June 2021

Evaluating Approaches for Dealing with Omitted Items in Large-Scale Assessments

Seong Eun Hong
University of Massachusetts Amherst

Evaluating Approaches for Dealing with Omitted Items in Large-Scale Assessments

Seong Eun Hong

Evaluating Approaches for Dealing with Omitted Items in Large-Scale Assessments

A Dissertation Presented

by

SEONG EUN HONG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2021

College of Education

Research, Educational Measurement, and Psychometrics

Evaluating Approaches for Dealing with Omitted Items in Large-Scale Assessments

A Dissertation Presented

by

SEONG EUN HONG

Approved as to style and content by:

_____
Scott Monroe, Chair

_____
Craig S. Wells, Member

_____
John Staudenmayer, Member

_____
Jennifer Randall,
Associate Dean of Academic Affairs
College of Education

ACKNOWLEDGMENTS

I cannot express enough thanks to several people for their continued support and encouragement on this project. First, I feel incredibly grateful to my committee for their insightful comments, time and efforts: Dr. Scott Monroe, Dr. Craig S. Wells, and Dr. John Staudenmayer. I offer my sincere appreciation for the amazing learning opportunities.

In particular, I am most thankful for my academic advisor, Dr. Scott Monroe, who gave me the foundation for this topic, explained complex technical questions, and has been always supportive in my growing as a researcher.

In addition to my committee, I would like to express my sincere gratitude to my academic mentor, Professor Ronald K. Hambleton, who helped me grow both academically and professionally. It has been an honor to work with you.

Last but not least, my completion of this project could not have been accomplished without the support of my family, friends and REMP family. My parents always believed in me and encouraged me to give my best. My friends, Eun and Jean provided constant support and encouragement. REMP students have been incredibly supportive that I would never forget our lunches, kind advice, and countless conversations.

ABSTRACT

Evaluating Approaches for Dealing with Omitted Items in Large-Scale Assessments

MAY 2021

SEONG EUN HONG, B.A., CORNELL UNIVERSITY

M.A., COLUMBIA UNIVERSITY, TEACHERS COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Scott Monroe

Large-scale assessments (LSAs), such as the National Assessment of Educational Progress (NAEP) are low-stakes tests for examinees; consequently, they might randomly guess or generate no responses. Such disengaged test-taking behavior can undermine the validity of test score interpretation. To account for such behavior, various methods have been proposed over the years, which can be classified as ad hoc or model-based. For instance, the Programme for the International Assessment of Adult Competencies (PIAAC) uses a common time threshold (e.g., 5 seconds) method for all items: if an examinee spends more than or equal to five seconds on an item, the omitted response is coded as incorrect; otherwise, it is coded as ignored. Recently, the speed-accuracy+omission (i.e., SA+O model) has been proposed for modeling the processes underlying response and nonresponse behavior. The present research aims to investigate the impact of omitted responses on item and person parameter estimates with the ad hoc and the model-based approaches in the context of LSAs. In a simulation study, we examine (a) how ad hoc and model-based approaches for handling omitted responses compare in terms of item and person parameter estimation in IRT and (b) whether there is

a practical difference between ad hoc and model-based approaches to handling omitted responses in real data analyses. Finally, we illustrate the practical implications of selecting a certain approach for handling the omitted items in LSAs through an empirical analysis.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Since the 21st century, the number of countries participating in international surveys has substantially increased. For the Trends in International Mathematics and Science Study (TIMSS), the number of participating countries has grown from 38 to 58 (NCES, 2020a), and for the Progress in Reading Literacy Study (PIRLS) and the Programme for International Student Assessment (PISA), it has grown from 34 to 49, and 43 to 79, respectively (NCES, 2020b; NCES, 2020c). With the growth of international large-scale assessments (LSA), more data have become available, and studies on secondary data sets have also notably increased. For instance, cross-sectional estimates of achievement are provided, in addition to student background information, including their homes, teachers, and schools. The substantial amount of data from various countries are oftentimes publicly available to researchers with online tools (e.g., IDB Analyzer, NAEP Data Explorer) for data extraction. Further, since the data are gathered at multiple levels (e.g., classroom, school, country), different units of analysis can be used for further investigation.

However, there are several limitations on the use of LSAs. First, the features of the LSAs do not explain cause-and-effect relationships; in other words, the data cannot be used to answer the following research question: what causes educational outcomes to change (Chudgar, & Luschei, 2016)? In addition, the technical complexities of LSA data prevent certain research questions. For example, simple random sampling is not used, and different sets of cognitive items are administered to each participant (OECD, 2013; OECD, 2017). To deal with this problem, sampling weights are used to reflect that some units (e.g., students, teachers, or

schools) are selected with different probabilities (Rutkowski et al., 2010). Moreover, since the individual proficiency estimates are biased in LSAs (Mislevy et al., 1992), the population-level proficiency estimates (i.e., plausible values) are reported, instead. Further, due to the stratified multistage sampling design used in LSAs, standard errors of estimates based on the random sampling assumption cannot be calculated; instead, special methods need to be employed to estimate the uncertainty associated with sampling (Rutkowski et al., 2010). Overall, researchers need to be well aware of the prominent issues associated with analyzing the LSAs, including level of analysis, sampling weights, plausible values, and variance estimation.

Another concern about the LSAs is examinee's lack of motivation. One of the purposes of LSAs is to measure group proficiency, which is based on examinee proficiency. For an accurate interpretation of proficiency estimates, it is assumed that examinees are actively engaged to answer every item correctly. For high-stakes tests, a lack of motivation is generally not a concern because test results have significant consequences for examinees such as receipt of a high school diploma, a scholarship, or a license to practice a profession. However, the LSAs are low-stakes tests that have little or no consequences for the examinees. Unlike the high-stakes tests, unmotivated examinees taking low-stakes tests might randomly guess on multiple-choice (MC) items or generate no responses without even reading or attempting an item (Wise & Gao, 2017). This can cause a serious threat to the validity of proficiency estimates in LSAs because test scores can be confounded with the level of disengagement (Braun, Kirsch, & Yamamoto, 2011).

Traditionally, missing responses have been dealt with using ad hoc approaches. For instance, missing responses are coded as incorrect, ignored, or partially correct for further

analysis. Unfortunately, there is no consensus among researchers regarding how to ideally deal with missing responses in IRT models.

