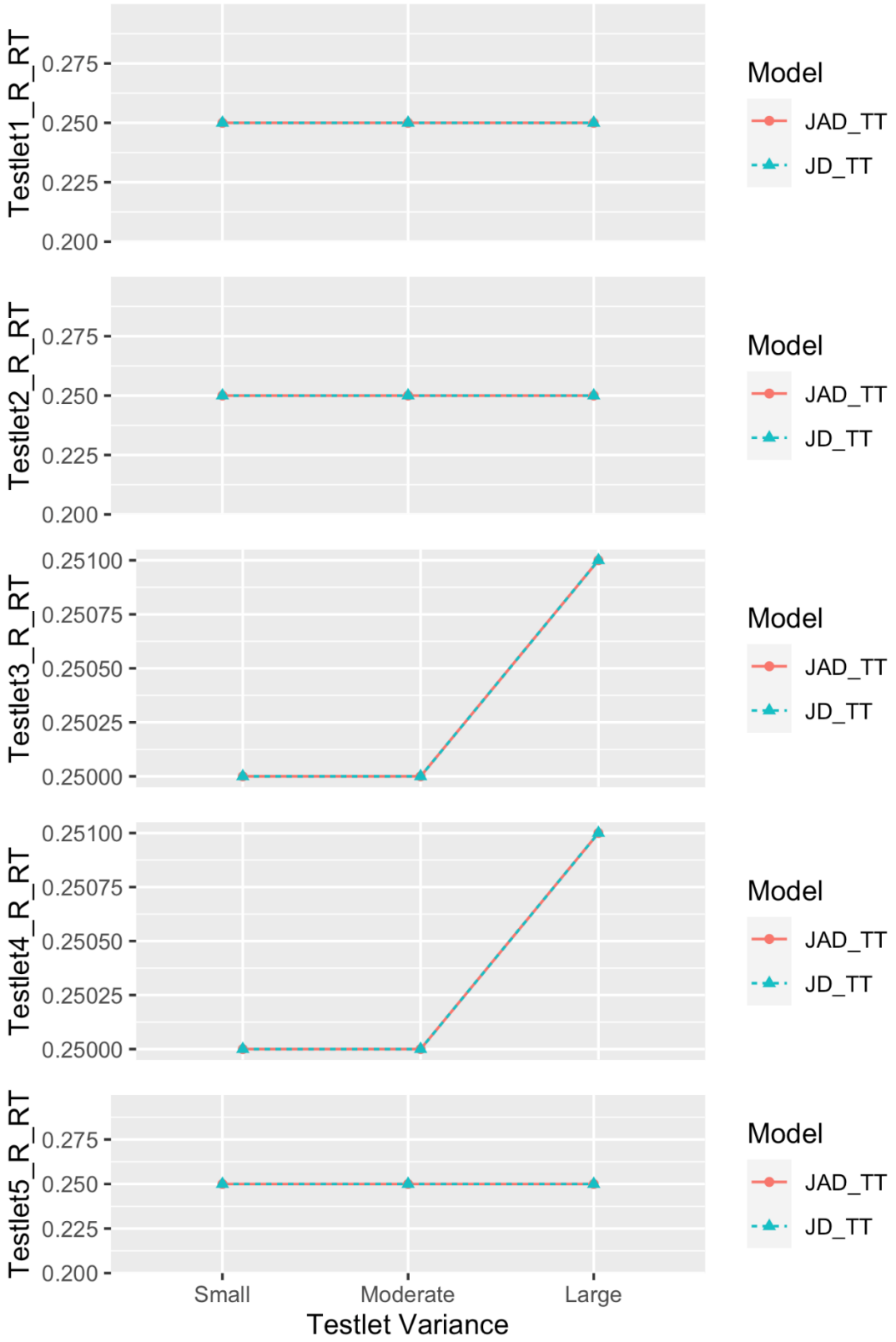

Figure 72 Marginal mean bias of the estimates of the covariance of the testlet response and response time effects for the five testlets at three testlet variance levels.

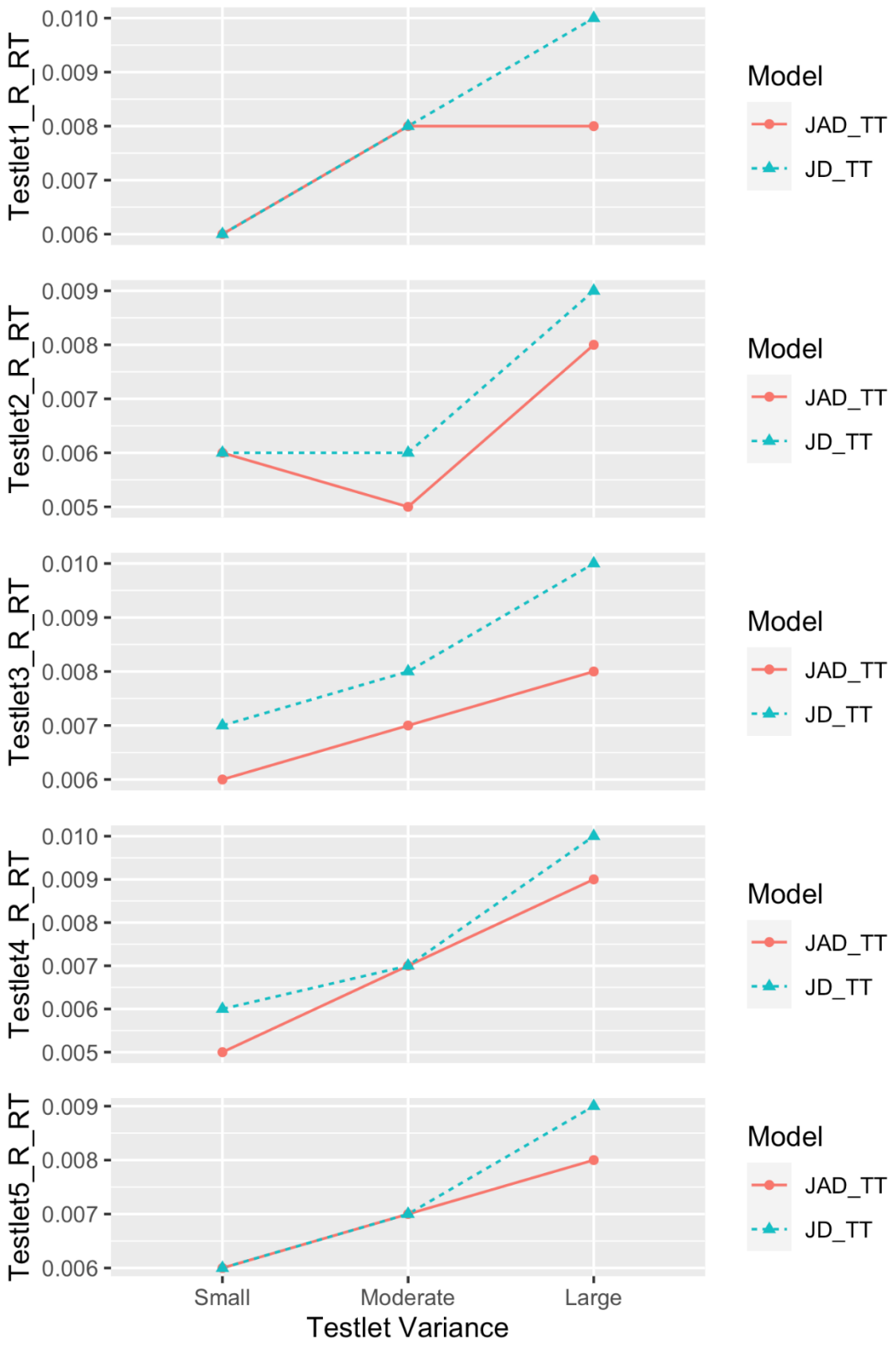

Figure 73 Marginal mean SE of the estimates of the covariance of the testlet response and response time effects for the five testlets at three testlet variance levels.

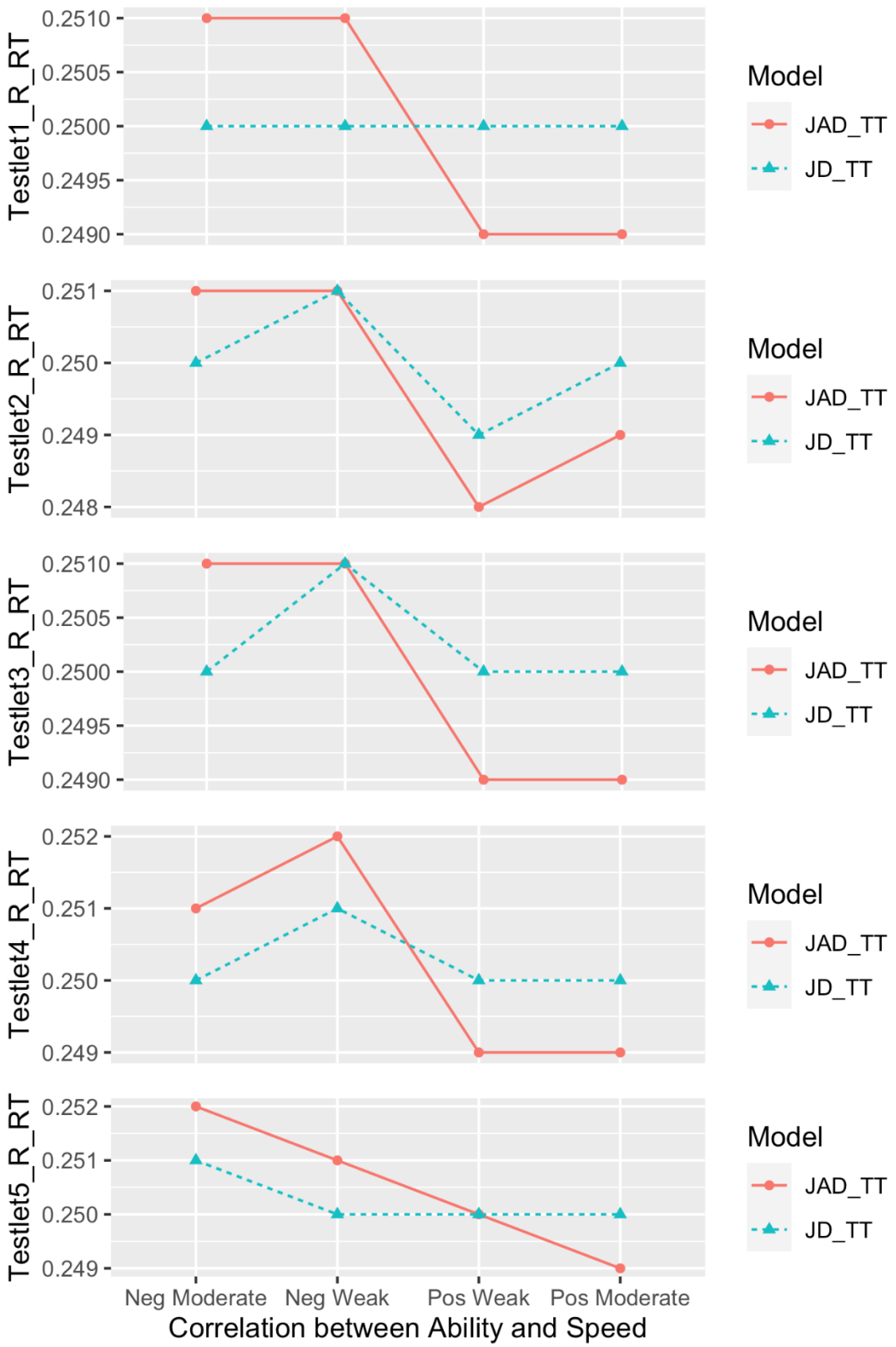

Figure 74 Marginal mean bias of the estimates of the covariance of the testlet response and response time effects for the five testlets at four correlation levels.

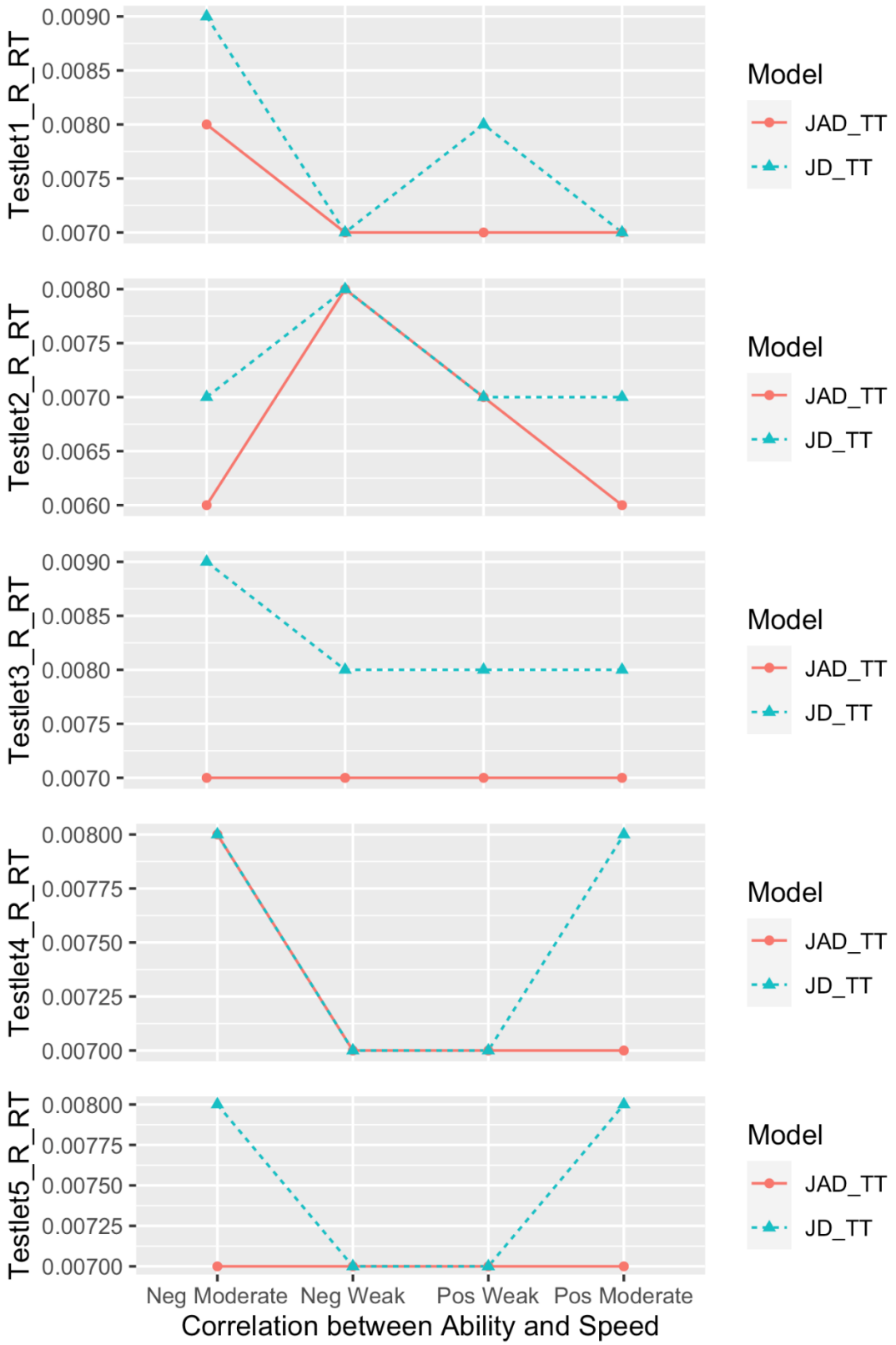

Figure 75 Marginal mean SE of the estimates of the covariance of the testlet response and response time effects for the five testlets at four correlation levels.

4.1.8 Summary of the Simulation Study Results

Through examining the performance of model fit indices, the simulation study identifies the proposed model as having a better model fit than the JRT-DINA-R/RT/AC model. ACCRs and PCCRs for the proposed model are comparable for the proposed model and the two alternative models. Mixed-effects ANOVA results suggest that the proposed mode significantly improves the marginal mean SEs of the person and item parameters and at specific levels of the manipulated factors, their marginal mean biases. Marginal mean plots indicate that the proposed model yields less biased and/or more accurate estimates of elements in the item mean vector, item and person variance and covariance matrix, and testlet variance and covariance matrix. The impact of the manipulated factor on parameter estimation is also explored in the simulation study. Section 4.2 presents results of the empirical data analysis.

4.2 Empirical Data Analysis

To demonstrate its application, the proposed model was fit to an empirical dataset consisted of 71 examinee's responses, RT, and AC patterns for 58 items from a computer-based mathematics assessment. The 55 items assess four attributes. Qmatrix identifying the items that assess each attribute is presented in Table 19. Four testlets are embedded in the assessment, each comprising two items. In addition, the JRT-DINA-R/RT/AC model ignoring the testlet effects was fit to the dateset to evaluate its fit relative to the proposed model. One limitation of the study is that the sample size of the empirical dataset is very small and is less than the number of model parameters. As such, model identification can be an issue. It was nevertheless employed in the empirical study to demonstrate the application of the proposed

model, as testing companies do not routinely extract answer changes data and efforts at acquiring a large-scale dataset consisting of responses, response time, and answer change patterns from a large testing company proved futile. Notwithstanding such limitation, the empirical data example using informative priors reported in Zhan et al, (2018a) and Zhan et al., (2018b) for Bayesian parameter estimation can demonstrate the estimation of model parameters and illustrate how they relate to each other.

This section presents the results of the empirical data analysis. $\hat{R}$ for the model parameters was ≤ 1.1. The effective sample size (ESS) ranges from 1, for instance, for the variance of θ_i which is constrained to 1, to 10,000. Section 4.2.1 presents the performance of the model fit indices. Estimation of the model parameters is presented in Section 4.2.2.