Recently, new technologies have been implemented in LSAs. For instance, the Program for the International Assessment of Adult Competencies (PIAAC) and PISA have changed the mode of administration from a paper-based assessment (PBA) to a computer-based assessment (CBA). CBA allows for introducing new item types, measuring new constructs, and increasing efficiency. CBA also collects log data such as click or touch event (e.g., using a button, link, or menu), keystroke event (e.g., entering text), focus-in and out event (e.g., scrolling, zooming) and view event (e.g., page is loaded and displayed). The log data have following properties: log data are event-based (i.e., events are collected based on examinee's behavior), events are of different types and events have time stamps representing the temporal relations of events (Kroehne & Goldhammer, 2018). Further, CBA allows for collection of the response time (RT), which provides information about response process. There has been extensive research of reaction times (e.g., speed-accuracy trade-offs) in psychology (van der Linden, 2007; Ulitzsch, von Davier, & Pohl, 2019; Ulitzsch, von Davier, & Pohl, 2020).

In particular, RTs can be used to investigate cognitive processes. RT modeling approaches can be classified into four categories (De Boeck & Jeon, 2019, p. 2):

1) RT models: RTs are used as the sole dependent variable (e.g., distribution models, explanatory models, and models with response accuracy as a covariate).
2) Joint models: RTs and response accuracy (RA) are joint dependent variables (e.g., hierarchical model, diffusion model, race models).

3) Dependency models: RTs and RA are jointly modeled with the possibility of dependencies beyond the relationship of latent variables and item parameters so that they can explain an extra dependency.
4) RTs as covariate models: RTs are used as a covariate and RA as a dependent variable.

Overall, RTs can be incorporated in modeling test data in various ways to identify and measure cognitive processes.

In educational measurement, RT is widely used to provide relevant information about examinees. First, RT is used to identify disengaged test-taking behavior-that is, a response occurs so rapidly that an examinee does not take the necessary time to read, understand, and fully consider the item (i.e., rapid guessing) (Schnipke, 1995; Wise & Kong, 2005). Second, RT is used to improve item and person parameter estimates (van der Linden, 2007; Guo et al., 2016; Wise & Kong, 2005). Most recently, RTs have accounted for omitted, not-reached items, and disengaged test-taking behavior to reduce bias of the item and person parameter estimates (Pohl, Ulitzsch, & von Davier, 2019; Ulitzsch et al., 2019; Ulitzsch et al., 2020).

1.2 Statement of the Problem and Its Significance

Since the LSAs are low-stakes tests, examinees can exhibit disengaged test-taking behavior. Unmotivated examinees might not take the necessary time to consider the item; instead, they are likely to omit some of the items (i.e., item-level nonresponse) or generate nonresponses for the last few items due to the time limit (i.e., not-reached items). As a result, the LSA data can contain a significant number of missing responses. In 2012 PIAAC, for example, the rate of omitted responses ranged from 2% for the numeracy domain in South Korea to 25.9% for the literacy domain in Chile (OECD, 2013). In the National Assessment of Educational Progress (NAEP) mathematics assessment of 1990, the highest rate of not-reached items was

45% (Koretz et al., 1993). In PISA 2006, the proportion of omitted responses and not reached items varied substantially from 1% in Netherlands to 16% in Kyrgyzstan and 0.3% in Azerbaijan to 13% in Colombia, respectively (OECD, 2009, p.220).

This considerable amount of missing responses needs to be taken into account in psychometric analysis of test data. In particular, ignoring or not appropriately dealing with omitted and not-reached items can lead to biased item parameter and proficiency estimates (Lord, 1974; Mislevy & Wu, 1996; Pohl, Gräfe, & Rose, 2014) as well as biased estimates of group statistics. Furthermore, the presence of nonignorable omitted and not-reached items in the data set can lead to a different country ranking and biased regression coefficients for predicting test performance from explanatory variables (Köhler, Pohl, & Carstensen, 2015a; Rose, von Davier, & Xu, 2010).

1.3 Purpose of Study

The purpose of the present research is to investigate the impact of omitted responses on item and person parameter estimates with the ad hoc and the model-based approaches in the context of LSAs. In particular, the performance of the simultaneous modeling of item response and nonresponse behavior as well as the associated RTs and NRTs (Ulitzsch et al., 2020) is compared with the current approaches on omitted response treatments in LSAs. To that end, this research addresses the following research questions.

First, how do ad hoc and model-based approaches for handling omitted responses compare in terms of item and person parameter estimation in IRT in a simulation study? In particular, what factors (e.g., omission rates, sample size, test length, correlation between proficiency and omission propensity) in simulation conditions substantially affect the item parameter estimates and proficiency estimates? Further, it is of interest to investigate the

accuracy and efficiency associated with item parameter estimates and proficiency estimates under different data-generating conditions (i.e., whether proficiency is related to omission propensity). In particular, the bias and root mean square error/mean absolute error of values of all parameters as well as the means of standard deviations of proficiency estimates are of interest.

Second, in real data analyses, is there a practical difference between ad hoc and model-based approaches to handling omitted responses? We illustrate the practical implications of selecting a certain approach for handling the omitted items in LSAs through an empirical analysis. It is also of interest to investigate the accuracy and efficiency associated with item parameter estimates and proficiency estimates under different conditions.

CHAPTER 2

LITERATURE REVIEW

2.1 Brief Review of Item Response Theory

In educational measurement, item response theory (IRT) models the probability of answering an item correctly, given the characteristics of item and examinee proficiency. In other words, IRT provides a scale for the underlying latent variable (i.e., proficiency), measured by the test items (Thissen, & Orlando, 2001). The common assumptions of IRT models include unidimensionality (e.g., there is a single dominant proficiency being measured), local independence (e.g., item responses are mutually independent given a proficiency level), and monotonicity (e.g., probability is the dependent variable; as the probability of a correct answer increases, the proficiency also increases) (Holland & Rosenbaum, 1986).

Compared to classical measurement models, IRT has several advantages (Hambleton, Swaminathan, & Rogers, 1991, p 5):

1) Item characteristics are not group-dependent;
2) Scores describing examinee proficiency that are not test-dependent;
3) A model that is expressed at the item level rather than at the test level;
4) A model that does not require strictly parallel tests for assessing reliability;
5) A model that provides a measure of precision for each ability score.

Given the desired features of IRT, it has been widely used in large-scale test development and scoring.

Let there be $i = 1, \dots, I$ items and $j = 1, \dots, N$ examinees. For examinee j, let Y_{ij} be the observed response and y_{ij} a possible value. Under the IRT model, the conditional probability function of the complete data for examinee j is defined as:

$$P\left(Y_{1j} = y_{1j}, \ldots, Y_{Ij} = y_{Ij} \middle| \theta_j, a_i, b_i, c_i\right) = \prod_{i=1}^{I} f_\theta(y_{ij}|a_i, b_i, c_i) \quad (2.1.1)$$

where θ_j represents the proficiency parameter of examinee j, a_i, b_i, and c_i denotes the discrimination, difficulty, and guessing parameters for item i, respectively (See Figure 4) and f_θ is a likelihood function $L(\theta|Y_{ij})$. More specifically, the three-parameter logistic model (i.e., 3PL model) is assumed and the probability of success of an item is defined as

$$P_i\left(Y_{ij} = 1|\theta_j, a_i, b_i, c_i\right) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad (2.1.2)$$

and $P_i\left(Y_{ij} = 0|\theta_j, a_i, b_i, c_i\right) = 1 - P_i\left(Y_{ij} = 1|\theta_j, a_i, b_i, c_i\right)$. When $c_i = 0$, the 3PL model reduces to a 2PL model. Further, when $c_i = 0$ and $a_i = 1$, the 3PL model reduces to the Rasch model. IRT models are commonly used to estimate examinee proficiency in LSAs. In TIMSS (Martin, Mullis, & Hooper, 2016) and PIRLS (Martin, Mullis, & Hooper, 2017), for dichotomously scored items, a 2PL model is used for the short-constructed response items and a 3PL model is used for the multiple-choice items. In addition, the Rasch model had been used in prior PISA cycles (2000-2012) for dichotomously scored responses, while the 2PL model is implemented in PISA 2015 (OECD, 2017). For this paper, the marginal estimation (Martin et al., 2016; Martin, et al., 2017) is used for ad hoc approaches, while the Bayesian estimation (Ulitzsch et al., 2019; van der Linden, 2007) is used for model-based approach.

2.2 Missing Data Mechanism and Ignorability

Missing data are inevitable in educational measurement. Even worse, any method for compensating for missing data requires unverifiable assumptions, and further, missing data complicate likelihood-based inferences (Little, 2009). To evaluate the consequences of missing data, it is important to consider potential reasons for the missingness. Missing data patterns are

characterized by three different processes: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976).

Let Y_i be the observed response for item i and $Y = (Y_1, \ldots Y_I)$ be the vector of observed response to all items. The missing data indicator D_i is defined:

$$D_i = \begin{cases} 0 & \text{if } Y_i \text{ is observed} \\ 1 & \text{otherwise.} \end{cases} \quad (2.2.1)$$

and $D = (D_1, \ldots D_I)$ be the vector of missing data indicator to all items. The Y is also partitioned into two parts: observed data Y^o for which $D_i = 1$ and missing data Y^m for which $D_i = 0$. Under the IRT model, the probability function of the complete data is defined as the Equation 2.1.1. Inferences about the proficiency, θ need to be based on the observed data, (Y^o, D) and our beliefs about the missing mechanism (Mislevy, 2017).

2.2.1 MCAR

When the probability of missingness is independent of the observed responses, the missing mechanism can be described as MCAR (Rubin, 1976). In other words, the probability of a missingness pattern does not depend on the missing responses and observed responses (Mislevy, 2017):

$$g_\phi(D|Y) = g_\phi(D). \quad (2.2.2)$$

where ϕ indicates the vector of parameters of missing mechanism and $g_\phi(D|Y)$ (i.e., $P\{D;\ \phi \mid Y\}$) represents the missing mechanism. Under a MCAR mechanism, the process of generating the missing values can be ignored. In other words, simple averages of the observed data provide unbiased estimates of the corresponding population means; thus, observed data can be treated as a random sample of the complete data set (Molenberghs & Kenward, 2007). For instance, when missing values result from a priori fixed incomplete test and calibration designs, they can be treated as MCAR, since the design has been fixed in prior (Holman, & Glas, 2005).

Another example of MCAR in educational measurement is a random assignment of alternate test forms (Mislevy, & Wu, 1996).

2.2.2 MAR

Under MAR, the probability of missingness pattern does not depend on the missing responses, but conditionally depends on the observed responses (Mislevy & Wu, 1996; Rubin, 1976). The process of MAR is defined as (Mislevy, 2017):

$$g_{\phi}(D|Y) = g_{\phi}(D|Y^{o}). \quad (2.2.3)$$

For instance, given the income data on gender and age, the missing value on income for a male aged 40 or more can be predicted to have a high income because other males aged 40 or more have high income. If a probabilistic relationship can be derived from the observed data, the missingness is MAR. In the IRT context, the data collected from the computerized adaptive testing (CAT) and multistage testing (MST) are MAR because the items administered are completely directed by the observed responses, but independent from the unobserved responses (Holman & Glas, 2005).

2.2.3 MNAR

Under MNAR, the probability of a missingness pattern may depend on unobserved variables. Specifically, if neither MCAR nor MAR assumption holds, it can be addressed as MNAR. Inferences can be drawn by making further assumptions about which observed responses carry no information, and there are three popular models for handling a MNAR mechanism: selection models (Heckman, 1976, 1979), pattern-mixture models (Little, 1993, 2009), and shared-parameter models (Follmann, & Wu, 1995). For instance, when the probability that an item response is missing is due to the response itself, the missing data are MNAR (Lord, 1974; Mislevy, & Wu, 1996). As an example, missing values might occur when

examinees fail to give responses to specific items due to a lack of motivation (e.g., omitted responses).

2.2.4 Ignorability

For IRT, we can consider the joint parameter space (θ, λ, ϕ). If the joint parameter space of (θ, λ, ϕ) factors into a θ, λ and ϕ space, they are independent and missingness process is ignorable. In other words, likelihood estimation and inference for θ and λ can be carried out while ignoring the missing data mechanism. On the other hand, if the missingness process depends on θ or λ, the parameter space cannot be factored out and the missingness process is nonignorable (Mislevy, 2017). Further, within the likelihood framework, both MCAR and MAR are ignorable missing processes (Rubin, 1976). Taken together, if both MAR (or MCAR) assumption and distinctiveness of parameters hold, ignorability results where likelihood function can be factorized into the likelihood for observed responses and that for missing mechanism (Feldman, & Rabe-Hesketh, 2012). However, if neither MCAR nor MAR holds, the missingness process is MNAR. For instance, if $g_{\phi}(D|Y)$ also depends on θ.

To address MCAR, listwise deletion (e.g., each individual with any missing value is excluded) or pairwise deletion (e.g., given a pair of variables, examinees missing either item in the pair are excluded) are adequate approaches. However, MAR and MNAR mechanisms make untestable assumptions, and there is no valid method to select the most appropriate model (Little & Rubin, 2002; Molenberghs, Beunckens, Sotto & Kenward, 2008). Thus, it is only possible to compare the model fit (or parameter estimates) under MAR and MNAR (Sterba & Gottfredson, 2015). Consequently, model choice should be based on theories, and sensitivity analyses are recommended for investigating the impact of different assumptions on the substantive conclusions (Feldman & Rabe-Hesketh, 2012).