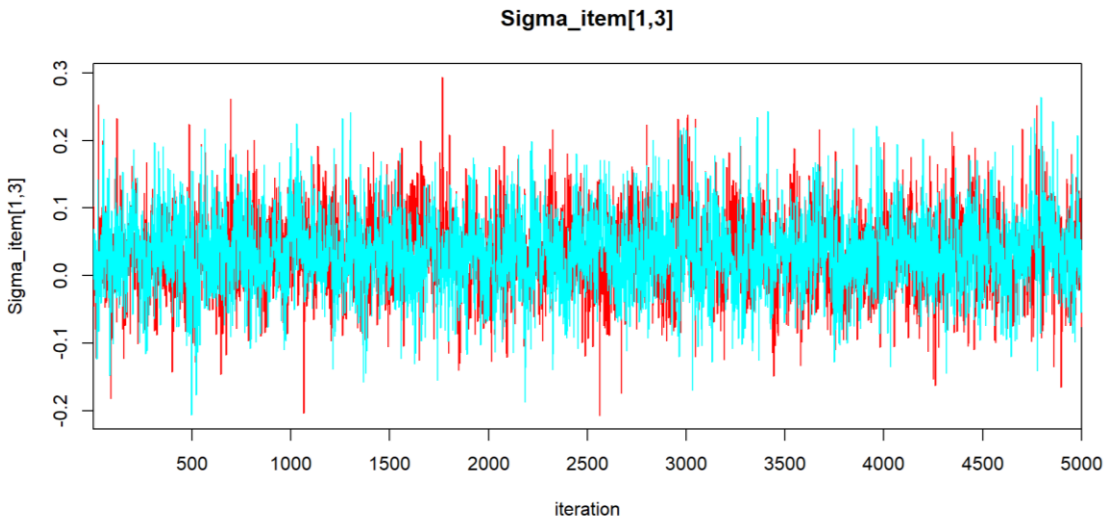

Figure 76. Sample traceplot for the covariance of item intercept and item time intensity parameter

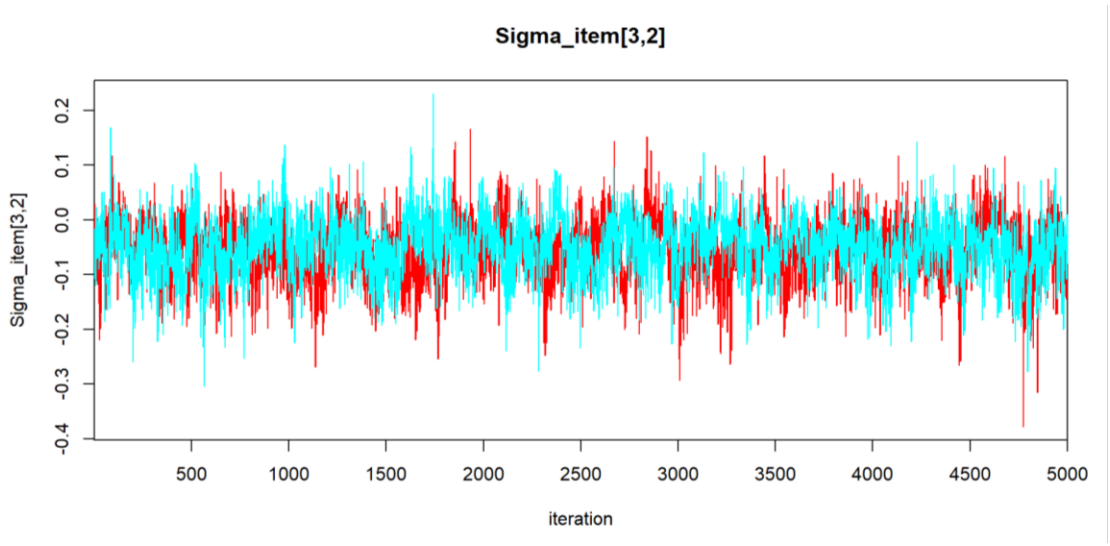

Figure 77 Sample traceplot for the covariance of item time intensity and item interaction parameter

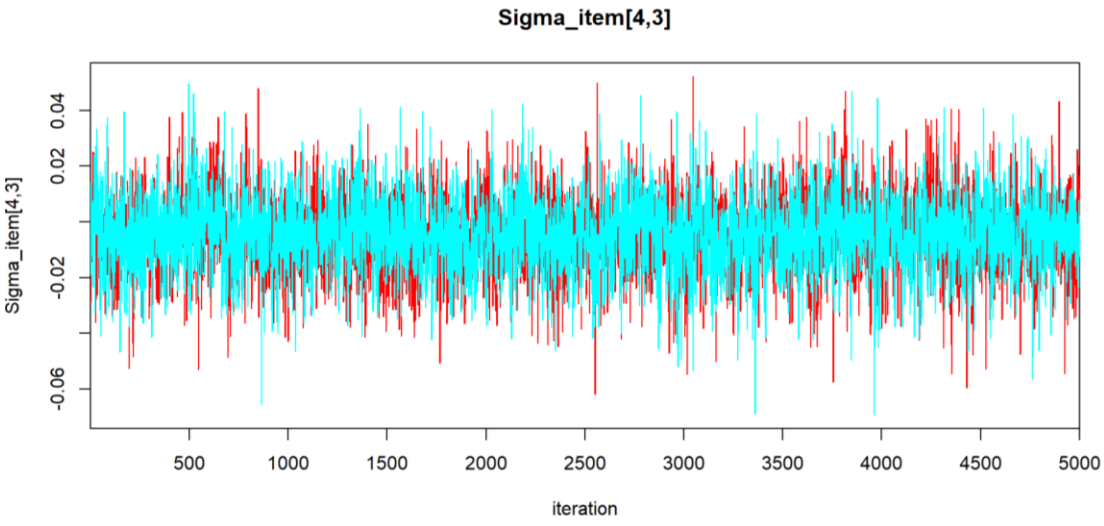

Figure 78 Sample traceplot for the covariance of item difficulty and item time intensity parameter

4.2.1 Performance of the Model Fit Indices

Table 18 presents the fit indices, -2 log-likelihood, AIC, BIC and DIC for the two models. AIC and BIC are smaller for the proposed model than for the JRT-

DINA-R/RT/AC model, indicating a better model fit. Section 4.2.2 presents the recovery of the model parameters by the proposed model.

Table 18
Model fit Indices for the Proposed Model and the JRT-DINA-R/RT/AC model

Model	-2 log-likelihood	AIC	BIC	DIC
JAD-TT	19207.377	20317.377	20234.825	20668.753
JAD	19551.364	20645.364	20564.002	20308.521

Table 19
Q-matrix for the 58 Computer-Based Mathematics Items

Items	Q-matrix			
	α_1	α_2	α_3	α_4
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0
5	0	1	0	0
6	0	1	0	0
7	0	1	0	0
8	0	1	0	0
9	0	1	0	0
10	0	1	0	0
11	0	1	0	0
12	0	1	0	0
13	0	1	0	0
14	0	1	0	0
15	1	1	0	0
16	0	1	0	0
17	0	0	1	0
18	0	1	0	0
19	0	0	1	0
20	1	0	0	0
21	0	0	0	1
22	0	1	0	0
23	0	0	1	0
24	0	1	0	0
25	0	0	0	1
26	1	0	0	0
27	0	0	0	1
28	0	1	0	0
29	0	1	0	0
30	0	1	0	0
31	0	1	0	0
32	0	0	0	1

33	0	0	0	1
34	0	0	0	1
35	0	0	0	1
36	0	1	0	0
37	0	0	0	1
38	0	1	0	0
39	0	1	0	0
40	1	0	0	0
41	0	0	0	1
42	0	0	0	1
43	1	0	0	1
44	0	0	1	0
45	0	1	0	0
46	1	0	0	0
47	0	0	0	1
48	0	1	0	0
49	0	1	0	0
50	0	1	0	0
51	0	1	0	0
52	0	1	0	0
53	1	0	0	1
54	1	0	0	0
55	0	0	1	0
56	0	1	0	0
57	1	1	0	0
58	1	1	0	0

4.2.2 Estimation of the Model Parameters

Table 20 presents the estimated person variance and covariance matrix. $\rho_{\theta\tau}$ is estimated to be -0.75, suggesting that the late ability parameter and the latent speed parameter. Zhan et al. (2018a) and Zhan et al. (2018b) similarly reported negative correlations between latent ability and latent speed, interpreting it as meaning low-performing students completing an assessment within a short timeframe and generating a great number of incorrect responses. The variance of the latent speed parameter is 1.300.

Table 20
Person Variance/Covariance Matrix Estimates for the Computer-Based Mathematics Items

Σ_{person}	θ	τ
θ	1	-0.750
τ	-0.855 (0.240)	1.300 (0.426)

Note: Covariance is in the lower triangular matrix; correlation coefficient is in the upper triangular matrix; standard errors are in the parentheses.

Table 21 presents the estimated item mean vector and item variance and covariance matrix. $\rho_{\beta\delta}$ is estimated to be -0.682, indicating an inverse relationship between the item intercept parameters and item interaction parameters. This result is consistent with Zhan et al (2018b) similarly reporting a negative $\rho_{\beta\delta}$. $\rho_{\beta\zeta}$ is estimated to be 0.089, indicating a positive weak relationship between the item intercept parameter and the item time-intensity parameter. $\rho_{\delta\zeta}$ is estimated to be -0.170, suggesting that items with higher interaction parameters tend to have lower time-intensity parameters. $\rho_{\beta b}$ and $\rho_{\zeta b}$ are estimated to be -0.755 and -0.053, which indicates an inverse relationship between the item intercept parameter and the item difficulty parameter, and between the item time-intensity parameters and the item difficulty parameter. $\rho_{\delta b}$ is estimated to be 0.3558, which means higher item interaction parameters are associated with higher item difficulty parameters.

Table 21
Item Mean Vector and Variance/Covariance Matrix Estimates for the Computer-Based Mathematics Items

	μ_{Item}	Σ_{item}	β	δ	ζ	b
μ_β	-0.127 (0.164)	β	1.253 (0.322)	-0.682	0.089	-0.755
μ_δ	2.469 (0.151)	δ	-0.705 (0.272)	0.854 (0.300)	-0.170	0.3558
μ_ζ	2.906 (0.080)	ζ	0.033 (0.056)	-0.052 (0.055)	0.110 (0.023)	-0.053
μ_b	-0.324 (0.061)	b	-0.239 (0.061)	0.093 (0.050)	-0.005 (0.014)	0.080 (0.017)

Note: Covariance is in the lower triangular matrix; correlation coefficients are in the upper triangular matrix; standard errors are in the parentheses.

Table 22 presents the four estimated testlet effect variance and covariance matrix. The variances of the four response testlet effect parameters are estimated to be 1.322, 1.000, 0.526, and 0.775, ranging from moderate to large. The variances of the RT testlet effects are estimated to be 0.250, 0.173, 0.481, and 0.381, ranging from small to moderate. The correlations between each pair of response and RT testlet effect parameters are estimated to be -0.325, -0.036, -0.278, and -0.031, suggesting an inverse relationship between each pair of response and RT testlet effect parameters.

Table 22
Testlet Effect Variance/Covariance Matrix Estimates for the Computer-Based Mathematics Items

$\Sigma_{testlet}$	Testlet 1		Testlet 2		Testlet 3		Testlet 4	
	γ	λ	γ	λ	γ	λ	γ	λ
γ	1.322 (1.017)	-0.325	1.000 (0.835)	-0.036	0.526(0.372)	-0.278	0.775 (0.840)	-0.031
λ	-0.187 (0.159)	0.250 (0.071)	-0.015 (0.115)	0.173 (0.057)	-0.140 (0.188)	0.481 (0.144)	-0.017 (0.188)	0.381 (0.105)

Note: Standard error is in the parentheses.

Table 23 presents the estimates for the higher-order structural parameters. All attribute intercept parameters are estimated to be negative, and all attribute interaction parameters are estimated to be high. As is explained by Zhan et al. (2018a), studies have indicated that accurate estimation of the second-order parameters tends to be harder to obtain than estimation of the first-order parameters. Additionally, the degree to which they are accurately estimated is dependent upon attribute estimation, which may further affect estimation accuracy.

Table 23
Higher-Order Structural Parameter Estimates for the Computer-Based Mathematics Items

Attribute	Intercept	Interaction
1	-0.947 (0.500)	6.042 (1.190)
2	-0.853 (0.489)	6.580 (1.189)
3	-0.827 (0.540)	6.089 (1.200)
4	-0.304 (0.500)	6.609 (1.211)

Note: Standard error is in the parentheses.

Table 24 presents the percentage by which each attribute is mastered by the 71 examinees. They range from 35.2% to 42.3%. Table 25 presents the estimated attribute profiles for the examinees. 11 out of the 16 possible attribute profiles are observed, with 36 examinees mastering all four attributes and 21 examinees mastering none.

Table 24
Percentage of Attribute Mastery for the Examinees

Attribute	Mastery (Count)	Mastery (Percentage)
1	25	0.352
2	27	0.380
3	26	0.366
4	30	0.423

Table 25
Estimated Attribute Profiles for the Examinees

Profile	Count
0000	21
0010	1
0100	2
0111	1
1001	1
1010	2
1011	2
1100	1
1101	1
1110	3
1111	36

Chapter 5: Discussion

Testlet-based assessments are a widely used format for assessing knowledge, skills, and abilities at various educational levels and across content domains. Advances in educational technology have made available a variety of rich process data the use of which can help improve diagnostic inferences and parameter estimation. This research extends the current joint model of responses, RT, and ACs to specifically address the scenario of testlet-based CDMs. It proposes a model that integrates ACs data in addition to modeling responses and RT and incorporates both testlet response and RT parameters to specifically address dual responses and RT dependency in CDM.

The simulation study in this research was conducted to investigate the performance of the proposed model in the context of model comparison involving two alternative models. Throughout the study, the proposed model and the two alternative models were compared in terms of model fit, ACCRs, PCCRs, and their accuracy and precision for estimating the model parameters quantified by biases and SEs. Application of this model was demonstrated with the analysis of an empirical dataset. Sections 5.1 and 5.2 are responses to the four research questions guiding this research and summarize key findings from the simulation study and the empirical data analysis. Section 5.3 discusses the limitations and directions for future research.

5.1 Findings from the Simulation Study

The simulation study was conducted to examine the extent to which accounting for dual dependency in responses and RT and including ACs as an

additional data source potentially affect model performance as evaluated by model fit, classification accuracy, and parameter recovery. It further manipulated three factors, sample size, $\rho_{\theta\tau}$, and the magnitude of the testlet variance, to investigate how they potentially affect the impact of the modeling approaches on the performance of the evaluation criteria.

5.1.1 Impact of Including of ACs as an Additional Data Source

The first two research question guiding the simulation study were: 1) how does the proposed joint model of responses, RTs, and ACs for testlet-based cognitive diagnostic assessment perform compared to testlet-based joint model of responses and RT in terms of attribute and attribute profile classification accuracy, and parameter estimates? and 2) how do the factors manipulated in this study, i.e., $\rho_{\theta\tau}$, testlet effects size, and sample size, affect comparisons of the joint model of response, RTs, and ACs for testlet-based cognitive diagnostic assessment and testlet-based joint model of responses and RT?

Attribute and Attribute Profile Classification Accuracy Rates Results of the simulation study indicates that the ACCRs and PCCRs are comparable for the proposed model, the JRT-AC-DINA for Testlets, and its comparison model excluding ACs in model specification, the Joint Testlet-DINA model, with the Joint Testlet-DINA model having slightly better marginal mean attribute and attribute profile classification rates. Further, as is shown in Figures 1 through 4, of the three manipulated factors, the only factor having a consistent impact is the magnitude of the testlet variance: as it increases from small to large, both ACCRs and PCCRs decrease across the attributes and the models.

These results suggest that the two models perform similarly as evaluated by their marginal mean classification rates at the attribute and attribute profile level. As their marginal mean ACCRs are above 93% across the attributes, and their marginal mean PCCRs are above 75%, at this level, the two models perform comparably well in recovering the attributes and attribute profiles. The recovery rates are consistent with the ACCRs and PCCRs reported in (Zhan et al., 2018b). Thus, the inclusion of ACs as an additional data source does not necessarily result in improved classification accuracy rates if the baseline comparison model is already well-performing in this respect.

Person Parameters Although the proposed modeling approach does not result in marked improvement in the classification accuracy rates at the attribute and attribute profile level, significant improvement in the mean SE of θ_i is shown for the proposed model incorporating ACs in addition to specifying testlet effects. Compared with the Joint Testlet-DINA model, at the sample size level of 200, consistent across the levels of the testlet variance, mean SEs of θ_i is markedly lower for estimation by the proposed model than by the Joint Testlet-DINA model. At the sample size of 500, consistent across the levels of testlet variance and correlation between person ability and person speed, the mean SEs of θ_i is evidently lower for the proposed model than for the Joint Testlet-DINA model. These findings suggest that incorporating ACs in the testlet-based joint model of responses and RT reduces random error in recovering θ_i at all levels of the testlet variance at both sample sizes, and at the sample size of 500, reduces random error in recovering θ_i at all levels of the correlation between speed and ability. The effect of model specification on the marginal mean biases of

the other person parameter, τ_i, is dependent on the interaction of the testlet variance and person-speed correlation: across the two sample sizes, the proposed model shows reduced systematic error in recovering this parameter at the interaction of specific levels of the two other manipulated factors. The marginal mean SEs of τ_i for the two models are identical at the sample size of 500.

Person variance/covariance matrix As is presented in Chapter 4, the marginal mean biases and SEs for estimates of the two elements in the person variance/covariance matrix: σ_τ^2 and $\sigma_{\theta\tau}$, are higher for the proposed model than the for the Joint Testlet-DINA model, suggesting that modeling ACs does not necessarily improve systematic error and random error for estimating these two elements. Further, the marginal mean bias of the estimates of $\sigma_{\theta\tau}$ and the SE of both σ_τ^2 and $\sigma_{\theta\tau}$ are lower at the sample size of 500 than at the sample size of 200, suggesting that increased sample size results in reduced random error in recovering these two elements, and reduced systematic error in recovering $\sigma_{\theta\tau}$.

Higher-Order Structural Parameters As is descried in Section 4.1.4, impact of model specification and the manipulated factors on the recovery of the attribute easiness parameters and the attribute discrimination parameters is dependent on the attribute being specified and on the intersection of the levels of the manipulated factors. Reduced systematic error attributable to the inclusion of ACs data is only shown for specific attributes and for specific levels of the manipulated factors. The marginal mean biases of the estimates for the attribute discrimination parameters are invariant across the levels of the manipulated factors, whereas the marginal mean SEs of the estimates of both attribute easiness and attribute discrimination parameters are

shown as reduced for the proposed model only for specific attributes and at specific levels of the manipulated factors. The only factor that consistently affects the recovery of these parameters is the sample size: as it increases from 200 to 500, the marginal mean SEs of the attribute easiness and discrimination parameters decrease, suggesting reduced random error associated with a larger sample.

Item Parameters As is described in Section 4.1.5, improvement in the marginal mean biases and SEs for the item parameters is shown for the proposed model that incorporate ACs data in addition to specifying responses and RT testlet effects. Compared with the Joint Testlet-DINA model, the proposed model generated smaller marginal mean SEs for estimates of the item intercept parameter across the levels of testlet variance at all four $\rho_{\theta\tau}$ levels and levels of $\rho_{\theta\tau}$ at both sample sizes. It additionally generates smaller marginal mean SEs for estimates of the item interaction parameter across the levels of the testlet variance at both sample sizes and all four levels of $\rho_{\theta\tau}$. Further, at both sample sizes and across the $\rho_{\theta\tau}$ levels, improvement in the marginal mean biases of the item interaction parameter is also shown for the proposed model in comparison with the Joint Testlet-DINA model, but only for small and moderate testlet variance where they are nonnegative. The inclusion of the ACs data, however, has little effect on the marginal mean biases and SEs for the item time intensity parameter: they are identical for the proposed model and the Joint Testlet-DINA model.

The impact of testlet variance on estimation of the item parameters is evident: as testlet variance increases from small to large, the marginal mean SEs of the item intercept parameter decrease across the levels of $\rho_{\theta\tau}$. The marginal mean SEs of the

item interaction parameter similarly decrease but only under the following conditions: at the negative moderate $\rho_{\theta\tau}$, and at the intersection of the sample size of 500 and positive $\rho_{\theta\tau}$ and of the sample size of 200 and negative weak correlation. The marginal mean SEs of the estimates of the item time intensity parameter, however, increase as testelt variance increases from small to large. The absolute value of the marginal mean biases of the item interaction parameter decreases as testlet variance increases from small to moderate but rises to the highest level as testlet variance reaches large.

Item Mean Vector and Item Variance/Covariance Matrix As is described in Section 4.1.6, overall, the inclusion of ACs as an additional data source by the proposed model improves the marginal mean SEs of the estimates for the item mean intercept and interaction parameters, but not necessarily their marginal mean biases. Only at specific levels of the manipulated factors, small testlet variance, as an example, are the marginal mean biases of the estimates for the two elements smaller as generated by the proposed model compared to the Joint Testlet-DINA model. No difference attributable to model specification is observed for in the marginal mean biases and SEs of the mean item intensity parameter. Evidently smaller marginal mean SEs in both parameters but slightly larger marginal mean biases are observed for a larger sample size of 500 than for 200.

Similar effect of model specification is observed for the marginal SEs of the elements of the item variance and covariance matrix: they are smaller for the proposed model than for the Joint Testlet-DINA model. Improvement in the marginal mean biases by the proposed model is only shown for specific elements of the item

variance and covariance matrix, and for specific levels of the manipulated factors, for instance, at the small and moderate testlet variance level. Recovery of the elements that relate to item time intensity is not affected by the inclusion of ACs. The impact of sample size on the marginal mean SEs of the estimates is similar to its impact on elements of the item mean vector: they are generally smaller for the larger sample size of 500.

Testlet Variance/Covariance Matrix Impact of including ACs as an additional data source in testlet-based joint model of responses and RT on the recovery of testlet variance/covariance matrix resembles the impact on the item variance and covariance matrix. The proposed model sees improved marginal mean SEs but higher marginal mean biases for estimates of the variance of the response testlet effect parameters. However, it has little impact on the recovery of the variance of the RT testlet effect parameter and the covariance of the response and RT testlet effect parameters as they are close identical for the two models. Impact of the manipulated factors is consistent across the five testlets: larger sample size is associated with larger marginal mean biases but smaller marginal mean SEs for estimates of the variance of the response and RT testlet effect parameters. Testlet variance has a similar impact: as it increases from small to large, larger marginal mean biases are shown for the estimates of the response and RT testlet effects parameters and larger marginal mean SEs are shown only for the estimates of the response testlet effects parameters.