2.3 Omitted and Not-Reached Items in LSAs

If missingness process is MAR or MNAR, the missing data raise various concerns for statistical analyses because they reduce statistical power and cause bias in parameter estimates. In theory, the impact of missing data can be statistically adjusted (i.e., treated as ignorable), if all the process variables associated with the missing data are identified and modeled appropriately (Little & Rubin, 1987). In practice, however, it is extremely challenging to find all the process variables for the corresponding missing data. In educational measurement, test data usually include missing responses due to various reasons: 1) items that are not administered, 2) omitted items, or 3) not-reached items because of time limits.

In particular, the amount of missing data in LSAs is not negligible. In 2012 PIAAC, for example, the average proportion of nonresponse (i.e., omitted or not-reached) for computer-based items is 7.2% for literacy domain and 4.9% for numeracy domain; however, the proportion of nonresponses varies markedly across countries, ranging from 2% for the numeracy domain in South Korea to 25.9% for the literacy domain in Chile (OECD, 2013). In PISA 2006, the average proportion of omitted responses and not reached items varied substantially from 1% in Netherlands to 16% in Kyrgyzstan and 0.3% in Azerbaijan to 13% in Colombia, respectively (OECD, 2009, p.220). For the mathematics in NAEP 1990, the average rate of omitted items differed by grades: 9% for Grade 12, and only 5% for Grade 4, and the rate of not-reached items was 8% (Koretz et al., 1993). For the science in TIMSS 2003, the average proportion of not-reached items noticeably differed by booklets from 0.4% to 17% (Mullis, Martin, & Diaconu, 2004, p. 249).

2.3.1 Definitions of Omitted and Not-Reached Items

In LSAs, omitted and not-reached items are defined in different ways. In TIMMS and PIRLS, omitted responses are defined as "the respondent had a chance to answer the question but did not do so, leaving the corresponding item or question blank" and not-reached items as the "items that student did not attempt due to a lack of time" (Martin, Mullis, & Hooper, 2016, p. 390-391; Martin, Mullis, & Hooper, 2017, p.9.11). Not-reached items are identified as following: First, the last answer given by an examinee is identified and then the first missing response after this last answer is treated as omitted, while all the following missing responses are treated as not-reached items. For instance, the response pattern "1 9 3 2 9 9 9" (e.g., "9" indicates missing responses) is recoded as "1 M 3 2 M N N" (e.g., "M" indicates omitted, while "N" indicates not-reached items).

In NAEP, omitted responses are defined as "a missing response prior to the last observed response" and not-reached items as "an item to which the student did not response because the time limit" (NCES, 2018). For instance, a single missing response at the end of the test is coded an omitted response (Ludlow & O'Leary, 1999). In other words, not-reached items are identified when there are at least two consecutive missing responses at the end of test.

In PISA, omitted and not-reached items are defined likewise as "students did not answer the given question but answered at least one subsequent question" and "students did not answer the given item nor the subsequent items within that cluster," respectively (OECD, 2017, p.133). In PIACC, omitted responses are defined as "any missing response followed by a valid response," whereas not-reached responses as "missing responses at the end of a block" (OECD, 2013, pp. 417-418). Overall, there is a consensus on definition and treatment of omitted items in LSAs, while the question of how many missing responses at the end of test indicate not-reached

items instead of omitted items has been a controversial issue (Rose, von Davier, & Nagengast, 2017).

Oftentimes, item nonresponses due to omission or time-constraints do not occur randomly (Mislevy & Wu, 1996); instead, they are correlated with examinee proficiency (Holman & Glas, 2005; Rose, Von Davier, & Xu, 2010), up to -0.45, indicating the more proficient a person, the smaller number of missing responses (Pohl, Gräfe, & Rose, 2014). To treat omitted responses appropriately, the missingness mechanism is modeled by an additional latent variable which represents the examinee's propensity to omit items (i.e., omission propensity) (Holman & Glas, 2005; Rose et al., 2010). For instance, the omission propensity can be included in the IRT model and computed like the examinee proficiency through jointly modeling responses and nonresponses. Further, the covariance of proficiency and omission propensity is computed for the population model. Likewise, Rose et al. (2010) reported a negative correlation (e.g., -0.33) between observed item responses and nonresponses (i.e., easier items are more likely to be answered than difficult items). Further, in the analysis of PISA data, relatively high correlations between proficiency and omission propensity are found across all countries, domains, and cycles which implies that missing data due to omission are nonignorable in all data sets (Sachse, Mahler, & Pohl, 2019). Through the empirical analysis, it is shown that item nonresponses depend on examinee proficiency and item characteristics; accordingly, it is not safe to assume that the missingness mechanism is ignorable.

The item nonresponses due to omissions and not-reached items result from different test-taking behaviors. On a timed test, for example, examinees may not reach the end of the test due to time limits (e.g., not-reached items). On the other hand, item omissions merely result from examinees' decision and may occur due to lack of motivation (Cosgrove, 2011; Köhler et al.,

2015a; Wise & Gao, 2017). The omitted and not-reached items differ in two important ways. First, not-reached items occur at the end of the test, whereas omitted items can occur anywhere in the test. Second, not-reached items can be considered as to be independent of item content and of the response that would be obtained if the item has been reached; however, omitted items occur when examinee has the opportunity to consider the item, but decides not to generate a response (Tijmstra, & Bolsinova, 2018). Since the process underlying omissions and not-reached items has different characteristics, they should be treated differently in IRT measurement models (Rose, 2013). Further, not adequately modeling the omitted and not-reached items may lead to biased parameter estimates in IRT (De Ayala, Plake, & Impara, 2001; Hohensinn & Kubinger, 2011; Lord, 1974; Mislevy & Wu, 1996) and considerably affects trend estimation, especially when omissions are scored as incorrect (Sachse et al., 2019).

Several methods have been proposed to treat omitted and not-reached items in LSAs. In the following sections, those methods are sorted by whether timing information is incorporated or not. Each method is described conceptually or its advantages and limitations are highlighted.

2.3.2 Omitted and Not-Reached Items, and RTs

The goal of testing is to provide valid scores from a test administration. To attain this goal, it is critical to have motivated examinees who actively and effortfully engage with test items. Test-taking effort is associated with RT (Schnipke & Scrams, 2002) and can be distinguished by two distinct test-taking behaviors: solution behavior and rapid guessing behavior. Examinees can show either solution behavior that they apply their knowledge, skills, and abilities to attempt the item, or rapid guessing behavior that they randomly guess the correct response without applying their effort (Wise, 2017). The empirical evidence supports this argument that the amount of time examinees spend on each behavior is considerably different

(Wise, 2017) and solution behavior and rapid guessing behavior have different accuracy rates (Schnipke, 1995). Test engagement can be explained as solution behavior, while disengaged behavior can be explained as rapid guessing (Schnipke, 1995; Wise, & Kong, 2005) and nonresponse behaviors (Ulitzsch et al., 2020).

The reason why examinees show rapid-guessing behavior varies by the different types of tests. For high-stakes tests, test scores may have important consequences for the examinees; thus, rapid guessing behavior can be explained as test speededness, that is, as examinees respond rapidly to items due to time limits, the accuracy will be at or near the chance level (Schnipke, 1995). In contrast, LSAs such as NAEP, PISA, PIAAC, TIMSS, and PIRLS are low-stakes tests that test scores have little or no consequences for examinees. Consequently, examinees can exhibit disengaged behavior due to lack of motivation (Köhler, Pohl, & Carstensen, 2015a; Wise & Gao, 2017).

More specifically, a rapid guessing can be explained by three different scenarios (Wise, 2017). First, motivated examinees can exhibit rapid guessing as a strategic behavior due to the time limit during high-stakes tests. Second, unmotivated examinees may show rapid guessing as a random guessing behavior during low-stakes tests. Third, when attempting an item, examinees recognize that they don't have the required skills or knowledge to solve the problem and they may respond with a random guess. Under the first two scenarios, rapid guessing can be considered as uninformative; however, under the third scenario, it is informative on examinee proficiency (Wise, 2017). Likewise, nonresponse behavior can be explained by a similar fashion: lack of motivation vs. skill-related reasons. Since the presence of rapid guessing in test data may lead to biased item parameter estimates (Schnipke, 1999) and aggregated scores (Rios, Guo,

Mao, & Liu, 2017), it is critical to identify the level of disengagement for drawing valid inference on examinee proficiency.

Then, how can we identify disengaged behavior? The amount of time examinees spends responding to an item differs by individuals due to various factors, including proficiency level, reading speed, or motivation. However, the response process underlying disengaged behavior is different from that of engaged behavior and this distinction can be revealed by the associated RT distribution, which oftentimes results in bimodal frequency distributions (Schnipke, 1995; Wise & Kong, 2005). By analyzing the NAEP data, Lee and Jia (2014) show that rapid responses are uncorrelated to the examinee proficiency, whereas responses, made after a certain time period are positively correlated with proficiency. Further, Weeks, von Davier, and Yamamoto (2016) found that the RT distribution for nonresponse behavior is distinctive from that for engaged behavior across countries by using the PIAAC data. Hence, test engagement can be evaluated at individual item responses by using the RT information.

The rise of CBAs allows the collection of RT information. The RT, defined as the time an examinee spends on an item, has been used as a method to obtain information about mental activity for an extensive period of time, as long as the field of psychology itself (Schnipke, & Scrams, 2002). Recently, analysis of RT has gained increasing attention in educational measurement. For high-stakes tests, for example, RT information has been used to improve item selection method in computerized adaptive testing (van der Linden, 2008), to detect differential speededness (van der Linden, Scrams, & Schnipke, 1999), and to detect cheating between pairs of examinees (van der Linden, 2009b). For low-stakes tests, RT information has been used for monitoring examinee effort and motivation such as solution behavior index (Wise, 2006; Wise, Kingsbury, Thomason, & Kong, 2004) and RT effort (Wise, & Kong, 2005).

RT information can also be used to improve item parameter estimation in LSAs. RTs, for instance, can be used to detect data fabrication (i.e., unmotivated responses) by investigating the cases which RT may be too short or inconsistent with expected times across different countries (Yamamoto, & Lennon, 2018). By using the RT information, the problematic responses can be excluded and this data cleaning procedure can improve item parameter and proficiency estimation (Wise & DeMars, 2006). Likewise, RTs can provide meaningful information about item-level nonresponse behavior and improve item parameter and proficiency estimates (Pohl, Ulitzsch, & von Davier, 2019; Ulitzsch, von Davier, & Pohl, 2019; Ulitzsch, von Davier, & Pohl, 2020).

Further, RT information can be included in the population model as additional covariates to enhance the modeling of group-level proficiency distributions in LSAs. The RT information needs to be incorporated in the population model for the following two reasons. First, given a substantial relationship between RTs and proficiency, ignoring the RT information in the population model can result in biased estimates of correlations between proficiency estimates and RTs in secondary analyses (von Davier et al., 2019). Secondly, RTs can help to classify examinees into groups that can be associated with test-taking strategies and motivation (Lee & Jia, 2014; Weeks, von Davier, & Yamamoto, 2016). To examine whether the RT data are comparable across countries, von Davier et al. (2019) analyzed PISA 2015 data and found that the item-level RT distribution in each domain appears similar across countries, suggesting that data cleaning and data quality analyses need to be conducted at the country level, instead of aggregate-level of all countries. Building upon the support of incorporating RTs in population modeling, more research is still needed on how to include RT information in the conditioning model (von Davier et al., 2019).

Detecting disengaged behavior is critical because it indicates the presence of item responses that are uninformative about an examinee's proficiency level; besides, rapid guessing leads to a negative bias of proficiency estimates because the correct rate of item responses for rapid guessing is substantially lower than that for solution behavior (Wise & Ma, 2012). To address this problem, several methods have been developed to identify disengaged behavior (i.e., rapid-guessing, unmotivated responses) by using the RTs.

2.4 Ad-Hoc Methods for Omitted and Not-Reached Items

This section discusses the treatment of omitted and not-reached items in LSAs, when the item and person parameters are estimated. The ad-hoc methods can be broadly differentiated by ignoring or incorporating timing information. Estimation then proceeds using typical marginal maximum likelihood (MML) estimation (e.g., Bock & Aitkin, 1981).

2.4.1 Methods that Ignore Timing Information

There are three ad hoc methods: partially-correct scoring, score as incorrect, or ignore. These ad hoc methods are commonly used to treat omitted and not-reached items in LSA. In NAEP, the omitted responses to multiple-choice items are scored as partially correct (i.e., reciprocal of the number of response alternatives) throughout the analysis and if the item is not a multiple-choice, the omitted response is scored as the lowest response category, whereas the not-reached items are ignored for both item and person parameter estimation (NCES, 2018).