5.1.2 Impact of Addressing Dual Response and RT Dependency

The third and fourth research questions guiding the simulation study were: 1) how does the proposed joint model of responses, RTs, and ACs accounting for testlet effects perform compared to the alternative model ignoring these effects in terms of model fit, attribute and attribute profile classification accuracy, and parameter estimation? and 2) how do the factors manipulated in this study, i.e., correlation between speed and ability, testlet effects size, and sample size, affect comparisons of the joint model of response, RT, and ACs accounting for testlet effects and the joint model of responses, RT, and ACs ignoring testlet effects?

Performance of Model Fit Indices Across the simulated conditions, AIC and BIC for the proposed model are consistently smaller than for the alternative model, the JRT-DINA-R/RT/AC model which does not explicitly account for local item dependency and item time dependency. The DICs, however, are larger for the proposed model than for the JRT-DINA-R/RT/AC model. As is discussed by Gelman et al (2013), results given by DIC can be nonsensical when the posterior mean is not well summarized by its mean. Levy and Mislevy (2016) similarly note that DIC may not be appropriate if the posterior mean is not a reasonable summary of the posterior, as in the case when the latent variables in latent class are discrete and nominal in nature. Judging by the AIC and BIC, the proposed model accounting for dual item and item time dependency has a better fit than the JRT-DINA-R/RT/AC model.

Statistical literature on predicative accuracy summarizes inference for θ not by a posterior distribution but by a point estimate $\hat{\theta}$, the maximum likelihood estimate

(Gelman et al., 2014). AIC and BIC both use the log posterior density of the observed data y given the point estimate $\hat{\theta}$ and correct for bias due to overfitting. DIC is a Bayesian version of AIC that replaces the maximum likelihood estimate of $\hat{\theta}$ with the posterior mean $\dot{\hat{\theta}}_{Bayes} = E(\theta|y)$ and k with the effective number of parameters. Since for continuous variables in simple models with true values not on a boundary, posterior means approach MLEs as sample size increases, but not necessarily so for hierarchical models or models with discrete variables. Spiegelhalter, Best, Carlin and van der Linde (2002) further note the distinction between conditional and marginal DICs when performing model comparisons, stating the appropriacy of the use of conditional likelihood if the parameters include the latent variables to justify inferences confined to existing clusters. If otherwise, marginal likelihood should be used. Merke, Furr, and Rabe-Hesketh (2018) along the same line recommend the use of marginal DIC if models being compared differ in the specification of the distribution for the latent variables. This dissertation study uses AIC and BIC to compare and evaluate the fit of the proposed model and the the JRT-DINA-R/RT/AC model, while at the same time noting the departure from the assumptions underlying the two fit indices and the caution that needs to be taken when using and interpreting them.

Attribute and Attribute Profile Classification Accuracy Rates As is presented in Section 4.1.2, marginal mean ACCRs and PCCRs for the proposed model and the JRT-DINA-R/RT/AC model are comparable, with the proposed model having higher classification accuracy rates. This result is similar to the impact of including ACs as an additional data source: when the baseline comparison model is well-performing in

recovering attributes and attribute profiles, explicitly modeling testlet effects may not necessarily improve model performance in this regard.

Person Parameters Significant improvement in the mean SE of τ_i is shown at both sample size levels for the proposed model explicitly accounting for testlet effects in responses and RT, compared with the JRT-DINA-R/RT/AC model. At the sample size of 200, mean SEs of θ_i is lower as estimated by the proposed model at the moderate and large testlet variance levels. At the sample size of 500, consistent across the levels of the testlet variance, and at the positive and negative moderate $\rho_{\theta\tau}$, the mean SEs of θ_i is lower as estimated by the proposed model. These findings suggest that accounting for dual item and item time dependency reduces random error in recovering θ_i at moderate and large testlet variance levels at the sample size of 200, and at the sample size of 500, reduces random error in recovering θ_i at all levels of the testlet variance and at the positive and negative moderate $\rho_{\theta\tau}$. The impact of model specification on the marginal mean biases of the other person parameter, τ_i, is dependent on the interaction of the testlet variance and person-speed correlation: across the two sample sizes, the proposed model is shown as reducing systematic error in recovering this parameter at the interaction of specific levels of the two other manipulated factors. Of the three manipulated factors, testlet variance covaries with the marginal mean SEs of the estimates of θ_i: larger marginal mean SEs are observed for larger testlet variance, suggesting that increased testlet variance results in larger random error for estimating this parameter.

Person variance/covariance matrix Of the two elements of the person variance and covariance matrix, the marginal mean biases for estimates of σ_τ^2 are

lower as estimated by the proposed model, suggesting reduced systematic error in recovering this parameter. As is stated in the discussions regarding the first two research questions, the marginal mean bias of the estimates of $\sigma_{\theta\tau}$ and the SE of both σ_{τ}^{2} and $\sigma_{\theta\tau}$ are lower at the sample size of 500 than at the sample size of 200, suggesting that increased sample size results in reduced random error in recovering these two elements, and reduced systematic error in recovering $\sigma_{\theta\tau}$. Additionally, as the correlation between person ability and speed progresses from negative moderate to positive moderate, the marginal mean SEs of the estimates of $\sigma_{\theta\tau}$ also increases, suggesting increased random error associated with positive correlations between person ability and person speed.

Higher-Order Structural Parameters Overall the marginal mean biases and SEs for estimates of the higher-order structural parameters are comparable for the proposed model and the JRT-DINA-R/RT/AC model. Reduced systematic and random error attributable to the modeling of the testlet effects is only shown for specific attributes and specific levels of the manipulated factors. As is stated above, sample size consistently affects the recovery of their marginal mean SEs: as it increases from 200 to 500, the marginal mean SEs of the attribute easiness and discrimination parameters decrease, suggesting reduced random error associated with a larger sample.

Item Parameters As is described in Section 4.1.5, improvement in the marginal mean SEs and biases for the item parameters is shown for the proposed model that explicitly account for dual item and item time dependency. Compared with the JRT-DINA-R/RT/AC model, the proposed model generated smaller marginal

mean SEs for estimates of the item intercept parameter across the levels of testlet variance at all four $\rho_{\theta\tau}$ levels and across the $\rho_{\theta\tau}$ levels at both sample sizes. It additionally generates smaller marginal mean SEs for estimates of the item interaction parameter across the levels of the testlet variance at both sample sizes and at all four levels of $\rho_{\theta\tau}$. Further, across the levels of the testlet variance and at both sample sizes and all four $\rho_{\theta\tau}$ levels, improvement in the marginal mean biases of the item interaction parameter is also shown for the proposed model in comparison with the JRT-DINA-R/RT/AC model. Explicit modeling of the testlet response and RT effects, however, has little effect on the marginal mean biases and SEs for the item time intensity parameter: they are identical for the proposed model and the Joint Testlet-DINA model. Impact of the manipulated factors is discussed above under 5.1.1.

Item Mean Vector and Item Variance/Covariance Matrix Overall, explicit modeling of the response and RT testlet effects by the proposed model improves the marginal mean biases of the estimates for the item mean intercept and interaction parameters, but not necessarily their marginal mean SEs. Compared with the JRT-DINA-R/RT/AC model, mean marginal biases of the estimates for the two elements of the item mean vector are lower for estimates generated by the proposed model than the JRT-DINA-R/RT/AC model. The mean marginal SEs of these estimates, however, are higher as estimated by the proposed model, suggesting that explicitly accounting for response and RT testlet effects reduces systematic error in recovering the elements, but increases random error. Similar impact is observed for elements of the item variance and covariance matrix that do not relate to the item time intensity

parameter. No difference attributable to model specification is observed for in the marginal mean and biases of the mean item intensity parameter. The impact of the sample size is discussed above under Section 5.1.1.

5.2 Findings from the Empirical Data Analysis

The proposed model was fitted to an empirical dataset consisting of responses, RT, and ACs of 71 examinees to the 58 items on a mathematics assessment. Comparison was drawn between the proposed model accounting for dual item and item time dependency and the JRT-DINA-R/RT/AC model ignoring testlet effects. Model fit indices suggest that the proposed model has a better model fit than the JRT-DINA-R/RT/AC model. Based upon this, recovery of the model parameters by the proposed model was presented and discussed.

Recovery of the person variance and covariance matrix suggests that the correlation between the latent ability and latent speed is estimated to be -0.75, suggesting that low ability students may be completing the assessment with a high speed, possibly due to the inability to correctly respond to the assessment items. The variance of τ is estimated to be 1.300. Recovery of the item mean vector and item variance presents in detail estimates of the inverse or positive relationship between the item intercept, item interaction, item time intensity parameter, and the item difficulty parameter. An example of this is the estimated inverse relationship between the item intercept and item difficulty parameters and the estimated positive relationship between the item interaction parameter and the item difficulty parameter.

Recovery of the testlet variance and covariance matrix indicates the magnitude of the response testlet effect and of the RT testlet effect. They range from

moderate to large for the response testlet effect and from small to moderate for the RT testlet effects. The correlation between the pair of testlet effect parameters is estimated to be negative, suggesting an inverse relationship between the pair of testlet parameters. These results are consistent with those reported in Zhan et al. (2018b), similarly reporting negative correlation between pairs of response testlet effect parameters and RT testlet effect parameters.

Recovery of the attribute mastery status provide diagnostic information regarding the percentage of students mastering each attribute. With the empirical dataset, the percentages range from 35% to 42%. Finally, recovery of the attribute profiles is indicative of where students belong in terms of mastery or nonmastery of the skills assessed by the mathematics test. Results of the empirical data analysis demonstrate the percentage of students who have mastered all four skills, and percentage of students who have not mastered a particular skill or sets of skills. This information is especially useful in allowing for finetuning classroom or web-based instruction to specifically target the skills or skill sets not yet mastered by the students having a specific attribute profile.

5.3 Limitations and Future Directions

To address the scenario of testlet-based assessment, this research proposes a joint model of responses, RT, and ACs for testlet-based cognitive diagnostic assessments. A simulation study was conducted to assess the impact of accounting for dual item and item dependency and of incorporating ACs as an additional data source on model fit, classification accuracy at the attribute and attribute profile level, and parameter estimation. Through manipulating three factors, the simulation study

examined the extent to which the manipulated factors impact the performance of the proposed model and two comparison models in recovering model parameters. Application of the proposed model was demonstrated with an empirical dataset.

Limitations and extensions of this study are addressed as follows. The first stems from the use of the DINA model as the measurement model in this research study, which assumes that all attributes are required to produce a correct response. As is discussed in Chapter 1, other examples of CDMs are DINO, LLM, RUM, and GDM (Rupp et al., 2010), each assuming a specific theory of cognitive processes for responses to the items. The modeling approach proposed in this research can readily be extended to use other CDMs as measurement models for responses. Further, the measurement model for RT used in this research is the lognormal RT model. As is reviewed Chapter 2, alternative distribution models for RT exist that can be used for modeling RT. Future studies can explore the use of alternative models for RT for joint modeling of responses, RT, and an additional data source. Another possible extension is the use of alternative models for ACs. The nature of ACs is such that it can be modeled using different approaches, such as the Poisson distribution for count data, or the generalized IRT tree model (Jeon et al., 2017). Other extensions include the modeling of person clustering effects (Jiao & Zhang, 2015) and of polytomous attributes, polytomous response items (e.g., Ma & de la Torre, 2016), and mixed-format tests.

The proposed model incorporating ACs and addressing dual item and item time dependency yields less biased and more precise estimates of the model parameters. Further, through plotting marginal means of the biases and SEs of the

parameter estimates, the simulation study identifies the specific set of parameters for which the proposed modeling approach improves their estimation and the conditions under which the improvement is shown. Results of the empirical data analysis illustrates its use for identifying the magnitude and direction of the relationship among the model parameters and providing the diagnostic information for improving instruction. As such, this proposed model will likely serve as a modeling approach for integrating multiple response process data with the modeling of responses for cognitive diagnosis, and as a useful tool for deriving informative diagnostic inferences regarding students' learning and skill acquisition.

Appendix A: Classification Accuracy, Bias, and SE Results by Simulated Conditions

Table A.1. 1

Attribute Correct Classification Rate (ACCR) and Pattern Correct Classification Rate (PCCR) Under 24 Simulated Conditions(N=200)

Condition No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ACCR					PCCR
				α_1	α_2	α_3	α_4	α_5	
1		0.25	JRT-AC-DINA for Testlet	0.9778	0.9552	0.9568	0.9722	0.9948	0.8668
			JRT-DINA-R/RT/AC	0.9765	0.9538	0.9560	0.9720	0.9943	0.8645
			Joint Testlet-DINA	0.9770	0.9557	0.9552	0.9722	0.9952	0.8660
2	-0.5	0.5	JRT-AC-DINA for Testlet	0.9647	0.9355	0.9428	0.9623	0.9888	0.8205
			JRT-DINA-R/RT/AC	0.9623	0.9347	0.9417	0.9595	0.9873	0.8132
			Joint Testlet-DINA	0.9667	0.9347	0.9475	0.9640	0.9887	0.8275
3		1	JRT-AC-DINA for Testlet	0.9387	0.9127	0.9213	0.9288	0.9665	0.7532
			JRT-DINA-R/RT/AC	0.9367	0.9132	0.9158	0.9235	0.9642	0.7402
			Joint Testlet-DINA	0.9415	0.9177	0.9250	0.9327	0.9687	0.7683
4		0.25	JRT-AC-DINA for Testlet	0.9797	0.9492	0.9635	0.9745	0.9943	0.8727
			JRT-DINA-R/RT/AC	0.9797	0.9508	0.9627	0.9735	0.9942	0.8728
			Joint Testlet-DINA	0.9803	0.9488	0.9620	0.9742	0.9955	0.8727
5	-0.3	0.5	JRT-AC-DINA for Testlet	0.9738	0.9335	0.9508	0.9610	0.9867	0.8300
			JRT-DINA-R/RT/AC	0.9725	0.9335	0.9487	0.9597	0.9857	0.8257
			Joint Testlet-DINA	0.9742	0.9340	0.9515	0.9635	0.9872	0.8347
6		1	JRT-AC-DINA for Testlet	0.9410	0.9237	0.9280	0.9425	0.9708	0.7770
			JRT-DINA-R/RT/AC	0.9375	0.9237	0.9248	0.9382	0.9670	0.7652
			Joint Testlet-DINA	0.9423	0.9242	0.9318	0.9427	0.9717	0.7842

Table A.1. 2
Attribute Correct Classification Rate (ACCR) and Pattern Correct Classification Rate (PCCR) Under 24 Simulated Conditions (N=200)

Condition No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ACCR					PCCR
				α_1	α_2	α_3	α_4	α_5	
7		0.25	JRT-AC-DINA for Testlet	0.9795	0.9543	0.9652	0.9762	0.9952	0.8798
			JRT-DINA-R/RT/AC	0.9798	0.9543	0.9633	0.9750	0.9947	0.8773
			Joint Testlet-DINA	0.9793	0.9560	0.9650	0.9752	0.9955	0.8808
8	0.3	0.5	JRT-AC-DINA for Testlet	0.9653	0.9393	0.9570	0.9635	0.9875	0.8368
			JRT-DINA-R/RT/AC	0.9633	0.9378	0.9543	0.9608	0.9867	0.8283
			Joint Testlet-DINA	0.9660	0.9398	0.9575	0.9652	0.9882	0.8423
9		1	JRT-AC-DINA for Testlet	0.9398	0.9062	0.9272	0.9405	0.9663	0.7592
			JRT-DINA-R/RT/AC	0.9372	0.9048	0.9228	0.9360	0.9622	0.7468
			Joint Testlet-DINA	0.9408	0.9075	0.9267	0.9418	0.9680	0.7688
10		0.25	JRT-AC-DINA for Testlet	0.9787	0.9555	0.9607	0.9673	0.9935	0.8672
			JRT-DINA-R/RT/AC	0.9790	0.9560	0.9613	0.9650	0.9933	0.8663
			Joint Testlet-DINA	0.9802	0.9577	0.9602	0.9683	0.9938	0.8708
11	0.5	0.5	JRT-AC-DINA for Testlet	0.9702	0.9375	0.9550	0.9613	0.9842	0.8358
			JRT-DINA-R/RT/AC	0.9697	0.9395	0.9535	0.9605	0.9823	0.8332
			Joint Testlet-DINA	0.9730	0.9427	0.9558	0.9615	0.9862	0.8452
12		1	JRT-AC-DINA for Testlet	0.9420	0.9082	0.9382	0.9248	0.9667	0.7580
			JRT-DINA-R/RT/AC	0.9403	0.9055	0.9338	0.9222	0.9645	0.7473
			Joint Testlet-DINA	0.9457	0.9133	0.9407	0.9282	0.9680	0.7725

Table A.1. 3
Attribute Correct Classification Rate (ACCR) and Pattern Correct Classification Rate (PCCR) Under 24 Simulated Conditions (N=500)

Condition No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ACCR					PCCR
				α_1	α_2	α_3	α_4	α_5	
13		0.25	JRT-AC-DINA for Testlet	0.9820	0.9509	0.9643	0.9705	0.9937	0.8713
			JRT-DINA-R/RT/AC	0.9813	0.9515	0.9635	0.9701	0.9931	0.8697
			Joint Testlet-DINA	0.9829	0.9507	0.9649	0.9721	0.9941	0.8746
14	-0.5	0.5	JRT-AC-DINA for Testlet	0.9712	0.9392	0.9528	0.9573	0.9865	0.8317
			JRT-DINA-R/RT/AC	0.9712	0.9389	0.9507	0.9549	0.9859	0.8264
			Joint Testlet-DINA	0.9731	0.9421	0.9567	0.9605	0.9869	0.8421
15		1	JRT-AC-DINA for Testlet	0.9422	0.9139	0.9315	0.9411	0.9661	0.7687
			JRT-DINA-R/RT/AC	0.9396	0.9130	0.9261	0.9381	0.9646	0.7581
			Joint Testlet-DINA	0.9460	0.9171	0.9341	0.9444	0.9685	0.7812
16		0.25	JRT-AC-DINA for Testlet	0.9789	0.9523	0.9637	0.9730	0.9927	0.8710
			JRT-DINA-R/RT/AC	0.9787	0.9528	0.9631	0.9717	0.9925	0.8694
			Joint Testlet-DINA	0.9803	0.9543	0.9626	0.9729	0.9929	0.8740
17	-0.3	0.5	JRT-AC-DINA for Testlet	0.9678	0.9414	0.9529	0.9649	0.9861	0.8373
			JRT-DINA-R/RT/AC	0.9666	0.9421	0.9512	0.9633	0.9858	0.8345
			Joint Testlet-DINA	0.9687	0.9423	0.9537	0.9665	0.9867	0.8434
18		1	JRT-AC-DINA for Testlet	0.9449	0.9109	0.9248	0.9349	0.9664	0.7595
			JRT-DINA-R/RT/AC	0.9416	0.9109	0.9217	0.9315	0.9643	0.7507
			Joint Testlet-DINA	0.9454	0.9134	0.9277	0.9371	0.9675	0.7702

Table A.1. 4

Attribute Correct Classification Rate (ACCR) and Pattern Correct Classification Rate (PCCR) Under 24 Simulated Conditions (N=500)

Condition No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ACCR					PCCR
				α_1	α_2	α_3	α_4	α_5	
19		0.25	JRT-AC-DINA for Testlet	0.9802	0.9454	0.9617	0.9682	0.9947	0.8613
			JRT-DINA-R/RT/AC	0.9791	0.9455	0.9605	0.9681	0.9947	0.8597
			Joint Testlet-DINA	0.9818	0.9473	0.9616	0.9697	0.9951	0.8671
20	0.3	0.5	JRT-AC-DINA for Testlet	0.9688	0.9415	0.9527	0.9597	0.9877	0.8336
			JRT-DINA-R/RT/AC	0.9683	0.9411	0.9498	0.9575	0.9871	0.8296
			Joint Testlet-DINA	0.9704	0.9447	0.9525	0.9617	0.9889	0.8419
21		1	JRT-AC-DINA for Testlet	0.9461	0.9190	0.9366	0.9412	0.9719	0.7743
			JRT-DINA-R/RT/AC	0.9450	0.9162	0.9325	0.9372	0.9695	0.7635
			Joint Testlet-DINA	0.9489	0.9225	0.9416	0.9445	0.9736	0.7903
22		0.25	JRT-AC-DINA for Testlet	0.9809	0.9439	0.9677	0.9677	0.9955	0.8661
			JRT-DINA-R/RT/AC	0.9807	0.9440	0.9667	0.9667	0.9950	0.8637
			Joint Testlet-DINA	0.9811	0.9453	0.9703	0.9679	0.9958	0.8709
23	0.5	0.5	JRT-AC-DINA for Testlet	0.9689	0.9384	0.9543	0.9650	0.9861	0.8373
			JRT-DINA-R/RT/AC	0.9674	0.9391	0.9527	0.9633	0.9853	0.8322
			Joint Testlet-DINA	0.9691	0.9420	0.9554	0.9665	0.9875	0.8429
24		1	JRT-AC-DINA for Testlet	0.9469	0.9058	0.9330	0.9401	0.9670	0.7647
			JRT-DINA-R/RT/AC	0.9457	0.9041	0.9301	0.9365	0.9647	0.7552
			Joint Testlet-DINA	0.9491	0.9101	0.9383	0.9428	0.9691	0.7787