On the other hand, in TIMSS (Martin, Mullis, & Hooper, 2016, p.13.12) and PIRLS (Martin, Mullis, & Hooper, 2017, p.12.7), omitted responses are coded as incorrect throughout the analysis, but not reached items are ignored for item parameter estimation, but coded as incorrect for person parameter estimation.

Similarly, in PISA, the omitted responses are scored as incorrect and the not-reached items are ignored for item parameter estimation. However, not-reached items are accounted for a covariate in the latent regression model as a part of the proficiency estimation in the generation of plausible values (OECD, 2017).

2.4.2 Methods that Incorporate Timing Information

There are several ad hoc methods that incorporate timing information: 5-second rule, visual inspection method, normative threshold method, and combining RT and response accuracy method. First, a 5-second rule uses a common time threshold (5 seconds) for all items. It is a special case of the common time threshold method, which uses a common time threshold (usually 3-5 seconds) for all items (OECD, 2013; Wise, Kingsburry, Thomason, & Kong, 2004) The advantage of this method is its simplicity, as it does not require any information about RT distribution or item's surface features. PIAAC, for example, uses a common time threshold method (e.g., 5-second rule) which is illustrated as a red vertical line in Figure 1. That is, if examinees spend more than or equal to five seconds on an item, the missing response is treated as incorrect, while examinees spend less than 5 seconds, the missing response is ignored (Yamamoto, Khorramdel, & von Davier, 2016). However, this method often produces variation in classification errors across items as RT distribution typically varies by items (Goldhammer, Martens, Christoph, & Lüdtke, 2016).

Second, a visual inspection (VI) method inspects the RT distribution to identify a threshold-gap which separates two distinct behaviors (i.e., engaged and disengaged behavior) for each item (Schnipke, 1995; Wise, 2006). In Figure 1, for instance, such a gap occurs at around 7 seconds. The VI method is congruent with the theoretical conceptualization of rapid guessing

and solution behavior; however, observed RT distribution is not always bimodal, especially, when item is easy that solution behavior occurs at relatively short period of time (Wise, 2017).

Third, normative threshold (NT) method (Wise & Ma, 2012) examines the mean RT for each item and then evaluates different percentage values (e.g., NT10, NT15, NT20) to find a threshold value which may reflect random guessing, up to a maximum threshold value of 10 seconds. For example, if examinees take 50 seconds on average to respond to a particular item, a 10 percent threshold (NT10) would be 5 seconds, while a NT15 would be 7.5 seconds. In particular, Wise and Ma (2012) recommend for using the NT10 method (i.e., 10% of the mean RT for each item) for a computer adaptive test.

Lastly, combining RT and response accuracy method extends the VI method by incorporating the response accuracy information. The previous research show that when examinees exhibit rapid guessing behavior, the correct rate of item responses is expected to be similar to that produced by random guessing (Lee & Jia, 2012; Schnipke, & Scrams, 2002; Wise & Ma, 2012). Bringing all together, this method searches for the point at which accuracy exceeds what would be expected from random guessing-the reciprocal of the number of response options (Goldhammer et al., 2016; Guo et al., 2016; Lee & Jia, 2012). For instance, this value would be 0.2 with five response options. The disadvantage of this method, however, is that it requires significant amount of response data per item to accurately identify the increase in accuracy needed to find the threshold (Wise, 2017).

Recently, Weeks et al. (2016) explored how missing values can be evaluated with RTs by using a logistic regression. However, they only reported the quantile values for expected probability levels without suggesting a guideline for setting a threshold value. To reduce classification errors, it is critical to identify a reliable time threshold value for each item which

can accurately detect disengaged item responses. However, the sparse observations in the short time intervals and fluctuation of examinee accuracy across the RT range make it difficult to determine the threshold values which oftentimes lead to the subjective choice (Guo et al., 2016). Even though various threshold identification methods have been proposed, none of the methods is flawless. Hence, researchers need to be well aware of the limitations of each method and apply them in practice.

2.5 Model-Based Methods for Omitted and Not-Reached Items

There is a growing body of literature on the model-based approaches for dealing with omitted and not-reached items in IRT models. In the model-based approaches, the nonignorable missing propensity for omitted and not-reached items (Mislevy & Wu, 1996) is accounted for item and person parameter estimates. The performance of such models depends on the appropriate assumptions of missing mechanisms; however, due to the untestable assumptions that underlie MNAR mechanism, there is no valid method to select the most appropriate model (Little & Rubin, 2002). Hence, to address the nonignorable missing mechanism caused by omitted and not-reached items, a model choice should be based on theories and sensitivity analyses need to be conducted for exploring the impact of different assumptions on the item or person parameter estimation.

Several models have been developed to treat omitted and not-reached responses in IRT models. Typically, the missing propensity is included either via models that ignore timing information (e.g., Holman, & Glas, 2005; O'muircheartaigh, & Moustaki, 1999; Rose et al., 2010) or by models that incorporate timing information (Pohl, Ulitzsch, & von Davier, 2019; Ulitzsch, von Davier, & Pohl, 2019; Ulitzsch, von Davior & Pohl, 2020). In the following sections, each method is described conceptually by highlighting its advantages and limitations.

2.5.1 Models That Ignore Timing Information

2.5.1.1 Latent approach

In social science, the survey literature addresses the concern about the nonresponses that may lead to biased parameter estimates. To deal with this problem, O'muircheartaigh and Moustaki (1999) attempt to obtain the information about the latent variable from nonresponses, by fitting the extended two-dimension factor model (Albanese & Knott, 1992). They assume that in the survey analysis there are two dimensions: an attitude dimension and a second dimension of response propensity, which underlies the respondent's response decision for each item. For modeling responses and nonresponses, two parallel matrices of binary data are created and a mixed model is fitted to handle the nonresponses.

In line with this research, Holman, and Glas (2005) propose a model-based approach for handling the omitted and not-reached items by using IRT models. In this approach, the extended version of generalized partial credit model (GPCM) is fitted to include more latent traits and this approach can access the extent to which the missing data are nonignorable from the factor loadings of the probability of missingness or observed responses on latent traits. In simulation studies, it is shown that ignoring the missing data mechanism leads to substantial bias in the item parameter estimates and further this bias increases as a function of the correlation between the proficiency and the latent variable governing the missing data process (Holman, and Glas, 2005).