Table A. 2

Correlation between Generated and Estimated Higher-Order Ability and Person Parameters in the Simulation Study

Condition No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	θ			τ		
				JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	Joint Testlet-DINA	JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	Joint Testlet-DINA
1	200	-0.5	0.25	0.8343	0.8339	0.7991	0.9805	0.9806	0.9805
2			0.5	0.8237	0.8230	0.7907	0.9808	0.9809	0.9808
3			1	0.7918	0.7908	0.7641	0.9801	0.9802	0.9801
4		-0.3	0.25	0.8270	0.8268	0.7789	0.9809	0.9810	0.9809
5			0.5	0.8217	0.8212	0.7709	0.9817	0.9818	0.9817
6			1	0.7947	0.7937	0.7387	0.9806	0.9807	0.9806
7		0.3	0.25	0.8328	0.8325	0.7853	0.9810	0.9812	0.9810
8			0.5	0.8133	0.8127	0.7640	0.9801	0.9802	0.9801
9			1	0.7927	0.7916	0.7383	0.9809	0.9810	0.9809
10		0.5	0.25	0.8267	0.8263	0.7918	0.9799	0.9801	0.9799
11			0.5	0.8207	0.8200	0.7910	0.9817	0.9818	0.9817
12			1	0.7910	0.7900	0.7548	0.9795	0.9797	0.9796
13	500	-0.5	0.25	0.8356	0.8353	0.8068	0.9808	0.9809	0.9809
14			0.5	0.8223	0.8217	0.7942	0.9805	0.9806	0.9805
15			1	0.7937	0.7926	0.7641	0.9805	0.9806	0.9805
16		-0.3	0.25	0.8302	0.8298	0.7852	0.9805	0.9805	0.9805
17			0.5	0.8155	0.8150	0.7729	0.9807	0.9807	0.9807
18			1	0.7908	0.7898	0.7385	0.9809	0.9809	0.9809
19		0.3	0.25	0.8327	0.8324	0.7898	0.9806	0.9807	0.9807
20			0.5	0.8185	0.8180	0.7724	0.9809	0.9809	0.9809
21			1	0.7975	0.7965	0.7526	0.9811	0.9812	0.9811
22		0.5	0.25	0.8365	0.8362	0.8090	0.9809	0.9809	0.9809
23			0.5	0.8220	0.8214	0.7952	0.9806	0.9807	0.9806
24			1	0.7940	0.7929	0.7638	0.9806	0.9807	0.9806

Table A. 3

Bias of the Person Variance and Covariance Matrix in the Simulation Study

Condition No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	σ_τ^2			$\sigma_{\theta\tau}$		
				JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	Joint Testlet-DINA	JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	Joint Testlet-DINA
1	200	-0.5	0.25	0.0317	0.0386	-0.0029	-0.0381	-0.0383	-0.0079
2			0.5	0.0255	0.0322	-0.0069	-0.0312	-0.0307	-0.0058
3			1	0.0179	0.0244	-0.0085	-0.0089	-0.0076	0.0198
4		-0.3	0.25	0.0166	0.0230	0.0045	-0.0067	-0.0061	0.0126
5			0.5	0.0166	0.0230	0.0035	-0.0220	-0.0214	-0.0025
6			1	0.0065	0.0134	-0.0027	0.0050	0.0057	0.0221
7		0.3	0.25	0.0154	0.0221	0.0030	0.0168	0.0166	-0.0036
8			0.5	0.0097	0.0166	-0.0038	0.0254	0.0258	0.0040
9			1	0.0115	0.0184	0.0013	0.0038	0.0033	-0.0163
10		0.5	0.25	0.0337	0.0401	-0.0012	0.0317	0.0310	-0.0105
11			0.5	0.0457	0.0517	0.0104	0.0407	0.0395	0.0090
12			1	0.0102	0.0167	-0.0132	-0.0097	-0.0097	-0.0371
13	500	-0.5	0.25	0.0313	0.0357	-0.0030	-0.0345	-0.0347	5.00E-04
14			0.5	0.0302	0.0343	-6.00E-04	-0.0224	-0.0218	0.0067
15			1	0.0262	0.0303	-0.0018	-0.0142	-0.0135	0.0132
16		-0.3	0.25	0.0084	0.0127	-0.0045	-0.0223	-0.0221	-0.0015
17			0.5	0.0073	0.0117	-0.0038	-0.0119	-0.0116	0.0071
18			1	0.0132	0.0175	0.0029	-0.0075	-0.0071	0.0103
19		0.3	0.25	0.0135	0.0178	-5.00E-04	0.0301	0.0300	0.0105
20			0.5	0.0102	0.0145	-0.0011	0.0134	0.0131	-0.0020
21			1	0.0088	0.0131	-5.00E-04	0.0027	0.0024	-0.0134
22		0.5	0.25	0.0366	0.0409	0.0014	0.0359	0.0357	0.0010
23			0.5	0.0291	0.0335	-0.0027	0.0285	0.0283	-0.0042
24			1	0.0245	0.0287	-0.0015	0.0070	0.0063	-0.0246

Table A. 4

SE of the Person Variance and Covariance Matrix in the Simulation Study

Condition No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	σ_τ^2			$\sigma_{\theta\tau}$		
				JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	Joint Testlet-DINA	JRT-AC-DINA for Testlet	JRT-DINA-R/RT/AC	Joint Testlet-DINA
1	200	-0.5	0.25	0.0367	0.0369	0.0268	0.0478	0.0482	0.0433
2			0.5	0.0251	0.0253	0.0232	0.0342	0.0341	0.0321
3			1	0.0289	0.0288	0.0225	0.0379	0.0376	0.0361
4		-0.3	0.25	0.0272	0.0271	0.0241	0.0505	0.0498	0.0480
5			0.5	0.0291	0.0293	0.0258	0.0407	0.0403	0.0404
6			1	0.0233	0.0237	0.0209	0.0384	0.0396	0.0363
7		0.3	0.25	0.0238	0.0239	0.0225	0.0416	0.0416	0.0386
8			0.5	0.0327	0.0327	0.0301	0.0446	0.0447	0.0452
9			1	0.0345	0.0342	0.0301	0.0543	0.0542	0.0591
10		0.5	0.25	0.0380	0.0381	0.0318	0.0456	0.0449	0.0418
11			0.5	0.0366	0.0362	0.0272	0.0462	0.0458	0.0463
12			1	0.0297	0.0296	0.0227	0.0503	0.0497	0.0402
13	500	-0.5	0.25	0.0193	0.0192	0.0160	0.0294	0.0294	0.0279
14			0.5	0.0219	0.0216	0.0178	0.0328	0.0324	0.0286
15			1	0.0231	0.0229	0.0174	0.0303	0.0300	0.0281
16		-0.3	0.25	0.0168	0.0168	0.0147	0.0297	0.0298	0.0304
17			0.5	0.0141	0.0142	0.0121	0.0263	0.0260	0.0251
18			1	0.0180	0.0181	0.0163	0.0340	0.0340	0.0338
19		0.3	0.25	0.0194	0.0194	0.0159	0.0320	0.0320	0.0303
20			0.5	0.0206	0.0204	0.0190	0.0277	0.0277	0.0260
21			1	0.0183	0.0183	0.0167	0.0274	0.0275	0.0255
22		0.5	0.25	0.0278	0.0276	0.0191	0.0415	0.0416	0.0347
23			0.5	0.0211	0.0212	0.0151	0.0318	0.0320	0.0309
24			1	0.0180	0.0180	0.0162	0.0270	0.0267	0.0279

Table A.5. 1
Bias of the High-Order Structural Parameters (N=200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ι					κ				
				α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
1		0.25	JAD-TT	-0.1289	-0.0037	0.0222	-0.1800	0.3810	1.2821	0.7866	0.7642	1.2339	2.6157
			JAD	-0.1379	-0.0196	0.0082	-0.1911	0.3598	1.2764	0.7448	0.7668	1.2314	2.6073
			JD-TT	-0.0131	0.0572	-0.0034	-0.0847	0.0345	0.1850	0.1430	0.0232	0.3127	0.1346
2	-0.5	0.5	JAD-TT	-0.0387	-0.0606	-0.0018	-0.0366	0.3769	1.2464	0.8340	0.7651	1.1207	2.4817
			JAD	-0.0338	-0.0756	-0.0189	-0.0442	0.3543	1.2288	0.7821	0.7660	1.1235	2.4633
			JD-TT	0.0766	-0.0575	-0.0235	0.0181	0.0202	0.0608	0.2685	0.0337	0.2283	0.0549
3		1	JAD-TT	-0.1524	-0.0378	-0.0727	-0.0457	0.2974	1.3407	0.9983	0.9418	1.0237	2.4126
			JAD	-0.1310	-0.0014	-0.0711	-0.0382	0.3096	1.3234	0.9166	0.9527	1.0366	2.4261
			JD-TT	-0.0386	-0.0072	0.0055	0.0532	-0.0033	0.2280	0.4259	0.3419	0.0973	0.1006
4		0.25	JAD-TT	-0.1662	-0.0579	-0.0053	-0.1214	0.4810	1.2453	0.8818	0.8015	1.0850	2.6039
			JAD	-0.1500	-0.0578	-0.0076	-0.1201	0.4861	1.2254	0.8261	0.7983	1.0950	2.5994
			JD-TT	-0.0141	-0.0142	0.0121	-0.008	0.0849	0.0760	0.3947	0.1085	0.1416	0.0285
5	-0.3	0.5	JAD-TT	-0.1384	-0.0528	0.0075	-0.0152	0.2784	1.3339	0.8212	0.7164	1.0257	2.4650
			JAD	-0.1059	-0.0376	0.0067	0.0051	0.2999	1.3128	0.7535	0.7130	1.0393	2.4454
			JD-TT	-0.0161	0.0028	0.0230	0.0716	-0.0187	0.2372	0.1928	-0.0079	0.0798	0.0722
6		1	JAD-TT	-0.0990	-0.0188	0.0278	-0.012	0.2505	1.2204	0.8941	1.0075	1.1785	2.3369
			JAD	-0.0741	-0.0032	0.0301	-0.0017	0.2419	1.1815	0.824	1.0096	1.1848	2.3358
			JD-TT	0.0181	0.0085	0.0431	0.0532	-0.0093	0.1179	0.246	0.2941	0.3089	0.1512

Table A.5. 2
Bias of the High-Order Structural Parameters (N=200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$		ι					κ				
			Model	α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
7		0.25	JAD-TT	-0.1839	0.0188	0.0323	-0.1122	0.4654	1.2976	0.7501	0.8468	1.1826	2.6927
			JAD	-0.1866	0.0121	0.0163	-0.1171	0.4499	1.2904	0.7052	0.8439	1.1795	2.6771
			JD-TT	-0.0428	0.0552	0.0505	-0.0068	0.0986	0.1885	0.1861	0.1519	0.1804	0.1830
8	0.3	0.5	JAD-TT	-0.1395	-0.0145	0.0074	-0.0985	0.4129	1.3352	0.9495	0.8737	1.0686	2.7881
			JAD	-0.1306	-0.0363	-0.0040	-0.0998	0.4060	1.3161	0.8984	0.8601	1.0755	2.7788
			JD-TT	-0.0171	0.0307	0.0234	-0.0349	0.0595	0.1871	0.3014	0.1743	0.1648	0.2622
9		1	JAD-TT	-0.1391	-0.0882	-0.0962	-0.1858	0.2150	1.2786	1.0047	0.9481	1.156	2.6721
			JAD	-0.1449	-0.0770	-0.1210	-0.1664	0.2107	1.2678	0.9336	0.949	1.1729	2.6407
			JD-TT	-0.0235	-0.0153	-0.0234	-0.0901	-0.0440	0.3035	0.4731	0.3374	0.2735	0.2999
10		0.25	JAD-TT	-0.0988	-0.0486	0.0078	-0.0440	0.4051	1.2633	0.8521	0.6453	1.0369	2.9256
			JAD	-0.0913	-0.0404	0.0098	-0.0390	0.4068	1.2455	0.804	0.6405	1.0335	2.9216
			JD-TT	0.0222	0.0042	3.00E-04	0.0459	0.0350	0.1505	0.2353	-0.0685	0.0989	0.2369
11	0.5	0.5	JAD-TT	-0.1892	-0.0502	0.0395	-0.1669	0.3058	1.3662	0.8042	0.8626	1.0654	2.5309
			JAD	-0.1616	-0.0486	0.0430	-0.1571	0.3173	1.3353	0.7439	0.8633	1.0638	2.5325
			JD-TT	-0.0297	-0.0319	0.0563	-0.0900	0.0103	0.1590	0.2639	0.1508	0.1549	0.1073
12		1	JAD-TT	-0.1512	0.0050	-0.0411	0.0489	0.2787	1.4266	0.9561	0.8649	1.0123	2.3068
			JAD	-0.1368	-0.0113	-0.0562	0.0482	0.2592	1.4060	0.8869	0.8731	1.0277	2.2813
			JD-TT	-0.0262	0.0883	-0.0249	0.1263	-0.0110	0.3357	0.4446	0.1847	0.1186	0.0420

Table A.5. 3
Bias of the High-Order Structural Parameters (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ι					κ				
				α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
13		0.25	JAD-TT	-0.1149	-0.0351	0.0385	-0.1240	0.5070	1.2495	0.6987	0.8504	1.1263	2.9639
			JAD	-0.1102	-0.0452	0.0313	-0.1282	0.5007	1.2451	0.6564	0.8524	1.1331	2.9560
			JD-TT	0.0335	-0.0033	0.0191	-0.0089	0.0376	0.0499	0.0627	0.1056	0.0521	0.0796
14	-0.5	0.5	JAD-TT	-0.1569	-0.0403	-0.0257	-0.0834	0.4298	1.2623	0.8512	0.8283	1.0457	2.6709
			JAD	-0.1521	-0.0466	-0.0351	-0.0790	0.4271	1.2454	0.7892	0.8231	1.0453	2.6700
			JD-TT	-0.0047	0.0018	-0.0343	0.0289	0.0192	0.0588	0.1916	0.0659	0.0161	-0.0283
15		1	JAD-TT	-0.1143	-0.0262	-0.0495	-0.0848	0.2965	1.2714	0.8922	0.8563	1.1207	2.6221
			JAD	-0.1153	-0.0534	-0.0744	-0.0926	0.2739	1.2627	0.8362	0.8625	1.1293	2.6209
			JD-TT	0.0030	-0.0099	-0.0439	0.0138	-0.0616	0.1250	0.2240	0.1127	0.1012	0.0503
16		0.25	JAD-TT	-0.245	-0.0229	0.0029	-0.1272	0.4157	1.2704	0.7929	0.7574	1.1301	2.9854
			JAD	-0.2333	-0.0197	-0.0044	-0.1222	0.4161	1.2592	0.7544	0.7485	1.1316	2.9795
			JD-TT	-0.0704	-0.0030	-0.0303	-0.0064	-0.0034	0.0679	0.1170	-0.0166	0.1201	0.1519
17	-0.3	0.5	JAD-TT	-0.1399	-0.0216	0.0565	-0.1036	0.4404	1.2746	0.8481	0.8191	1.0677	2.7101
			JAD	-0.1368	-0.0307	0.0353	-0.1062	0.4201	1.2564	0.7975	0.8150	1.0648	2.7048
			JD-TT	0.0178	0.0208	0.0435	0.0065	0.0333	0.0748	0.2083	0.0637	0.0614	0.0329
18		1	JAD-TT	-0.1778	-1.00E-04	-0.0478	-0.0549	0.3141	1.2846	0.9772	0.9369	1.1269	2.5974
			JAD	-0.1479	0.0271	-0.0444	-0.0422	0.3278	1.2683	0.9096	0.9525	1.1326	2.5904
			JD-TT	-0.0071	0.0459	-0.0344	0.0428	-0.0229	0.0652	0.3599	0.1873	0.1746	0.0399

Table A.5. 4
Bias of the High-Order Structural Parameters (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ι					κ				
				α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
19		0.25	JAD-TT	-0.1345	-0.0247	0.0291	-0.0836	0.4497	1.3443	0.6835	0.8387	1.0993	2.7527
			JAD	-0.1355	-0.0293	0.015	-0.0900	0.4393	1.3302	0.6453	0.8368	1.1013	2.7527
			JD-TT	0.0165	0.0016	-0.0029	0.0250	0.0202	0.1156	0.0273	0.0718	0.0837	0.0303
20	0.3	0.5	JAD-TT	-0.1363	-0.0792	0.0345	-0.0804	0.4322	1.2904	0.8482	0.863	0.9966	2.7393
			JAD	-0.1380	-0.0841	0.0079	-0.0881	0.4080	1.2686	0.7974	0.8563	1.0016	2.7398
			JD-TT	3.00E-04	-0.0399	0.0197	0.0169	0.0054	0.1169	0.1843	0.1116	0.0021	0.0157
21		1	JAD-TT	-0.0858	-0.0295	0.0258	-0.0586	0.4282	1.2594	0.8615	0.8401	1.1885	2.7179
			JAD	-0.0672	-0.0387	0.0131	-0.0577	0.4234	1.2413	0.8082	0.8564	1.1868	2.7153
			JD-TT	0.0709	0.0259	0.0241	0.0276	0.0303	0.0551	0.2302	0.1280	0.2356	0.0946
22		0.25	JAD-TT	-0.1871	-0.0828	0.0909	-0.1044	0.4427	1.3317	0.7881	0.9476	1.0322	2.8382
			JAD	-0.1951	-0.0912	0.0713	-0.1160	0.4222	1.3248	0.7525	0.9412	1.0301	2.8416
			JD-TT	-0.0139	-0.0540	0.0705	0.0089	0.0234	0.0512	0.1208	0.1523	0.0188	0.0587
23	0.5	0.5	JAD-TT	-0.1510	-0.0279	0.0285	-0.0950	0.3985	1.2776	0.8774	0.8280	1.0717	2.7672
			JAD	-0.1525	-0.0294	0.0161	-0.0990	0.3835	1.2599	0.8261	0.8229	1.0751	2.7605
			JD-TT	0.0058	0.0111	0.0099	0.0176	0.0125	0.0913	0.1939	0.0506	0.0763	0.0990
24		1	JAD-TT	-0.1348	-0.0349	0.0178	-0.0618	0.3871	1.3231	1.0559	0.9542	1.1282	2.5075
			JAD	-0.1347	-0.0472	0.0036	-0.0574	0.3699	1.3069	0.9606	0.9618	1.1309	2.5175
			JD-TT	0.0123	0.0180	0.0250	0.0193	0.0117	0.1135	0.4535	0.2247	0.1554	0.0159