In summary, in latent approach, the missing tendency is included via a latent missing propensity by fitting a multidimensional IRT model, which is depicted in Figure 2, where Y_i indicates observed responses on the test items, D_i represents missing indicators, θ indicates an examinee's latent proficiency and ξ represents the latent missing propensity. There are several limitations on latent approaches. First, this model makes an assumption that missing indicators

fit a unidimensional measurement model. Since the items are not constructed that the missing indicators meet a unidimensional IRT model, this assumption may not be negligible (Pohl, Gräfe, & Rose, 2014). Second, if a possible multidimensionality of the latent missing propensity is not accounted for, the latent approach may fail to address the omitted and not-reached responses (Rose, 2013). Lastly, the latent approach may result in estimation problems, when the sample size and the proportion of missing responses are small (Rose, 2013).

2.5.1.2 Manifest approach

To address the estimation problems of latent approach, the manifest approach (Rose et al., 2010) is proposed. As shown in Figure 3, the average number of missing responses (i.e., $\bar{D}$) is included in the measurement model. Compared to the latent approach, the manifest approach is easier to implement and there are fewer estimation problems.

Rose et al. (2010) also compare the performance of model-based approaches (e.g., between and within MIRT models, manifest approach) and ad hoc methods (e.g., IRT model that ignores the missing data and that treats omissions always as wrong). In the between and the within MIRT models, a second latent trait (i.e., response propensity) is incorporated to capture the nonresponse information, while the latent regression based on missing data model uses a predictor based on the observed count of omitted responses to improve the proficiency estimation. Findings from the simulation studies show that model-based approaches are equally appropriate to account for omitted responses; however, the simple IRT model that ignores the omissions also shows relatively good performance under the condition of moderate amounts of missing data (Rose et al., 2010).

There are several limitations on the manifest approach. First, it is implicitly assumed that there is a unidimensional missing propensity, which cannot be tested in the model (Pohl et al.,

2014). Further, if a fallible measure of missing propensity is included, this may distort the correlation and lead to a less-efficient bias reduction (Lord, 1960).

2.5.2 Models That Incorporate Timing Information

Examinee's responses on test items, as well as corresponding RTs, reveal important information about proficiency. To account for a speed-accuracy tradeoff, van der Linden (2007) proposed a hierarchical framework. Hierarchical regression model is useful to incorporate predictors at different levels of variation (Gelman et al., 2013). For instance, the achievement test may include information about individual students (e.g., math and verbal scores), class-level information (e.g., characteristics of teachers), and school-level information (e.g., types of schools). With predictors at multiple levels, the classical regression is extended to introduce as predictors a set of indicator variables for each of the higher-level units in the data (Gelman et al., 2013).

Most recently, a hierarchical framework for modeling speed and accuracy (van der Linden, 2007) is applied to account for the not-reached items (Pohl, Ulitzsch, & von Davier, 2019), omitted responses (Ulitzsch, von Davier, & Pohl, 2020), and disengaged behavior (e.g., guessing and omission) (Ulitzsch, von Davior & Pohl, 2019). Instead of using ad hoc methods (e.g., ignoring, scoring as wrong, common time threshold method), these models can simultaneously account for either the omitted or not-reached items in the estimation of item parameter and proficiency.

In this section, we introduce the speed and accuracy (SA) model (van Der Linden, 2007), SA+Engagement (SA+E) model (Ulitzsch et al., 2019), and SA+Omission (SA+O) model (Ulitzsch et al., 2020).

2.5.2.1 Speed-Accuracy (SA) Model (van der Linden, 2007)

The concept of a speed-accuracy tradeoff is based on the observation that examinees need to choose between working faster with lower accuracy or working slower with higher accuracy, while taking the test. This notion is motivated by the fact that examinees have control of their working speed and have to accept the accuracy, followed by the choice of speed (van der Linden, 2009a). Typically, speed is negatively (nonlinear) correlated with accuracy and a speed-accuracy tradeoff is a within-person relationship (van der Linden, 2007).

In educational measurement, the main interest is generally in measuring an examinee's (latent) proficiency rather than (manifest) accuracy; hence, IRT model is commonly used for capturing the "effective proficiency" of examinee that does not necessarily match the proficiency level that the test intended to measure (i.e., "target proficiency") (Tijmstra & Bolsinova, 2018). For instance, when an examinee has low motivation, the effective proficiency can be lower than the target proficiency of the test. Due to the speed-accuracy trade-off phenomenon, the effective proficiency is likely to be influenced by the speed level while working on items.

To formulate this trade-off, it is useful to consider the hierarchical modeling framework that jointly models effective speed and effective proficiency (van der Linden, 2007). Further, different levels for person parameters need to be specified: the fixed-person level (e.g., the parameters remain constant), and the random-person level (e.g., there is a distribution of parameter values across persons) (van Der Linden, 2007). The hierarchical framework can incorporate the three distinctive levels and a structure for simultaneously modeling item responses and RTs. For example, the measurement models for item response and RTs are specified on the first level, while the joint distribution of person and item parameters is specified on the second level.

The assumptions of hierarchical modeling framework are as follows: RTs follow a lognormal distribution (van der Linden, 2007) and an examinee's effective speed and effective proficiency are constant throughout the test (van der Linden, 2009). Further, it is noteworthy to mention that the correlation between examinee proficiency and speed on the second level concerns the between-person association. As a result, it is possible that more capable examinees who choose to work at a higher speed than the average result in a positive correlation between speed and proficiency (van der Linden, 2007), even though the within-person association between effective speed and effective proficiency can generally be assumed to be negatively correlated (van der Linden, 2009).

Item responses are modeled as in Section 2.1 and let T_{ij} be the response-time to the *i*th item. For the RTs T_{ij}, a lognormal model is assumed:

$$\ln T_{ij} = \beta_i - \tau_j + \varepsilon_{tij}, \qquad \varepsilon_{tij} \sim N(0, \alpha_i^{-2}) \tag{2.5.1}$$

where β_i indicates the time intensity for item *i*, τ_j indicates the speed parameter of examinee *j*, and α_i indicates a time discrimination parameter (i.e., the reciprocal of the standard deviation of the RTs on item *i*). When α_i is large, the proportion of the RT variance due to the differences in speed across examinees also becomes large.

On the second level, a dependency of speed and accuracy across examinees is estimated by a joint distribution of these random effects by allowing for a correlation between proficiency and speed (See Figure 5). The SA model is an extension of IRT model (See Figure 4) which allows a dependency between speed and proficiency at both item and population levels. For population model, a vector of latent person parameters, λ_j are randomly drawn from a multivariate normal distribution:

$$\lambda_j \sim f(\lambda_j|\mu_P, \Sigma_P) \tag{2.5.2}$$

with corresponding mean vector

$$\mu_P = (\mu_\theta, \mu_\tau) \tag{2.5.3}$$

and covariance matrix

$$\Sigma_P = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{2.5.4}$$

For SA+O model, λ_j will be expanded to include additional person parameters.