Table A.6. 1
SE of the Higher-Order Structural Parameters (N=200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ι					κ				
				α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
1		0.25	JAD-TT	0.2855	0.3028	0.2333	0.2639	0.3644	0.3998	0.3226	0.4544	0.4994	0.5727
			JAD	0.2851	0.2987	0.2423	0.2682	0.3607	0.4009	0.3308	0.4693	0.5002	0.5767
			JD-TT	0.3166	0.3254	0.236	0.2763	0.2763	0.4840	0.2600	0.4373	0.4273	0.4503
2	-0.5	0.5	JAD-TT	0.3031	0.2798	0.2636	0.2339	0.3177	0.3679	0.4025	0.3425	0.3613	0.5579
			JAD	0.2955	0.2446	0.2627	0.2414	0.2916	0.3527	0.3830	0.3459	0.3541	0.5584
			JD-TT	0.2808	0.2756	0.2934	0.2098	0.2362	0.3364	0.4947	0.4062	0.4503	0.4733
3		1	JAD-TT	0.3220	0.3177	0.3158	0.3073	0.3568	0.4737	0.4476	0.3677	0.4250	0.4179
			JAD	0.3210	0.2890	0.3312	0.3151	0.3540	0.4710	0.4193	0.3696	0.4101	0.4233
			JD-TT	0.2967	0.3563	0.3256	0.3019	0.2700	0.4225	0.5083	0.4435	0.3810	0.3639
4		0.25	JAD-TT	0.2608	0.2503	0.2784	0.2112	0.3030	0.3493	0.4005	0.3615	0.4479	0.5028
			JAD	0.2506	0.2338	0.2776	0.2247	0.3288	0.3454	0.3977	0.3618	0.4598	0.5033
			JD-TT	0.2298	0.3118	0.2860	0.2053	0.2397	0.3587	0.5035	0.4578	0.4573	0.3530
5	-0.3	0.5	JAD-TT	0.3617	0.2572	0.2077	0.2441	0.3160	0.4034	0.4699	0.3801	0.3682	0.5256
			JAD	0.3399	0.2705	0.1881	0.2460	0.3393	0.3907	0.4398	0.3844	0.3693	0.5323
			JD-TT	0.3363	0.2699	0.2112	0.2255	0.2244	0.4367	0.5383	0.5106	0.3001	0.3978
6		1	JAD-TT	0.2887	0.3212	0.2337	0.3005	0.3486	0.4418	0.4105	0.4856	0.4390	0.5999
			JAD	0.2909	0.2865	0.2378	0.2764	0.3423	0.4173	0.3949	0.4753	0.4476	0.5913
			JD-TT	0.2824	0.3043	0.2402	0.2778	0.2284	0.4922	0.4864	0.5101	0.5157	0.4202

Table A.6. 2
SE of the Higher-Order Structural Parameters (N=200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ι					κ				
				α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
7		0.25	JAD-TT	0.2663	0.2548	0.2878	0.2252	0.3155	0.4265	0.4046	0.3167	0.4180	0.4709
			JAD	0.2570	0.2453	0.2832	0.2207	0.3090	0.4244	0.3805	0.3103	0.4161	0.4577
			JD-TT	0.2197	0.2657	0.2900	0.1985	0.2266	0.4497	0.5281	0.4008	0.4739	0.3989
8	0.3	0.5	JAD-TT	0.2306	0.2585	0.2569	0.2294	0.3642	0.4931	0.4366	0.4720	0.4070	0.5481
			JAD	0.2212	0.2480	0.2419	0.2404	0.3509	0.4759	0.4137	0.4679	0.4073	0.5436
			JD-TT	0.1985	0.2482	0.2504	0.2253	0.2303	0.5080	0.5443	0.5288	0.4588	0.4273
9		1	JAD-TT	0.3481	0.2831	0.3092	0.2948	0.3166	0.4691	0.4285	0.4304	0.4913	0.5409
			JAD	0.3779	0.2717	0.3113	0.3057	0.3348	0.4791	0.4078	0.4164	0.4925	0.5353
			JD-TT	0.3217	0.2485	0.3126	0.2706	0.2488	0.5225	0.4793	0.5333	0.5769	0.4904
10		0.25	JAD-TT	0.3150	0.2855	0.2856	0.2786	0.3440	0.4843	0.3771	0.3605	0.4217	0.6266
			JAD	0.3078	0.2478	0.2788	0.2687	0.3454	0.4693	0.3689	0.3565	0.4199	0.6362
			JD-TT	0.2952	0.2640	0.2829	0.2610	0.2394	0.4374	0.4421	0.3569	0.4433	0.4757
11	0.5	0.5	JAD-TT	0.3237	0.2793	0.2111	0.3120	0.3795	0.4389	0.3534	0.3827	0.4358	0.4256
			JAD	0.3050	0.2516	0.2113	0.3185	0.3991	0.4337	0.3511	0.3767	0.4382	0.4419
			JD-TT	0.2875	0.2786	0.2310	0.3053	0.2807	0.4035	0.5603	0.4194	0.4695	0.3276
12		1	JAD-TT	0.3328	0.2732	0.3409	0.2257	0.3443	0.4654	0.4386	0.4452	0.3241	0.4460
			JAD	0.3455	0.2566	0.3553	0.2353	0.3541	0.4556	0.4404	0.4332	0.3262	0.4611
			JD-TT	0.3084	0.2983	0.3279	0.2201	0.2635	0.5046	0.5839	0.4409	0.3747	0.4041

Table A.6. 3
SE of the Higher-Order Structural Parameters (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ι					κ				
				α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
13		0.25	JAD-TT	0.1423	0.1843	0.1762	0.1830	0.2879	0.2627	0.2365	0.2044	0.2669	0.3660
			JAD	0.1483	0.1672	0.1643	0.1791	0.2827	0.2629	0.2410	0.2031	0.2715	0.3732
			JD-TT	0.1361	0.1939	0.1811	0.1617	0.1823	0.2746	0.3120	0.2365	0.2401	0.2308
14	-0.5	0.5	JAD-TT	0.1535	0.1600	0.1723	0.1436	0.2933	0.2897	0.2957	0.3171	0.2533	0.5773
			JAD	0.1419	0.1491	0.1678	0.1445	0.3080	0.2787	0.2704	0.3133	0.2450	0.5789
			JD-TT	0.1428	0.1849	0.1668	0.1270	0.1799	0.2613	0.3449	0.3237	0.2280	0.2992
15		1	JAD-TT	0.2492	0.2362	0.1804	0.2189	0.2803	0.3091	0.2311	0.2777	0.2735	0.4073
			JAD	0.2441	0.2311	0.1661	0.2203	0.2872	0.3078	0.2211	0.2774	0.2813	0.4012
			JD-TT	0.2154	0.2229	0.1741	0.2074	0.1768	0.2906	0.2692	0.3360	0.2465	0.2737
16		0.25	JAD-TT	0.1812	0.1453	0.1424	0.1912	0.2496	0.2205	0.3254	0.1927	0.2654	0.4614
			JAD	0.1868	0.1477	0.1330	0.1849	0.2302	0.2176	0.2970	0.1942	0.2645	0.4517
			JD-TT	0.1632	0.1404	0.1456	0.1537	0.1338	0.2502	0.2978	0.2721	0.2321	0.2298
17	-0.3	0.5	JAD-TT	0.2054	0.1840	0.2296	0.2173	0.2165	0.2703	0.2374	0.2284	0.3594	0.3678
			JAD	0.1852	0.1710	0.2263	0.2227	0.2043	0.2655	0.2365	0.2291	0.3646	0.3578
			JD-TT	0.1825	0.1832	0.2391	0.1968	0.1396	0.2274	0.3015	0.2528	0.3205	0.2549
18		1	JAD-TT	0.2169	0.2226	0.2036	0.1694	0.2229	0.3118	0.2820	0.3391	0.3083	0.3905
			JAD	0.2222	0.2160	0.2079	0.1706	0.2236	0.3017	0.2625	0.3320	0.2993	0.4004
			JD-TT	0.2039	0.2105	0.1977	0.1607	0.1401	0.2774	0.3429	0.3588	0.3137	0.2214

Table A.6. 4
SE of the Higher-Order Structural Parameters (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	ι					κ				
				α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
19		0.25	JAD-TT	0.157	0.1967	0.1658	0.1576	0.2556	0.2844	0.2808	0.1905	0.2444	0.3783
			JAD	0.1618	0.2042	0.1691	0.1614	0.2417	0.2817	0.2685	0.1854	0.2453	0.3725
			JD-TT	0.1494	0.1968	0.1591	0.1416	0.1527	0.2782	0.3342	0.2380	0.2021	0.2974
20	0.3	0.5	JAD-TT	0.1924	0.1632	0.1736	0.1659	0.2476	0.2966	0.299	0.2528	0.3119	0.425
			JAD	0.1819	0.1633	0.1852	0.1820	0.2395	0.2862	0.2769	0.2496	0.3098	0.4271
			JD-TT	0.1781	0.1734	0.1707	0.1408	0.1564	0.2664	0.3428	0.3451	0.2844	0.2652
21		1	JAD-TT	0.2122	0.2012	0.1854	0.2201	0.2822	0.2981	0.2289	0.2500	0.3186	0.3990
			JAD	0.2225	0.2214	0.1913	0.2308	0.2776	0.2917	0.2106	0.2499	0.3113	0.4041
			JD-TT	0.1902	0.1864	0.1940	0.2161	0.1558	0.2660	0.2617	0.2767	0.3344	0.2269
22		0.25	JAD-TT	0.2500	0.1624	0.1432	0.1645	0.2543	0.3046	0.2534	0.2742	0.2457	0.4093
			JAD	0.2459	0.1707	0.1375	0.1602	0.2589	0.2985	0.2427	0.2715	0.2468	0.4127
			JD-TT	0.2080	0.1473	0.1453	0.1471	0.1600	0.2027	0.2939	0.2756	0.2033	0.2264
23	0.5	0.5	JAD-TT	0.2148	0.2030	0.1486	0.1775	0.2334	0.3544	0.2785	0.2193	0.2823	0.4038
			JAD	0.2005	0.2009	0.1667	0.1795	0.2227	0.3515	0.2733	0.2207	0.2777	0.4242
			JD-TT	0.1555	0.1900	0.1493	0.1531	0.1457	0.2813	0.3250	0.2717	0.2888	0.2773
24		1	JAD-TT	0.2021	0.2016	0.2022	0.2245	0.1685	0.2612	0.2939	0.3366	0.3007	0.3933
			JAD	0.2130	0.2021	0.2005	0.2306	0.1962	0.2662	0.2615	0.3359	0.3046	0.4023
			JD-TT	0.1678	0.1947	0.1731	0.2051	0.1041	0.2482	0.4369	0.3577	0.2573	0.1978

Table A. 7

Correlation between Generated and Estimated Item Parameters in the Simulation Study

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma^2_{\gamma_2}/\sigma_\lambda$	β			δ			ζ			b		
				JAD-TT	JAD	JD-TT	JADTT	JAD	JD-TT	JAD-TT	JAD	JD-TT	JAD-TT	JAD	JD-TT
1	200	-0.5	0.25	0.9218	0.9230	0.9256	0.8404	0.8419	0.8480	0.9970	0.9970	0.9970	0.7058	0.7058	
2			0.5	0.9114	0.9094	0.9158	0.8508	0.8489	0.8598	0.9968	0.9968	0.9968	0.6947	0.6951	
3			1	0.9097	0.9092	0.9134	0.8221	0.8192	0.8288	0.9971	0.9971	0.9971	0.7429	0.7427	
4		-0.3	0.25	0.9193	0.9204	0.9201	0.8427	0.8434	0.8452	0.9971	0.9971	0.9971	0.7181	0.7181	
5			0.5	0.9202	0.9187	0.9245	0.8345	0.8341	0.8385	0.9973	0.9973	0.9973	0.7311	0.7314	
6			1	0.8995	0.8978	0.9058	0.8141	0.8144	0.8186	0.9973	0.9973	0.9973	0.7079	0.7082	
7		0.3	0.25	0.9225	0.9200	0.9197	0.8554	0.8542	0.8556	0.9972	0.9972	0.9972	0.6883	0.6879	
8			0.5	0.9209	0.9196	0.9224	0.8604	0.8588	0.8608	0.9975	0.9975	0.9975	0.6991	0.6993	
9			1	0.8967	0.8933	0.9007	0.8130	0.8077	0.8191	0.9972	0.9972	0.9972	0.7168	0.7170	
10		0.5	0.25	0.9151	0.9160	0.9177	0.8542	0.8566	0.8555	0.9972	0.9972	0.9972	0.7282	0.7283	
11			0.5	0.9126	0.9129	0.9143	0.8555	0.8556	0.8601	0.9974	0.9974	0.9974	0.7139	0.7141	
12			1	0.9080	0.9055	0.9169	0.8301	0.8232	0.8430	0.9971	0.9971	0.9971	0.6978	0.6975	
13	500	-0.5	0.25	0.9665	0.9662	0.9668	0.9293	0.9284	0.9313	0.9989	0.9989	0.9989	0.7167	0.7170	
14			0.5	0.9598	0.9592	0.9619	0.9282	0.9279	0.9299	0.9986	0.9986	0.9986	0.7240	0.7240	
15			1	0.9477	0.9461	0.9528	0.9115	0.9088	0.9155	0.9989	0.9989	0.9989	0.7356	0.7357	
16		-0.3	0.25	0.9646	0.9649	0.9646	0.9190	0.9211	0.9213	0.9989	0.9989	0.9989	0.7164	0.7168	
17			0.5	0.9573	0.9580	0.9583	0.9217	0.9236	0.9237	0.9989	0.9989	0.9989	0.7154	0.7153	
18			1	0.9496	0.9470	0.9519	0.9037	0.9018	0.9030	0.9990	0.9990	0.9990	0.7302	0.7303	
19		0.3	0.25	0.9562	0.9567	0.9572	0.9267	0.9292	0.9270	0.9989	0.9989	0.9989	0.6971	0.6971	
20			0.5	0.9587	0.9583	0.9609	0.9123	0.9115	0.9158	0.9988	0.9988	0.9988	0.7128	0.7130	
21			1	0.9524	0.9487	0.9551	0.8997	0.8939	0.9042	0.9989	0.9989	0.9989	0.7270	0.7267	
22		0.5	0.25	0.9696	0.9693	0.9704	0.9352	0.9351	0.9366	0.9989	0.9989	0.9989	0.7249	0.7250	
23			0.5	0.9622	0.9624	0.9636	0.9197	0.9201	0.9219	0.9990	0.9990	0.9990	0.7389	0.7385	
24			1	0.9492	0.9457	0.9541	0.9134	0.9087	0.9164	0.9990	0.9990	0.9990	0.7098	0.7096	

Table A. 8

Bias of the Item Mean Vector

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	μ_β			μ_δ			μ_ζ		
				JAD-TT	JAD	JD-TT	JAD-TT	JAD	JD-TT	JAD-TT	JAD	JD-TT
1	200	-0.5	0.25	-0.0177	0.0999	-0.0310	0.1344	-0.114	0.1643	0.0805	0.0815	0.0771
2			0.5	-0.0090	0.1311	-0.0398	0.0135	-0.2701	0.0516	-0.0969	-0.0939	-0.0956
3			1	0.1033	0.2576	0.0738	-0.1584	-0.4660	-0.1055	0.3801	0.3802	0.3781
4		-0.3	0.25	-0.0354	0.0838	-0.0591	0.1355	-0.1128	0.1658	-0.0204	-0.0202	-0.0207
5			0.5	0.0086	0.1373	-0.0182	6.00E-04	-0.2644	0.0419	0.0973	0.0982	0.0965
6			1	0.0747	0.2178	0.0466	-0.1326	-0.4150	-0.0896	-0.077	-0.0773	-0.0785
7		0.3	0.25	-0.0285	0.0891	-0.0435	0.1128	-0.1326	0.1425	0.1278	0.1282	0.1274
8			0.5	0.0101	0.1383	-0.0127	-0.0421	-0.3038	-0.0087	0.2466	0.2503	0.2472
9			1	0.1277	0.2759	0.1094	-0.1539	-0.4638	-0.1067	0.0346	0.0335	0.0332
10		0.5	0.25	-0.0320	0.0868	-0.0436	0.0498	-0.1898	0.0737	-0.1531	-0.1535	-0.1518
11			0.5	0.0175	0.1466	-0.0059	0.0073	-0.2605	0.0485	0.2717	0.2744	0.2713
12			1	0.1388	0.2859	0.1160	-0.1413	-0.4523	-0.0974	0.2149	0.2168	0.2178
13	500	-0.5	0.25	-0.0121	0.0744	-0.0275	0.0269	-0.1570	0.0521	-0.1573	-0.1562	-0.1571
14			0.5	0.0579	0.1698	0.0464	-0.0335	-0.2668	3.00E-04	0.1975	0.1987	0.1963
15			1	0.1167	0.2380	0.0893	-0.2218	-0.4804	-0.1681	0.1707	0.1749	0.1709
16		-0.3	0.25	0.0074	0.1018	-0.0076	0.0324	-0.1652	0.0570	0.0898	0.0913	0.0894
17			0.5	0.0467	0.1566	0.0350	-0.0484	-0.2786	-0.0184	0.0620	0.0638	0.0622
18			1	0.1141	0.2443	0.0931	-0.2190	-0.4821	-0.1741	-0.2807	-0.2812	-0.283
19		0.3	0.25	-0.0205	0.0714	-0.0358	0.0480	-0.1459	0.0723	-0.1224	-0.1229	-0.1227
20			0.5	0.0139	0.1185	-0.0028	-0.0453	-0.2598	-0.0105	0.1572	0.1558	0.157
21			1	0.0794	0.2026	0.0636	-0.1361	-0.3992	-0.0934	-0.1198	-0.1185	-0.1192
22		0.5	0.25	0.0284	0.1189	0.0203	-0.0304	-0.2198	-0.0150	-0.0493	-0.0523	-0.051
23			0.5	0.1019	0.2075	0.0867	-0.0814	-0.3014	-0.0490	-0.0411	-0.0426	-0.0415
24			1	0.1151	0.2453	0.0964	-0.2087	-0.4741	-0.1602	-0.0454	-0.0458	-0.0461

Table A. 9
SE of the Item Mean Vector

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	μ_β			μ_δ			μ_ζ		
				JAD-TT	JAD	JD-TT	JADTT	JAD	JD-TT	JAD-TT	JAD	JD-TT
1	200	-0.5	0.25	0.2604	0.2467	0.2639	0.2806	0.2578	0.2739	0.4571	0.4601	0.4582
2			0.5	0.1773	0.1623	0.1753	0.1867	0.1737	0.1854	0.5314	0.5309	0.5309
3			1	0.1603	0.1492	0.1746	0.2158	0.1938	0.2198	1.0541	1.0509	1.0535
4		-0.3	0.25	0.2373	0.2296	0.2470	0.3127	0.2983	0.3158	0.4047	0.4057	0.4073
5			0.5	0.1983	0.1864	0.2109	0.2120	0.1973	0.2225	0.7444	0.7460	0.7440
6			1	0.2581	0.2427	0.2552	0.2217	0.2155	0.2252	1.0324	1.0308	1.0333
7		0.3	0.25	0.2187	0.2093	0.2240	0.1760	0.1672	0.1677	0.4965	0.4976	0.4983
8			0.5	0.2203	0.2064	0.2294	0.2456	0.2286	0.2458	0.6607	0.6637	0.6610
9			1	0.2175	0.1999	0.2237	0.2171	0.1927	0.2217	0.9810	0.9808	0.9808
10		0.5	0.25	0.2120	0.2032	0.2139	0.2082	0.204	0.2128	0.5203	0.5223	0.5210
11			0.5	0.2050	0.1893	0.2131	0.2344	0.2044	0.2346	0.6219	0.6266	0.6250
12			1	0.2092	0.1859	0.2164	0.1925	0.1612	0.1942	0.9132	0.9118	0.9114
13	500	-0.5	0.25	0.1668	0.1545	0.1712	0.2161	0.1982	0.2162	0.4234	0.4221	0.4210
14			0.5	0.1758	0.1587	0.1760	0.1608	0.1421	0.1641	0.5330	0.5319	0.5323
15			1	0.1761	0.1611	0.1841	0.2221	0.1925	0.2258	0.9897	0.9915	0.9888
16		-0.3	0.25	0.2174	0.2040	0.2216	0.2078	0.1884	0.2087	0.5440	0.5431	0.5448
17			0.5	0.2281	0.2089	0.2371	0.1942	0.1739	0.1958	0.7726	0.7710	0.7723
18			1	0.1873	0.1668	0.1933	0.1850	0.1609	0.1837	1.1501	1.1507	1.1501
19		0.3	0.25	0.1823	0.1702	0.1823	0.2344	0.2107	0.2338	0.6022	0.6016	0.6010
20			0.5	0.1923	0.1916	0.1975	0.1954	0.1886	0.1947	0.7396	0.7402	0.7407
21			1	0.1602	0.1417	0.1601	0.2035	0.1762	0.2010	1.0160	1.0165	1.0171
22		0.5	0.25	0.1660	0.1572	0.1653	0.1880	0.1751	0.1900	0.5913	0.5928	0.5921
23			0.5	0.2171	0.2012	0.2200	0.1941	0.1741	0.1968	0.6858	0.6878	0.6862
24			1	0.1890	0.1780	0.1943	0.2044	0.1784	0.2055	0.7560	0.7567	0.7581

Table A.10. 1

Bias of the Item Variance and Covariance Matrix (N=200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
1		0.25	JAD-TT	0.1717	-0.0479	0.0043	0.1683	0.0992	-0.0016	-0.0351	0.0412	-0.2048	-0.6255
			JAD	0.0749	0.0323	0.0142	0.1833	-0.0040	-0.0079	-0.0419	0.0415	-0.2048	-0.6258
			JD-TT	0.2269	-0.1150	-1.00E-04	NA	0.1869	0.0041	NA	0.0413	NA	NA
2	-0.5	0.5	JAD-TT	-0.0531	0.1064	0.0117	0.2026	0.0078	0.0142	-0.0208	0.0431	-0.2112	-0.6267
			JAD	-0.1651	0.2087	0.0277	0.2198	-0.1325	-9.00E-04	-0.034	0.0428	-0.2117	-0.6269
			JD-TT	0.0027	0.0399	0.0058	NA	0.0971	0.0214	NA	0.0428	NA	NA
3		1	JAD-TT	-0.0197	0.1770	0.0136	0.1521	-0.1598	-0.0308	-0.0285	0.0466	-0.1966	-0.6110
			JAD	-0.1651	0.2998	0.0325	0.1811	-0.3049	-0.0470	-0.0507	0.0468	-0.1966	-0.6110
			JD-TT	0.0304	0.1088	0.0102	NA	-0.0680	-0.0239	NA	0.0466	NA	NA
4		0.25	JAD-TT	-0.0064	0.1328	0.0102	0.1827	-0.0780	-0.0124	-0.0508	0.0365	-0.2062	-0.6202
			JAD	-0.0910	0.2034	0.0203	0.1965	-0.1740	-0.0185	-0.061	0.0367	-0.2055	-0.6194
			JD-TT	0.0453	0.0652	0.0057	NA	0.0036	-0.0048	NA	0.0363	NA	NA
5	-0.3	0.5	JAD-TT	0.0312	0.1105	0.0215	0.1913	-0.0615	-0.0374	-0.0676	0.0566	-0.2088	-0.6445
			JAD	-0.0777	0.1967	0.0346	0.2103	-0.1707	-0.0458	-0.0791	0.0566	-0.2088	-0.6452
			JD-TT	0.0929	0.0295	0.0178	NA	0.0483	-0.0298	NA	0.0567	NA	NA
6		1	JAD-TT	-0.1295	0.2722	0.0336	0.2125	-0.2924	-0.0390	-0.0658	0.0479	-0.2105	-0.6493
			JAD	-0.2229	0.3420	0.0467	0.2279	-0.3835	-0.0473	-0.0737	0.0480	-0.2098	-0.6482
			JD-TT	-0.0899	0.2180	0.0282	NA	-0.2178	-0.0310	NA	0.0479	NA	NA

Table A.10. 2
Bias of the Item Variance and Covariance Matrix (N=200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
7		0.25	JAD-TT	0.0681	0.0425	0.0087	0.1753	0.0044	-0.0152	-0.0320	0.0513	-0.2085	-0.6138
			JAD	-0.0319	0.1275	0.0203	0.1921	-0.0985	-0.0236	-0.0433	0.0512	-0.2089	-0.6147
			JD-TT	0.0990	-0.0067	0.0072	NA	0.0766	-0.0092	NA	0.0511	NA	NA
8	0.3	0.5	JAD-TT	0.0431	0.0241	0.0303	0.1711	0.0728	-0.0058	-0.0019	0.0336	-0.2183	-0.6265
			JAD	-0.0724	0.1304	0.0449	0.1887	-0.0642	-0.0205	-0.0157	0.0336	-0.2184	-0.6265
			JD-TT	0.0918	-0.0401	0.0282	NA	0.1657	-0.0016	NA	0.0336	NA	NA
9		1	JAD-TT	-0.0831	0.2238	0.0196	0.1928	-0.1935	-0.0469	-0.0850	0.0736	-0.1919	-0.6157
			JAD	-0.2027	0.3259	0.0348	0.2115	-0.3253	-0.0550	-0.0937	0.0737	-0.1916	-0.6156
			JD-TT	-0.0382	0.1659	0.0150	NA	-0.1172	-0.0392	NA	0.0737	NA	NA
10		0.25	JAD-TT	-0.0353	0.1520	-9.00E-04	0.2156	-0.1374	-0.0101	-0.0461	0.0645	-0.1988	-0.6538
			JAD	-0.1177	0.2185	0.0110	0.2277	-0.2241	-0.0172	-0.0516	0.0647	-0.1988	-0.6534
			JD-TT	-7.00E-04	0.1085	-0.0037	NA	-0.0799	-0.0030	NA	0.0646	NA	NA
11	0.5	0.5	JAD-TT	0.0612	0.0361	-0.0103	0.1756	0.0288	0.0044	-0.0298	0.0563	-0.1992	-0.6065
			JAD	-0.0428	0.1282	0.0045	0.1938	-0.0879	-0.0076	-0.0445	0.0567	-0.1990	-0.6063
			JD-TT	0.1139	-0.0289	-0.0145	NA	0.1214	0.0115	NA	0.0567	NA	NA
12		1	JAD-TT	-0.1340	0.2241	0.0409	0.1953	-0.1675	-0.0455	-0.0410	0.0339	-0.1972	-0.6177
			JAD	-0.2508	0.3264	0.0567	0.2177	-0.3026	-0.0575	-0.0580	0.0340	-0.1973	-0.6181
			JD-TT	-0.0799	0.1527	0.0380	NA	-0.0646	-0.0390	NA	0.0341	NA	NA

Table A.10. 3

Bias of the Item Variance and Covariance Matrix (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
13		0.25	JAD-TT	0.0905	0.0049	0.0106	0.2059	0.0362	-0.0058	-0.0465	0.0457	-0.2114	-0.6313
			JAD	0.0256	0.0592	0.0173	0.2161	-0.0297	-0.0111	-0.0544	0.0459	-0.2113	-0.6312
			JD-TT	0.1149	-0.0233	0.0090	NA	0.0749	-0.0053	NA	0.0460	NA	NA
14	-0.5	0.5	JAD-TT	0.0169	0.0837	-0.011	0.2027	-0.0619	0.0092	-0.0593	0.0372	-0.1996	-0.6219
			JAD	-0.0697	0.1589	0.0012	0.2155	-0.1559	-3.00E-04	-0.0664	0.0368	-0.2000	-0.6220
			JD-TT	0.0525	0.0454	-0.0135	NA	-0.0066	0.0135	NA	0.0369	NA	NA
15		1	JAD-TT	-0.1392	0.2534	0.0125	0.2009	-0.2666	-0.0216	-0.0334	0.0757	-0.194	-0.6231
			JAD	-0.2366	0.3340	0.0264	0.2205	-0.3681	-0.0332	-0.0496	0.0761	-0.1942	-0.6239
			JD-TT	-0.0918	0.2000	0.0070	NA	-0.1901	-0.0137	NA	0.0758	NA	NA
16		0.25	JAD-TT	0.0459	0.0575	9.00E-04	0.2128	-0.0310	-0.0098	-0.0611	0.0451	-0.2122	-0.6440
			JAD	-0.0314	0.1245	0.0104	0.2234	-0.1105	-0.0178	-0.0678	0.0454	-0.2124	-0.6449
			JD-TT	0.0802	0.0226	-0.0022	NA	0.0141	-0.0057	NA	0.0454	NA	NA
17	-0.3	0.5	JAD-TT	0.0037	0.0679	0.0144	0.1678	-0.0178	-0.0182	-0.0071	0.0560	-0.1992	-0.6102
			JAD	-0.0820	0.1433	0.0251	0.1821	-0.1131	-0.0249	-0.0180	0.0564	-0.1989	-0.6107
			JD-TT	0.0375	0.0294	0.0114	NA	0.0411	-0.0153	NA	0.0560	NA	NA
18		1	JAD-TT	-0.1505	0.2964	0.0042	0.1902	-0.3380	-0.0430	-0.0688	0.0627	-0.1948	-0.5968
			JAD	-0.2440	0.3677	0.0199	0.2089	-0.4275	-0.0526	-0.0803	0.0627	-0.1946	-0.5965
			JD-TT	-0.1067	0.2470	-0.0012	NA	-0.2725	-0.0369	NA	0.0630	NA	NA

Table A.10. 4
Bias of the Item Variance and Covariance Matrix (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
19		0.25	JAD-TT	0.0225	0.0562	0.0261	0.2074	0.0065	-0.0324	-0.0508	0.0399	-0.2098	-0.6479
			JAD	-0.0466	0.1120	0.0336	0.2195	-0.0658	-0.0359	-0.0577	0.0397	-0.2097	-0.6474
			JD-TT	0.0700	0.0042	0.0238	NA	0.0713	-0.0295	NA	0.0397	NA	NA
20	0.3	0.5	JAD-TT	0.0372	0.0794	-0.0058	0.1768	-0.1158	-0.0047	-0.0139	0.0588	-0.1995	-0.6449
			JAD	-0.0461	0.1498	0.0055	0.1903	-0.1985	-0.0140	-0.0252	0.0586	-0.1995	-0.6452
			JD-TT	0.0805	0.0372	-0.0096	NA	-0.0617	-0.0021	NA	0.0587	NA	NA
21		1	JAD-TT	-0.1077	0.2349	0.0226	0.1985	-0.2588	-0.0540	-0.0569	0.0565	-0.1941	-0.617
			JAD	-0.1906	0.3028	0.0339	0.2134	-0.3550	-0.0590	-0.0622	0.0560	-0.1943	-0.6173
			JD-TT	-0.0788	0.2006	0.0201	NA	-0.2012	-0.0528	NA	0.0562	NA	NA
22		0.25	JAD-TT	0.1440	-0.0513	-0.0389	0.1603	0.0481	0.0392	0.0075	0.0536	-0.1994	-0.6279
			JAD	0.0628	0.0202	-0.0289	0.1731	-0.0376	0.0311	-0.0024	0.0536	-0.1993	-0.6279
			JD-TT	0.1718	-0.0853	-0.0412	NA	0.0948	0.0436	NA	0.0536	NA	NA
23	0.5	0.5	JAD-TT	0.0068	0.1011	0.0133	0.1937	-0.0952	-0.0213	-0.0405	0.0517	-0.2068	-0.6308
			JAD	-0.0770	0.1698	0.0234	0.207	-0.1746	-0.0289	-0.0507	0.0517	-0.2071	-0.6306
			JD-TT	0.0522	0.0454	0.0090	NA	-0.0212	-0.0147	NA	0.0517	NA	NA
24		1	JAD-TT	-0.0901	0.1855	0.0024	0.1612	-0.1918	-0.0065	-0.0134	0.0555	-0.1937	-0.5970
			JAD	-0.1942	0.2733	0.0186	0.1855	-0.2988	-0.0204	-0.0337	0.0552	-0.1939	-0.5970
			JD-TT	-0.0554	0.1407	-0.0013	NA	-0.1241	-0.0019	NA	0.0555	NA	NA

Table A.11. 1
SE of the Item Variance and Covariance Matrix (N-200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
1		0.25	JAD-TT	0.3794	0.3203	0.1345	0.1589	0.3363	0.1329	0.1305	0.0761	0.0719	0.0986
			JAD	0.3494	0.2915	0.1270	0.1498	0.3058	0.1216	0.1189	0.0761	0.0724	0.0996
			JD-TT	0.3970	0.3355	0.1386	NA	0.3560	0.1370	NA	0.0755	NA	NA
2	-0.5	0.5	JAD-TT	0.3495	0.4004	0.0968	0.1105	0.5103	0.1072	0.1162	0.0820	0.0697	0.1030
			JAD	0.2989	0.3378	0.0891	0.1051	0.4252	0.1009	0.1070	0.0816	0.0693	0.1031
			JD-TT	0.3678	0.4287	0.1006	NA	0.5511	0.1127	NA	0.0817	NA	NA
3		1	JAD-TT	0.3316	0.3064	0.1165	0.1461	0.3328	0.1182	0.1457	0.0803	0.0626	0.0917
			JAD	0.2663	0.2395	0.1050	0.1258	0.2664	0.1032	0.1234	0.0803	0.0624	0.0907
			JD-TT	0.3669	0.329	0.1227	NA	0.3463	0.1255	NA	0.0804	NA	NA
4		0.25	JAD-TT	0.3229	0.2878	0.0975	0.1278	0.3093	0.0950	0.1028	0.0659	0.0662	0.1062
			JAD	0.3100	0.2805	0.0919	0.1224	0.2925	0.0889	0.1002	0.0666	0.0667	0.1068
			JD-TT	0.3523	0.3250	0.1002	NA	0.3484	0.0978	NA	0.0663	NA	NA
5	-0.3	0.5	JAD-TT	0.3998	0.3980	0.1145	0.1450	0.4157	0.1276	0.1625	0.0649	0.061	0.0860
			JAD	0.3447	0.3374	0.1097	0.1380	0.3494	0.1199	0.1498	0.0646	0.0611	0.0860
			JD-TT	0.4295	0.4435	0.1171	NA	0.4689	0.1330	NA	0.0649	NA	NA
6		1	JAD-TT	0.2402	0.2117	0.1100	0.0966	0.2363	0.0953	0.0782	0.0763	0.0585	0.0814
			JAD	0.2082	0.1795	0.1033	0.0909	0.2008	0.0885	0.0705	0.0767	0.0594	0.0821
			JD-TT	0.2601	0.2342	0.1133	NA	0.2502	0.1006	NA	0.0765	NA	NA

Table A.11. 2

SE of the Item Variance and Covariance Matrix (N=200)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
7		0.25	JAD-TT	0.3583	0.3186	0.1401	0.1537	0.3535	0.1230	0.1522	0.0711	0.0628	0.0828
			JAD	0.3149	0.2833	0.1325	0.1462	0.3222	0.1159	0.1440	0.0714	0.0627	0.0815
			JD-TT	0.3633	0.3381	0.1435	NA	0.3814	0.1281	NA	0.0708	NA	NA
8	0.3	0.5	JAD-TT	0.3342	0.3027	0.0946	0.1418	0.3233	0.1180	0.1432	0.0445	0.0578	0.1010
			JAD	0.2834	0.2508	0.0896	0.1353	0.2659	0.1113	0.1341	0.0445	0.0581	0.1016
			JD-TT	0.3848	0.3533	0.0962	NA	0.3686	0.1227	NA	0.0448	NA	NA
9		1	JAD-TT	0.2964	0.2780	0.0946	0.1389	0.2842	0.0921	0.1288	0.0921	0.0767	0.0902
			JAD	0.2558	0.2322	0.0912	0.1289	0.2305	0.0894	0.1190	0.0916	0.0763	0.0905
			JD-TT	0.3155	0.2952	0.0964	NA	0.3005	0.0963	NA	0.0923	NA	NA
10		0.25	JAD-TT	0.3215	0.3423	0.1105	0.1105	0.4081	0.1143	0.1104	0.0765	0.0613	0.0788
			JAD	0.2882	0.2988	0.1047	0.1048	0.3551	0.1096	0.1049	0.0766	0.0615	0.0791
			JD-TT	0.3432	0.3606	0.1111	NA	0.4296	0.1158	NA	0.0764	NA	NA
11	0.5	0.5	JAD-TT	0.4476	0.4406	0.1367	0.1573	0.5063	0.1291	0.1430	0.0725	0.0848	0.1017
			JAD	0.3915	0.3798	0.1273	0.1485	0.4359	0.119	0.1386	0.0731	0.0849	0.1011
			JD-TT	0.4733	0.4772	0.1387	NA	0.5531	0.1353	NA	0.0726	NA	NA
12		1	JAD-TT	0.2869	0.3009	0.1164	0.1505	0.3440	0.1080	0.1313	0.0569	0.0638	0.0911
			JAD	0.2515	0.2510	0.1089	0.1398	0.2758	0.1017	0.1192	0.0567	0.0639	0.0917
			JD-TT	0.2978	0.3168	0.1169	NA	0.3685	0.1111	NA	0.0572	NA	NA

Table A.11. 3
SE of the Item Variance and Covariance Matrix (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
13		0.25	JAD-TT	0.3154	0.3052	0.1129	0.1306	0.3283	0.1094	0.1280	0.0741	0.0659	0.0898
			JAD	0.2883	0.2814	0.1074	0.1259	0.3045	0.1037	0.1245	0.0738	0.0658	0.0895
			JD-TT	0.3146	0.3122	0.1102	NA	0.3403	0.1095	NA	0.0742	NA	NA
14	-0.5	0.5	JAD-TT	0.3051	0.2598	0.1267	0.1429	0.2964	0.1086	0.1221	0.0737	0.0609	0.0921
			JAD	0.2679	0.2227	0.1182	0.1342	0.2602	0.0999	0.1110	0.0736	0.0607	0.0924
			JD-TT	0.3151	0.2625	0.1275	NA	0.2930	0.1086	NA	0.0735	NA	NA
15		1	JAD-TT	0.2398	0.1613	0.0909	0.112	0.1694	0.0858	0.0816	0.0845	0.0475	0.0751
			JAD	0.2184	0.1468	0.0868	0.1069	0.1447	0.0781	0.0785	0.0846	0.0473	0.0743
			JD-TT	0.2588	0.1791	0.0942	NA	0.1884	0.0904	NA	0.0846	NA	NA
16		0.25	JAD-TT	0.2951	0.2785	0.0959	0.0926	0.3011	0.0989	0.1009	0.0747	0.0598	0.0879
			JAD	0.2715	0.2568	0.0938	0.0897	0.2800	0.0970	0.0964	0.0753	0.0599	0.0876
			JD-TT	0.3050	0.2891	0.1004	NA	0.3137	0.1048	NA	0.0753	NA	NA
17	-0.3	0.5	JAD-TT	0.3381	0.2973	0.1100	0.1599	0.3229	0.1270	0.1511	0.0779	0.0670	0.0889
			JAD	0.3044	0.2756	0.1026	0.1512	0.3078	0.1179	0.1415	0.0784	0.0672	0.0887
			JD-TT	0.3515	0.3162	0.1138	NA	0.3408	0.1306	NA	0.0781	NA	NA
18		1	JAD-TT	0.2788	0.2015	0.1107	0.1395	0.2124	0.0849	0.1004	0.0686	0.0850	0.0899
			JAD	0.2412	0.1737	0.1051	0.1300	0.1741	0.0785	0.0913	0.0691	0.0851	0.0901
			JD-TT	0.2991	0.2277	0.1174	NA	0.2448	0.0896	NA	0.0689	NA	NA

Table A.11. 4
SE of the Item Variance and Covariance Matrix (N=500)

Cond. No.	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	Model	Item Variance/Covariance Matrix									
				σ_β^2	$\sigma_{\beta\delta}$	$\sigma_{\beta\zeta}$	$\sigma_{\beta b}$	σ_δ^2	$\sigma_{\delta\zeta}$	$\sigma_{\delta b}$	σ_ζ^2	$\sigma_{\zeta b}$	σ_b^2
19		0.25	JAD-TT	0.2884	0.2856	0.1066	0.1112	0.3148	0.1122	0.1190	0.0577	0.0574	0.0943
			JAD	0.2619	0.2574	0.1023	0.1027	0.2826	0.1070	0.1103	0.0576	0.0577	0.0943
			JD-TT	0.3131	0.3123	0.1082	NA	0.3463	0.1147	NA	0.0572	NA	NA
20	0.3	0.5	JAD-TT	0.3734	0.3294	0.1097	0.1458	0.3332	0.1168	0.1371	0.0727	0.0550	0.0768
			JAD	0.3436	0.3014	0.1057	0.1416	0.2984	0.1117	0.1339	0.0724	0.0551	0.0766
			JD-TT	0.3814	0.3438	0.1112	NA	0.3516	0.1188	NA	0.0726	NA	NA
21		1	JAD-TT	0.2504	0.2114	0.1026	0.1289	0.2672	0.096	0.1022	0.0797	0.0626	0.0841
			JAD	0.2249	0.1753	0.0985	0.1243	0.2111	0.0906	0.0951	0.0795	0.0625	0.0839
			JD-TT	0.2573	0.2256	0.1036	NA	0.2943	0.0975	NA	0.0795	NA	NA
22		0.25	JAD-TT	0.3255	0.3225	0.1344	0.1340	0.3909	0.1280	0.1186	0.0648	0.0684	0.0985
			JAD	0.3043	0.3024	0.1295	0.1288	0.3611	0.1240	0.1129	0.0649	0.0681	0.0981
			JD-TT	0.3424	0.3398	0.1377	NA	0.4094	0.1339	NA	0.0649	NA	NA
23	0.5	0.5	JAD-TT	0.2503	0.2049	0.0985	0.0917	0.2676	0.1039	0.0901	0.0626	0.0518	0.0816
			JAD	0.2217	0.1834	0.0947	0.0865	0.2429	0.0995	0.0847	0.0626	0.0517	0.0818
			JD-TT	0.2659	0.2211	0.1007	NA	0.2953	0.1078	NA	0.0627	NA	NA
24		1	JAD-TT	0.2891	0.2846	0.1129	0.1273	0.2847	0.1105	0.1219	0.0856	0.0667	0.0841
			JAD	0.2533	0.2481	0.1089	0.1183	0.2420	0.1089	0.1127	0.0854	0.0664	0.0846
			JD-TT	0.3009	0.3013	0.1143	NA	0.3092	0.1129	NA	0.0855	NA	NA

Table A.12. 1

Bias of the Testlet Variance and Covariance Matrix

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	$\sigma_{\gamma_1}^2$		$\sigma_{\lambda_1}^2$		$\sigma_{\gamma_1\lambda_1}$		$\sigma_{\gamma_2}^2$		$\sigma_{\lambda_2}^2$	
				JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT
1	200	-0.5	0.25	0.1225	0.1458	-0.2143	-0.2143	0.2506	0.2499	0.1360	0.1519	-0.2129	-0.2129
2			0.5	0.0199	0.0787	-0.4638	-0.4637	0.2502	0.2487	-0.1056	-0.0666	-0.4648	-0.4647
3			1	-0.4973	-0.4138	-0.9635	-0.9636	0.2489	0.2471	-0.5207	-0.4693	-0.9635	-0.9636
4		-0.3	0.25	0.1459	0.1636	-0.2138	-0.2138	0.2501	0.2497	0.1178	0.1343	-0.2143	-0.2143
5			0.5	-0.0727	-0.0470	-0.4638	-0.4639	0.2520	0.2512	-0.0999	-0.0658	-0.4629	-0.4629
6			1	-0.5466	-0.4882	-0.9630	-0.9630	0.2502	0.2492	-0.5265	-0.4562	-0.9640	-0.9640
7		0.3	0.25	0.1257	0.1639	-0.2125	-0.2126	0.2485	0.2489	0.1015	0.1206	-0.2136	-0.2136
8			0.5	-0.0489	-0.0004	-0.4649	-0.4649	0.2501	0.2506	-0.1068	-0.0877	-0.4650	-0.4651
9			1	-0.4541	-0.3783	-0.9630	-0.9631	0.2508	0.2508	-0.5857	-0.5291	-0.9630	-0.9631
10		0.5	0.25	0.1355	0.1620	-0.2134	-0.2134	0.2472	0.2474	0.1097	0.1394	-0.2137	-0.2138
11			0.5	-0.0659	-0.0231	-0.4636	-0.4635	0.2479	0.249	-0.0994	-0.0752	-0.4638	-0.4638
12			1	-0.5049	-0.4348	-0.9641	-0.9641	0.2498	0.2511	-0.5710	-0.5322	-0.9643	-0.9642
13	500	-0.5	0.25	0.0069	0.0231	-0.2263	-0.2264	0.2514	0.2512	0.0207	0.0316	-0.2255	-0.2255
14			0.5	-0.1321	-0.1101	-0.4755	-0.4755	0.2507	0.2493	-0.1534	-0.1116	-0.4757	-0.4757
15			1	-0.5784	-0.4995	-0.9763	-0.9763	0.2523	0.2521	-0.6448	-0.5826	-0.9761	-0.9761
16		-0.3	0.25	0.0561	0.0812	-0.2261	-0.2262	0.2509	0.2507	0.0415	0.0633	-0.2252	-0.2252
17			0.5	-0.1606	-0.1327	-0.4757	-0.4757	0.2500	0.2496	-0.1532	-0.1310	-0.4760	-0.4760
18			1	-0.5659	-0.4735	-0.9759	-0.9759	0.2499	0.2491	-0.6302	-0.5754	-0.9760	-0.9760
19		0.3	0.25	0.0425	0.0627	-0.2256	-0.2256	0.2479	0.2481	0.0214	0.0373	-0.2253	-0.2253
20			0.5	-0.1501	-0.1204	-0.4757	-0.4756	0.2482	0.2484	-0.2075	-0.1803	-0.4767	-0.4767
21			1	-0.5498	-0.4924	-0.9760	-0.9760	0.2504	0.2515	-0.6269	-0.5631	-0.9763	-0.9764
22		0.5	0.25	0.0361	0.0517	-0.2253	-0.2253	0.2503	0.2509	0.0184	0.0353	-0.2261	-0.2261
23			0.5	-0.1542	-0.1081	-0.4756	-0.4755	0.2499	0.2506	-0.2100	-0.1824	-0.4756	-0.4756
24			1	-0.5739	-0.4928	-0.9761	-0.9761	0.2494	0.2507	-0.6445	-0.5893	-0.9758	-0.9758

Table A.12. 2

Bias of the Testlet Variance and Covariance Matrix

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma^2_{\gamma_2}/\sigma^2_{\lambda}$	$\sigma_{\gamma_2\lambda_2}$		$\sigma^2_{\gamma_3}$		$\sigma^2_{\lambda_3}$		$\sigma_{\gamma_3\lambda_3}$		$\sigma^2_{\gamma_4}$	
				JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT
1	200	-0.5	0.25	0.2515	0.2510	0.2150	0.2766	-0.2137	-0.2137	0.2515	0.2504	0.1091	0.1196
2			0.5	0.2495	0.2484	-0.0475	0.0373	-0.4647	-0.4647	0.2494	0.2488	-0.0503	-0.0276
3			1	0.2523	0.2513	-0.4018	-0.2371	-0.9636	-0.9636	0.2525	0.2523	-0.4214	-0.3557
4		-0.3	0.25	0.2505	0.2504	0.1884	0.2448	-0.2139	-0.2139	0.2504	0.2498	0.0852	0.0891
5			0.5	0.2526	0.2526	-0.0343	0.0460	-0.4637	-0.4636	0.2538	0.2537	-0.1051	-0.079
6			1	0.2509	0.2505	-0.4978	-0.3785	-0.9641	-0.9641	0.2506	0.2502	-0.4743	-0.4337
7		0.3	0.25	0.2476	0.2475	0.1766	0.2154	-0.2146	-0.2146	0.2482	0.2488	0.0871	0.0953
8			0.5	0.2488	0.2489	-0.0549	0.0260	-0.4638	-0.4639	0.2501	0.2508	-0.1056	-0.0831
9			1	0.2460	0.2467	-0.4346	-0.2896	-0.9642	-0.9641	0.2506	0.2516	-0.4156	-0.3504
10		0.5	0.25	0.2505	0.2512	0.1395	0.1954	-0.2134	-0.2134	0.2494	0.2501	0.1209	0.1346
11			0.5	0.2483	0.2490	-0.0603	0.0134	-0.4645	-0.4645	0.2481	0.2494	-0.0226	0.0041
12			1	0.2491	0.2500	-0.4156	-0.2634	-0.9648	-0.9648	0.2484	0.2491	-0.4135	-0.3634
13	500	-0.5	0.25	0.2511	0.2505	0.1071	0.1706	-0.2258	-0.2258	0.2511	0.2508	0.0312	0.0413
14			0.5	0.2504	0.2496	-0.0968	0.0137	-0.4756	-0.4756	0.2507	0.2494	-0.1214	-0.0938
15			1	0.2518	0.2512	-0.5464	-0.4004	-0.9755	-0.9754	0.2521	0.2507	-0.5607	-0.4958
16		-0.3	0.25	0.2516	0.2512	0.0783	0.1173	-0.2255	-0.2255	0.2497	0.2496	0.0598	0.0772
17			0.5	0.2508	0.2500	-0.0778	0.0206	-0.4763	-0.4762	0.2509	0.2505	-0.1355	-0.114
18			1	0.2525	0.2523	-0.5464	-0.4006	-0.9761	-0.9761	0.2522	0.2524	-0.5219	-0.4694
19		0.3	0.25	0.2491	0.2495	0.0617	0.1111	-0.2257	-0.2257	0.2485	0.2488	0.0514	0.0655
20			0.5	0.2489	0.2493	-0.1283	-0.0605	-0.4756	-0.4755	0.2479	0.2482	-0.1503	-0.1187
21			1	0.2506	0.2510	-0.5565	-0.4236	-0.9761	-0.9761	0.2507	0.2511	-0.5445	-0.4866
22		0.5	0.25	0.2492	0.2501	0.0637	0.1167	-0.2256	-0.2256	0.2501	0.2508	0.0193	0.0360
23			0.5	0.2498	0.2510	-0.1005	-0.0114	-0.4756	-0.4756	0.2502	0.2515	-0.1171	-0.0783
24			1	0.248	0.2488	-0.4737	-0.2981	-0.9756	-0.9756	0.2481	0.2491	-0.5377	-0.4998

Table A.12. 3

Bias of the Testlet Variance and Covariance Matrix

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	$\sigma_{\lambda_4}^2$		$\sigma_{\gamma_4\lambda_4}$		$\sigma_{\gamma_5}^2$		$\sigma_{\lambda_5}^2$		$\sigma_{\gamma_5\lambda_5}$	
				JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT
1	200	-0.5	0.25	-0.2141	-0.2142	0.2501	0.2494	0.1053	0.1274	-0.2141	-0.2141	0.2512	0.2503
2			0.5	-0.4649	-0.4649	0.2522	0.2512	-0.0466	-0.0218	-0.4645	-0.4645	0.2533	0.2525
3			1	-0.9633	-0.9632	0.2520	0.2504	-0.5182	-0.4635	-0.9637	-0.9637	0.2534	0.2520
4		-0.3	0.25	-0.2146	-0.2145	0.2512	0.2509	0.1398	0.1503	-0.2144	-0.2144	0.2481	0.2477
5			0.5	-0.4650	-0.4650	0.2504	0.2499	-0.0562	-0.0282	-0.4638	-0.4638	0.2520	0.2518
6			1	-0.9635	-0.9635	0.2525	0.2522	-0.5370	-0.4996	-0.9633	-0.9633	0.2504	0.2499
7		0.3	0.25	-0.2137	-0.2137	0.2493	0.2498	0.1402	0.1567	-0.2133	-0.2131	0.2485	0.2489
8			0.5	-0.4631	-0.4630	0.2502	0.2507	-0.0984	-0.0889	-0.4641	-0.4640	0.2514	0.2521
9			1	-0.9640	-0.9640	0.2464	0.2482	-0.4535	-0.3964	-0.9633	-0.9633	0.2492	0.2498
10		0.5	0.25	-0.2151	-0.2151	0.2484	0.2491	0.1204	0.1280	-0.2143	-0.2143	0.2511	0.2518
11			0.5	-0.4646	-0.4645	0.2460	0.2473	-0.0439	-0.0052	-0.4639	-0.4640	0.2458	0.2464
12			1	-0.9639	-0.9639	0.2516	0.2536	-0.4244	-0.4027	-0.9634	-0.9634	0.2492	0.2511
13	500	-0.5	0.25	-0.2253	-0.2252	0.2501	0.2493	0.0175	0.0372	-0.2252	-0.2252	0.2516	0.2510
14			0.5	-0.4763	-0.4762	0.2512	0.2502	-0.1478	-0.1156	-0.4760	-0.4760	0.2504	0.2497
15			1	-0.9760	-0.9760	0.2531	0.2521	-0.5254	-0.4637	-0.9753	-0.9753	0.2494	0.2483
16		-0.3	0.25	-0.2258	-0.2258	0.2526	0.2524	0.0354	0.0483	-0.2262	-0.2262	0.2514	0.2510
17			0.5	-0.4759	-0.4758	0.2509	0.2505	-0.1692	-0.1467	-0.4756	-0.4755	0.2493	0.2489
18			1	-0.9757	-0.9757	0.2521	0.2513	-0.5523	-0.5092	-0.9752	-0.9753	0.2535	0.2530
19		0.3	0.25	-0.2261	-0.2261	0.2494	0.2500	0.0574	0.0681	-0.2259	-0.2259	0.2505	0.2506
20			0.5	-0.4757	-0.4758	0.2514	0.2519	-0.1476	-0.1225	-0.4763	-0.4762	0.2498	0.2501
21			1	-0.9766	-0.9766	0.2491	0.2498	-0.5713	-0.5262	-0.9763	-0.9762	0.2496	0.2497
22		0.5	0.25	-0.2257	-0.2257	0.2488	0.2494	0.0371	0.0473	-0.2254	-0.2254	0.2495	0.2500
23			0.5	-0.4759	-0.4759	0.2503	0.2515	-0.1767	-0.1562	-0.4757	-0.4757	0.2489	0.2496
24			1	-0.9758	-0.9758	0.2487	0.2499	-0.5319	-0.4831	-0.9760	-0.9761	0.2486	0.2496

Table A.13. 1
SE of the Testlet Variance/Covariance Matrix

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	$\sigma_{\gamma_1}^2$		$\sigma_{\lambda_1}^2$		$\sigma_{\gamma_1\lambda_1}$		$\sigma_{\gamma_2}^2$		$\sigma_{\lambda_2}^2$	
				JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT
1	200	-0.5	0.25	0.1036	0.1149	0.0028	0.0027	0.0063	0.006	0.1377	0.1486	0.0032	0.0033
2			0.5	0.1562	0.1889	0.0030	0.0031	0.0101	0.0108	0.1010	0.1211	0.0028	0.0028
3			1	0.1910	0.2614	0.0035	0.0035	0.0133	0.0144	0.2284	0.2741	0.0037	0.0036
4		-0.3	0.25	0.1041	0.1150	0.0031	0.0031	0.0071	0.0076	0.1081	0.1107	0.0034	0.0034
5			0.5	0.1351	0.1301	0.0029	0.0028	0.0090	0.0093	0.1019	0.1322	0.0028	0.0028
6			1	0.1670	0.2241	0.0030	0.0030	0.0084	0.0096	0.2127	0.2447	0.0027	0.0027
7		0.3	0.25	0.1136	0.1456	0.0032	0.0032	0.0077	0.0088	0.1011	0.1081	0.0034	0.0034
8			0.5	0.1427	0.1735	0.0022	0.0022	0.0073	0.0074	0.1187	0.1322	0.0023	0.0023
9			1	0.2199	0.2595	0.0042	0.0042	0.0127	0.0141	0.1278	0.1690	0.0027	0.0028
10		0.5	0.25	0.1477	0.1857	0.0024	0.0024	0.0068	0.0073	0.0934	0.1130	0.0040	0.0040
11			0.5	0.1233	0.1425	0.0031	0.0031	0.0086	0.0086	0.1384	0.1467	0.0027	0.0027
12			1	0.1927	0.2120	0.0018	0.0018	0.0086	0.0099	0.1467	0.1631	0.0027	0.0026
13	500	-0.5	0.25	0.0560	0.0639	0.0021	0.0021	0.0054	0.0058	0.0651	0.0716	0.0026	0.0027
14			0.5	0.1348	0.1337	0.0018	0.0017	0.0071	0.0075	0.1109	0.1292	0.0016	0.0016
15			1	0.1501	0.1781	0.0020	0.0020	0.0057	0.0069	0.1089	0.1355	0.0017	0.0017
16		-0.3	0.25	0.1082	0.1249	0.0017	0.0017	0.0034	0.0037	0.0827	0.0950	0.0021	0.0021
17			0.5	0.1114	0.1197	0.0025	0.0025	0.0064	0.0070	0.1302	0.1310	0.0022	0.0022
18			1	0.1616	0.2090	0.0020	0.0019	0.0060	0.0070	0.1242	0.1354	0.0021	0.0021
19		0.3	0.25	0.0768	0.0807	0.0016	0.0016	0.0043	0.0046	0.0742	0.0843	0.0022	0.0021
20			0.5	0.1126	0.1212	0.0019	0.0019	0.0056	0.0061	0.0644	0.0699	0.0018	0.0018
21			1	0.1247	0.1469	0.0022	0.0022	0.0060	0.0070	0.1171	0.1426	0.0020	0.0020
22		0.5	0.25	0.0576	0.0676	0.0027	0.0028	0.0044	0.0046	0.0627	0.0632	0.0023	0.0022
23			0.5	0.0914	0.1179	0.0019	0.0018	0.0060	0.0068	0.0808	0.0923	0.0018	0.0018
24			1	0.1405	0.1699	0.0022	0.0022	0.0066	0.0076	0.0803	0.1143	0.0023	0.0023

Table A.13. 2

SE of the Testlet Variance and Covariance Matrix

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	$\sigma_{\gamma_2\lambda_2}$		$\sigma_{\gamma_3}^2$		$\sigma_{\lambda_3}^2$		$\sigma_{\gamma_3\lambda_3}$		$\sigma_{\gamma_4}^2$	
				JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT
1	200	-0.5	0.25	0.0076	0.0075	0.172	0.1911	0.0028	0.0028	0.0088	0.0100	0.0982	0.0955
2			0.5	0.0059	0.0064	0.2183	0.288	0.0019	0.0020	0.008	0.0099	0.1295	0.1313
3			1	0.0092	0.0098	0.2107	0.3023	0.0039	0.0038	0.0088	0.0105	0.1989	0.2093
4		-0.3	0.25	0.0085	0.0088	0.1459	0.2154	0.0029	0.0029	0.0055	0.0059	0.0922	0.0927
5			0.5	0.0064	0.0072	0.1383	0.1686	0.0036	0.0035	0.0102	0.0111	0.1145	0.1445
6			1	0.0121	0.0126	0.1835	0.2767	0.0023	0.0023	0.0068	0.0093	0.1655	0.1817
7		0.3	0.25	0.0073	0.0072	0.1385	0.1441	0.0029	0.0030	0.0097	0.0101	0.0807	0.0850
8			0.5	0.0065	0.0072	0.1203	0.1804	0.0028	0.0027	0.0071	0.0079	0.1052	0.1131
9			1	0.0105	0.0109	0.1598	0.2221	0.0027	0.0027	0.0096	0.0115	0.2900	0.3242
10		0.5	0.25	0.0077	0.0079	0.1183	0.1575	0.0028	0.0028	0.0072	0.0085	0.1295	0.1444
11			0.5	0.0060	0.0063	0.1459	0.1836	0.0024	0.0025	0.0055	0.0063	0.1712	0.1872
12			1	0.0084	0.0094	0.3041	0.3940	0.0028	0.0028	0.0104	0.0119	0.2156	0.2298
13	500	-0.5	0.25	0.0045	0.0045	0.1165	0.1477	0.0021	0.0022	0.0057	0.0063	0.0703	0.0734
14			0.5	0.0050	0.0051	0.1512	0.2067	0.0024	0.0023	0.0069	0.0083	0.1214	0.1268
15			1	0.0056	0.0061	0.1072	0.1562	0.0017	0.0016	0.0068	0.0085	0.1301	0.1599
16		-0.3	0.25	0.0062	0.0064	0.0809	0.0951	0.0018	0.0019	0.0047	0.0048	0.0955	0.0974
17			0.5	0.0050	0.0056	0.1595	0.2252	0.0020	0.0020	0.0050	0.0058	0.1011	0.1132
18			1	0.0081	0.0085	0.1465	0.2023	0.0022	0.0022	0.0083	0.0099	0.1445	0.1540
19		0.3	0.25	0.0045	0.0048	0.0924	0.1205	0.0024	0.0024	0.0048	0.0054	0.0799	0.0868
20			0.5	0.0038	0.0040	0.1153	0.1603	0.0022	0.0023	0.0068	0.0080	0.1206	0.1360
21			1	0.0068	0.0075	0.1190	0.1721	0.0018	0.0018	0.0062	0.0074	0.1423	0.1595
22		0.5	0.25	0.0045	0.0047	0.0807	0.1016	0.0020	0.0019	0.0036	0.0039	0.0601	0.0651
23			0.5	0.0041	0.0047	0.1463	0.2019	0.0016	0.0017	0.0067	0.0078	0.1289	0.1490
24			1	0.0054	0.0060	0.1550	0.2401	0.0017	0.0017	0.0087	0.0104	0.1553	0.1580

Table A.13. 3

SE of the Testlet Variance and Covariance Matrix

Cond. No.	N	$\rho_{\theta\tau}$	$\sigma_\gamma^2/\sigma_\lambda^2$	$\sigma_{\lambda_4}^2$		$\sigma_{\gamma_4\lambda_4}$		$\sigma_{\gamma_5}^2$		$\sigma_{\lambda_5}^2$		$\sigma_{\gamma_5\lambda_5}$	
				JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT	JAD-TT	JD-TT
1	200	-0.5	0.25	0.0028	0.0028	0.0077	0.0074	0.0885	0.1035	0.0031	0.0032	0.0075	0.0078
2			0.5	0.0021	0.0021	0.0082	0.0089	0.1459	0.158	0.0037	0.0039	0.009	0.0094
3			1	0.0035	0.0036	0.0129	0.0137	0.1758	0.2074	0.0029	0.0030	0.0104	0.0111
4		-0.3	0.25	0.0023	0.0023	0.0047	0.0047	0.1317	0.1346	0.0032	0.0032	0.0069	0.0072
5			0.5	0.0027	0.0028	0.0065	0.0065	0.1419	0.1471	0.0026	0.0025	0.0079	0.0086
6			1	0.0028	0.0028	0.0101	0.0106	0.1194	0.1583	0.0032	0.0032	0.0081	0.0082
7		0.3	0.25	0.0024	0.0024	0.008	0.0079	0.1405	0.1502	0.0029	0.0030	0.0074	0.0076
8			0.5	0.0037	0.0037	0.0072	0.0076	0.1239	0.1246	0.0046	0.0045	0.0082	0.0088
9			1	0.0033	0.0033	0.0121	0.0128	0.2200	0.2641	0.0029	0.0029	0.0099	0.0105
10		0.5	0.25	0.0027	0.0027	0.0057	0.0057	0.0990	0.1050	0.0034	0.0035	0.0071	0.0069
11			0.5	0.0027	0.0027	0.0098	0.0101	0.1571	0.1819	0.003	0.0030	0.0076	0.0082
12			1	0.0028	0.0028	0.0107	0.0114	0.2795	0.2678	0.0035	0.0034	0.0104	0.0100
13	500	-0.5	0.25	0.0018	0.0018	0.0045	0.0049	0.0780	0.0997	0.0018	0.0018	0.0044	0.0044
14			0.5	0.0019	0.0020	0.0059	0.0062	0.0818	0.0948	0.0024	0.0024	0.0063	0.0065
15			1	0.0018	0.0018	0.0072	0.0077	0.1567	0.1755	0.0015	0.0016	0.0061	0.0065
16		-0.3	0.25	0.0022	0.0022	0.0053	0.0055	0.0755	0.0807	0.0020	0.0020	0.0045	0.0046
17			0.5	0.0021	0.0021	0.0057	0.0062	0.0950	0.1027	0.0020	0.0019	0.0056	0.0054
18			1	0.0022	0.0022	0.0072	0.0072	0.1491	0.1654	0.0022	0.0022	0.0083	0.0086
19		0.3	0.25	0.0019	0.0020	0.0033	0.0038	0.1288	0.1322	0.0018	0.0018	0.0046	0.0046
20			0.5	0.0016	0.0015	0.005	0.0056	0.0803	0.0891	0.0021	0.0021	0.0058	0.0062
21			1	0.0017	0.0018	0.0057	0.0065	0.1354	0.1490	0.0018	0.0018	0.0055	0.0056
22		0.5	0.25	0.0019	0.0018	0.0044	0.0046	0.0791	0.0821	0.0026	0.0025	0.0050	0.0051
23			0.5	0.0018	0.0018	0.0057	0.0064	0.0752	0.0716	0.0022	0.0022	0.0055	0.0057
24			1	0.0016	0.0017	0.0074	0.0080	0.1457	0.1467	0.0022	0.0022	0.0091	0.0094

Bibliography

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

Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing, 22*(4), 509–531. https://doi.org/10.1191/0265532205lt317oa

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian Networks in Educational Assessment*. Springer-Verlag. https://doi.org/10.1007/978-1-4939-2125-6

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (2014). *Standards for Educational & Psychological Testing*. Washington DC: American Educational Research Association.

Anders, R., Alario, F.-X., & Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods, 21*(3), 309–327. https://doi.org/10.1037/met0000066

Belov, D. I. (2011). Detection of Answer Copying Based on the Structure of a High-Stakes Test. *Applied Psychological Measurement, 35*(7), 495–517.

Belov, D. I. (2015). Robust Detection of Examinees With Aberrant Answer Changes. *Journal of Educational Measurement, 52*(4), 437–456. https://doi.org/10.1111/jedm.12094

Belov, D. I. (2017). On the Optimality of the Detection of Examinees With Aberrant Answer Changes. *Applied Psychological Measurement, 41*(5), 338–352.

WorldCat.org. https://doi.org/10.1177/0146621617692077

Birnbaum, A. (1968) Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: Lord, F.M. and Novick, M.R., Eds., *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, 397-479.

Bock, R. D. (1997). The Nominal Categories Model. In Wim J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 33–49). Springer. https://doi.org/10.1007/978-1-4757-2691-6_2

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bolsinova, M., de Boeck, P., & Tijmstra, J. (2017). Modelling Conditional Dependence Between Response Time and Accuracy. *Psychometrika*, *82*(4), 1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bolsinova, M., & Tijmstra, J. (2019). Modeling Differences Between Response Times of Correct and Incorrect Responses. *Psychometrika*, *84*(4), 1018–1046. https://doi.org/10.1007/s11336-019-09682-5

Bradlow, E. T., Wainer, H., & Wang, X. (1999a). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168. https://doi.org/10.1007/BF02294533

Bradlow, E. T., Wainer, H., & Wang, X. (1999b). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168. https://doi.org/10.1007/BF02294533

Brent Bridgeman, Jun Xu, Nan Kong, Ou Lydia Liu, & Lixiong Gu. (2015).

Investigation of Response Changes in the GRE Revised General Test. *Educational and Psychological Measurement*, *75*(6), 1002–1020. https://doi.org/10.1177/0013164415573988

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods* (pp. xvi, 265). Sage Publications, Inc.

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*(2), 117–133. https://doi.org/10.3758/BF03204271

Cho, S.-J., Brown-Schmidt, S., Boeck, P. D., & Shen, J. (2020). Modeling Intensive Polytomous Time-Series Eye-Tracking Data: A Dynamic Tree-Based Item Response Model. *Psychometrika*, *85*(1), 154–184. https://doi.org/10.1007/s11336-020-09694-6

Cox, D.R., & Hinkley, D.V. (1974). Theoretical Statistics. London, UK: Chapman and Hall.

Davier, M. von, & Lee, Y.-S. (Eds.). (2019). *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *The British Journal of Mathematical and Statistical Psychology*, *70*(2), 225–237. https://doi.org/10.1111/bmsp.12094

De Boeck, P., & Jeon, M. (2019). An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, *10*, 102. https://doi.org/10.3389/fpsyg.2019.00102

de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Hong, Y. (2010). Parameter Estimation With Small Sample Size A Higher-Order IRT Model Approach. Applied Psychological Measurement, 34(4), 267–285. https://doi.org/10.1177/0146621608329501

de la Torre, J., & Song, H. (2009). Simultaneous Estimation of Overall and Domain Abilities: A Higher-Order IRT Model Approach. *Applied Psychological Measurement*, 33(8), 620–639. https://doi.org/10.1177/0146621608326423

DeCarlo, L. T. (2011). On the Analysis of Fraction Subtraction Data: The DINA Model, Classification, Latent Class Sizes, and the Q-Matrix. *Applied Psychological Measurement*, *35*(1), 8–26. https://doi.org/10.1177/0146621610377081

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A. A. (2019). Application of Cognitive Diagnostic Models to Learning and Assessment Systems. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (pp. 437–460). Springer International

Publishing. https://doi.org/10.1007/978-3-030-05584-4_21

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*(2), 175–186. https://doi.org/10.1007/BF02294171

Ercikan, K., Pellegrino, J. W., & Pellegrino, J. W. (2017, March 27). *Validation of Score Meaning Using Examinee Response Processes for the Next Generation of Assessments*. Validation of Score Meaning for the Next Generation of Assessments; Routledge. https://doi.org/10.4324/9781315708591-1

Fang, G., Liu, J., & Ying, Z. (2017). On the Identifiability of Diagnostic Classification Models. *ArXiv:1706.01240 [Math, Stat]*. http://arxiv.org/abs/1706.01240

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. Springer-Verlag. https://doi.org/10.1007/978-1-4419-0742-4

Fox, J.-P., Wenzel, J., & Klotzke, K. (2020). The Bayesian Covariance Structure Model for Testlets. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998620941204

Fujimoto, K. A. (2018). A general Bayesian multilevel multidimensional IRT model for locally dependent data. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 536–560. https://doi.org/10.1111/bmsp.12133

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press.

Goldhammer, F., Steinwascher, M. A., Kroehne, U., & Naumann, J. (2017). Modelling individual response time effects between and within experimental speed conditions: A GLMM approach for speeded tests. *British Journal of*

Mathematical and Statistical Psychology, *70*(2), 238–256.

Haertel, E. H. (1989). Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items. *Journal of Educational Measurement*, *26*(4), 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

Henson, R. A., Templin, J. L., & Willse, J. T. (2008). Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika*, *74*(2), 191. https://doi.org/10.1007/s11336-008-9089-5

Higham, P. A., & Gerrard, C. (2005). Not all errors are created equal: Metacognition and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *59*(1), 28–34. https://doi.org/10.1037/h0087457

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*(4), 382–386. https://doi.org/10.1037/h0021740

Huang, H.-Y., & Wang, W.-C. (2013). Higher Order Testlet Response Models for Hierarchical Latent Traits and Testlet-Based Items. *Educational and Psychological Measurement*, *73*(3), 491–511.

Im, K. S. (2017). *The Hierarchical Testlet Response Time Model: Bayesian Analysis of a Testlet Model for Item Responses and Response Times*. Doctoral dissertation, University of Kansas, Lawrence, KS.

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling Answer Change Behavior: An Application of a Generalized Item Response Tree Model. *Journal of Educational and Behavioral Statistics*, 42(4), 467–490.

https://doi.org/10.3102/1076998616688015

Jiao, H., Ding, Y., & Yin, C. (2020, July). *Joint modeling of responses, response time, and answer change patterns for cognitive diagnosis*. Paper presented at the annual International Meeting of the Psychometric Society.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A Multilevel Testlet Model for Dual Local Dependence. *Journal of Educational Measurement*, *49*(1), 82–100.

Jiao, H., Liao, D., & Zhan, P. (2019). Utilizing Process Data for Cognitive Diagnosis. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (pp. 421–436). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_20

Jiao, H., & Lissitz, R. (eds.). (2020). *Innovative Psychometric Modeling and Methods*. Charlotte, NC: Information Publishing.

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. https://doi.org/10.1111/bmsp.12035

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement*, 38: 79-

93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

Kane, M., & Mislevy, R. (2017, March 27). *Validating Score Interpretations Based on Response Processes*. Validation of Score Meaning for the Next Generation of Assessments; Routledge. https://doi.org/10.4324/9781315708591-2

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *Ann. Math. Statist.*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: CRC Press.

Linden, W. J. V. D. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of Response Changes in the GRE Revised General Test. Educational and Psychological Measurement, 75(6), 1002–1020. https://doi.org/10.1177/0013164415573988

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*,

1171. https://doi.org/10.3389/fpsyg.2015.01171

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika, 76*(3), 487–503. https://doi.org/10.1007/s11336-011-9211-y

Luce, R. D., & Luce, V. S. T. P. of P. R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. OUP USA.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavior Statistics.* 2, 99–120. doi: 10.3102/10769986002002099

Man, K., & Harring, J. R. (2019). Negative Binomial Models for Visual Fixation Counts on Test Items. *Educational and Psychological Measurement*. 79(4), 617-635. doi:10.1177/0013164418824148

Maris, E. (1993a). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*(3), 445–469. https://doi.org/10.1007/BF02294651

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika,* 47, 149-174

Masters, G. N. (2018). *Partial Credit Model*. Handbook of Item Response Theory; Chapman and Hall/CRC. https://doi.org/10.1201/9781315119144-7

Masters, G. N., & Wright, B. D. (1997). The Partial Credit Model. In Wim J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 101–121). Springer. https://doi.org/10.1007/978-1-4757-2691-6_6

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-

Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817. https://doi.org/10.3758/PBR.16.5.798

Merkle, E., Furr, D., & Rabe-Hesketh, S. (2018). Bayesian model assessment: Use of conditional vs marginal likelihoods. arXiv preprint arXiv:1802.04452.

Milia, L. D. (2007). Benefiting from Multiple-Choice Exams: The positive impact of answer switching. *Educational Psychology*, *27*(5), 607–615. https://doi.org/10.1080/01443410701309142

Mislevy, R. J. (2018). *Sociocognitive Foundations of Educational Measurement*. New York, NY: Routledge.

Molenaar, D., & de Boeck, P. (2018). Response Mixture Modeling: Accounting for Heterogeneity in Item Characteristics across Response Times. *Psychometrika*, *83*(2), 279–297. https://doi.org/10.1007/s11336-017-9602-9

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment* (pp. x, 471). Lawrence Erlbaum Associates, Inc.

Oranje, A., Gorin, J., Jia, Y., Kerr, D., Gorin, J., Jia, Y., & Kerr, D. (2017, March 27). *Collecting, Analyzing, and Interpreting Response Time, Eye-Tracking, and Log Data*. Validation of Score Meaning for the Next Generation of Assessments; Routledge. https://doi.org/10.4324/9781315708591-4

Plummer, M. (2017). *JAGS Version 4.0.0 User Manual*. Lyon. Available online at: http://sourceforge.net/projects/mcmc-jags/

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of*

Philosophy, 14, 58–94.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ranger, J., & Kuhn, J.-T. (2015). Modeling Information Accumulation in Psychological Tests Using Item Response Times. *Journal of Educational and Behavioral Statistics*, *40*(3), 274–306.

Ranger, J., & Kuhn, J.-T. (2018). Estimating Diffusion-Based Item Response Theory Models: Exploring the Robustness of Three Old and Two New Estimators. *Journal of Educational and Behavioral Statistics*, *43*(6), 635–662.

Ranger, J., Kuhn, J.-T., & Gaviria, J.-L. (2015). A Race Model for Responses and Response Times in Tests. *Psychometrika*, *80*(3), 791–810. https://doi.org/10.1007/s11336-014-9427-8

Ratcliff, R., & McKoon, G. (2007). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In *Progress in mathematical psychology*, *1*. (pp. 151–174). Elsevier Science.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The Lognormal Race: A Cognitive-Process Model of Choice and Latency with Desirable Psychometric Properties. *Psychometrika*, *80*(2), 491–513. https://doi.org/10.1007/s11336-013-9396-3

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

Samejima, F. (1968). Estimation of Latent Ability Using a Response Pattern of Graded Scores1. *ETS Research Bulletin Series*, *1968*(1), i–169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

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

Sinharay, S., & Johnson, M. S. (2017). Three New Methods for Analysis of Answer Changes. *Educational and Psychological Measurement*, *77*(1), 54–81. WorldCat.org. https://doi.org/10.1177/0013164416632287

Sinharay S. (2018). Detecting Fraudulent Erasures at an Aggregate Level. *Journal of Educational and Behavioral Statistics*, *43*(3), 286–315. https://doi.org/10.3102/1076998617739626

Sinharay, S. (2018). Detecting Fraudulent Erasures at an Aggregate Level. *Journal of Educational and Behavioral Statistics*, *43*(3), 286–315. https://doi.org/10.3102/1076998617739626

Sinharay, S., Duong, M. Q., & Wood, S. W. (2017a). A New Statistic for Detection of Aberrant Answer Changes. *Journal of Educational Measurement*, *54*(2), 200–217. https://doi.org/10.1111/jedm.12141

Sinharay, S., Duong, M. Q., & Wood, S. W. (2017b). A New Statistic for Detection of Aberrant Answer Changes. *Journal of Educational Measurement*, *54*(2), 200–217. https://doi.org/10.1111/jedm.12141

Sinharay, S., Duong, M. Q., & Wood, S. W. (2017c). A New Statistic for Detection of Aberrant Answer Changes. *Journal of Educational Measurement*, *54*(2), 200–217. https://doi.org/10.1111/jedm.12141

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Su, Y. S., & Yajima, M. (2020). Package R2jags: Using R to run 'JAGS'. R package version 0.6‐1. *Available online at: https://cran.r-project.org/web/packages/R2jags/R2jags.pdf*.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In *Cognitively diagnostic assessment* (pp. 327–359). Lawrence Erlbaum Associates, Inc.

Thissen, D. (1983). Timed testing: An Approach Using Item Response Theory. In D. J. WEISS (Ed.), *New Horizons in Testing* (pp. 179–203). Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50019-6

Torre, J. de la, & Minchen, N. D. (2019). The G-DINA Model Framework. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (pp. 155–169). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_7

Torre, J. D. L. (2008). An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications. *Journal of Educational Measurement*, *45*(4), 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

Townsend, J. T., Ashby, F. G., & Ashby, F. G. (2014, January 14). *Methods of Modeling Capacity in Simple Processing Systems*. Cognitive Theory; Psychology Press. https://doi.org/10.4324/9781315802473-14

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650. https://doi.org/10.1007/s11336-000-0810-3

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. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, *46*(3), 247–272. WorldCat.org.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130. https://doi.org/10.1111/j.1745-3984.2007.00030.x

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical Tests of Conditional Independence Between Responses and/or Response Times on Test Items. *Psychometrika*, *75*(1), 120–139. https://doi.org/10.1007/s11336-009-9129-9

van der Linden, W. J., & Jeon, M. (2012). Modeling Answer Changes on Test Items. *Journal of Educational and Behavioral Statistics*, 37(1), 180–199. https://doi.org/10.3102/1076998610396899

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. https://doi.org/10.1177/0146621609349800

van der Linden, W. J., & Guo, F. (2008). Bayesian Procedures for Identifying Aberrant Response-Time Patterns in Adaptive Testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

van der Maas, H., 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. https://doi.org/10.1037/a0022749

van Gog, T., & Scheiter, K. (2020). *Eye tracking as a tool to study and enhance multimedia learning*. Learning and Instruction, 20(2), 95-99.

Wang, X., Bradlow E. T., & Wainer, H. (2002) A General Bayesian Model for Testlets: Theory and Applications. *Applied Psychological Measurement*. 26(1):109-128. doi:10.1177/0146621602026001007

Wainer, H., Bradlow, E. T., and Wang, X. H. (2007). *Testlet Response Theory and Its Applications*. New York, NY: Cambridge University Press.

Wainer, H. & Kiely, G. L. (1987), Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement,* 24: 185-201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Lewis, C. (1990). Toward a Psychometrics for Testlets. *Journal of*

Educational Measurement, *27*(1), 1–14. https://doi.org/10.1111/j.1745-3984.1990.tb00730.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203–220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

Wald, A. (1947). *Sequential Analysis*. New York: John Wiley & Sons.

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. https://doi.org/10.1111/bmsp.12054

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. https://doi.org/10.1177/0146621605275984

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models. *Psychological Methods*, *9*(1), 116–136. https://doi.org/10.1037/1082-989X.9.1.116

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A General Bayesian Model for Testlets: Theory and Applications. *Applied Psychological Measurement*, *26*(1), 109–128.

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting Test Tampering Using Item Response Theory. *Educational and Psychological Measurement*,

75(6), 931–953. https://doi.org/10.1177/0013164414568716

Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining Error Rates and Power for Detecting Answer Copying. *Applied Psychological Measurement*, 25(4), 385–404. https://doi.org/10.1177/01466210122032118

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zhan, P., Jiao, H., & Liao, D. (2018a). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, 71(2), 262–286. https://doi.org/10.1111/bmsp.12114

Zhan, P., Liao, M., & Bian, Y. (2018b). Joint Testlet Cognitive Diagnosis Modeling for Paired Local Item Dependence in Response Times and Response Accuracy. *Frontiers in Psychology, 9,* 607. https://doi.org/10.3389/fpsyg.2018.00607

Van Gog, T., & Scheiter, K. (2020). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, 20(2), 95-99.