For item domain model, the vector of item parameters, φ_i are randomly drawn from a multivariate normal distribution:

$$\varphi_i \sim f(\varphi_i|\mu_I, \Sigma_I) \tag{2.5.5}$$

with mean vector

$$\mu_I = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta) \tag{2.5.6}$$

and covariance matrix

$$\Sigma_I = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix}. \tag{2.5.7}$$

Lastly, a joint sampling distribution for the observed item responses and RTs, conditional on all of the item and person parameters is defined as:

$$L = \prod_{j=1}^{N} \prod_{i=1}^{I} f(Y_{ij}, T_{ij} \,|\lambda_j, \varphi_i) f\left(\lambda_j \middle| \mu_P, \Sigma_P\right) f(\varphi_i|\mu_I, \Sigma_I). \tag{2.5.8}$$

There are several assumptions in the SA model (van Der Linden, 2007):

1) Examinees operate at constant accuracy and speed across the test.

2) For each examinee, both the item response and RT are random variables.
3) Item responses and RTs are conditionally independent between different items.
4) Item responses and RTs on the same items are conditionally independent.

Overall, the item responses and RT distributions have separate sets of parameters in the first level and the only constraint on them is the shape of their distributions in the population of examinees and the domain of items in the second level. One advantage of this model is simultaneously but separately modeling item responses and RTs by using a hierarchical structure.

Recently, Pohl et al. (2019) investigated whether the speed-accuracy (SA) model (van der Linden, 2007) can account for the not-reached items due to time limits by making an assumption of the propensity of not-reached items as a working speed. They also show the close association between the SA model and the manifest missing approach for not-reached items (Rose, von Davier, & Xu, 2010), for which speed is indicated by the RTs per item and the number of not-reached items, respectively. The number of not-reached items can be viewed as a rough approximation of RT at the test level. For data analysis, Bayesian estimation with Gibbs sampling is used and missing values are imputed based on the specified model. If not-reached items occur due to the different speed levels of examinees, the SA model can account for missing data process. Further, the SA model can estimate effective proficiency, which results in target proficiency for both examinees with and without missing values (Pohl et al., 2019).

2.5.2.2 SA+Omission Model (Ulitzsch et al., 2020)

By adopting a hierarchical framework (van der Linden, 2007), Ulitzsch et al. (2020) introduce a joint modeling of response and nonresponse behavior (i.e., SA+O model). The SA+O model (See Figure 6) is an extension of the SA model (See Figure 5) in that it allows a

dependency of speed and proficiency across examinees. However, the dependency on item level is reduced; instead, it includes parallel structures for observed and missing responses and allows simultaneously modeling item responses and RTs.

The SA+O model has several advantages. First, the SA+O model provides the information on nonresponse behavior by accounting for the degree of nonignorability of missing values in item responses, RTs and nonresponse times (NRTs) (Ulitzsch et al., 2020). For instance, a latent omission propensity (Holman & Glas, 2005) as well as an omission speed factor are accounted by joint modeling of response behavior. Second, the SA+O model can provide insights on test-taking strategies by allowing examinees to operate differential speed levels for generating observed and missing responses. In other words, it provides a better understanding of examinee's test-taking behavior by taking into account response and nonresponse speed variables (e.g., assessing whether examinees use different pacing strategies) when they generate engaged or disengaged behavior. Given the empirical data analysis, observed RTs of two different psychometric properties-observed and omitted responses had distinctive distributions within the same item (Weeks et al., 2016). Thus, estimating the correlation between speed and omission speed can provide valuable information on how response processes are related to item omission processes (Ulitzsch et al., 2020). Lastly, given the SA+O model as the data-generating model, modeling nonresponse behavior jointly with response behavior results in less biased person parameter estimates (Ulitzsch et al., 2020), compared to the SA model (van der Linden, 2007).

Despite the advantages of the SA+O model, there are several limitations. First, previous studies show that model-based approaches for handling missing responses affect proficiency parameter estimates only under the conditions with a large proportion of item nonresponses and a

high degree of nonignorability (Holman & Glas, 2005; Pohl et al., 2014; Rose et al., 2010; Rose et al., 2017). Therefore, given the complexity of the SA+O model, it is recommended for use under conditions with a large sample size (N ≥ 750) or a high omission rate (≥ 17%) for small sample sizes (Ulitzsch et al., 2020). Second, the SA+O model assumes stationarity of proficiency, speed, omission propensity, and omission speed. This assumption might be violated when examinees increase their working speeds to finish the test on time in speeded tests. Further, the SA+O model assumes RT and NRT distributions to be lognormal; however, RT distribution differed dramatically across items within one test (Ranger & Kuhn, 2012). The violation of RT and NRT distribution assumptions might lead to biased item and person parameter estimates.

2.5.2.2.1 Modeling response behavior

For the dichotomous item responses Y_{ij}, the two-parameter logistic model (i.e., 2PL model) is assumed (See Equation 2.1.1). For the RTs T_{ij}, the lognormal model is assumed (See Equation 2.5.1).

2.5.2.2.2 Modeling nonresponse behavior

The missing data indicator is defined in Section 2.2. When the amount of omissions is small, the data for nonresponse behavior can be sparse; thus, the simplest IRT model (e.g., Rasch model) is assumed. The probability of item omission is modeled as a function of a latent omission propensity ξ_j and omission difficulty ν_i on item i:

$$P_i\left(D_{ij}=1|\xi_j,\nu_i\right)=\frac{e^{\left(\xi_j-\nu_i\right)}}{1+e^{\left(\xi_j-\nu_i\right)}}. \tag{2.5.9}$$

Likewise, for the nonRTs S_{ij}, the lognormal model is assumed as it is defined for the SA model (See Equation 2.5.1):

$$\ln S_{ij}=\ \delta_i-\zeta_j+\varepsilon_{tij}\,,\qquad \varepsilon_{tij}\sim N(0,\omega_i^{-2}) \tag{2.5.10}$$

where δ_i represents the omission time intensity for item i, ζ_j represents the omission speed that examinee j decides to omit items, and ω_i represents an omission time discrimination parameter.

Unlike the SA model (van der Linden, 2007), the first-level item parameters are assumed to be fixed effects, while person parameters are modeled as random effects. Hence, the vector of latent person parameters, λ_j are randomly drawn from a multivariate normal distribution:

$$\lambda_j \sim f(\lambda_j|\mu_P, \Sigma_P) \tag{2.5.11}$$

with mean vector

$$\mu_P = (\mu_\theta, \mu_\tau, \mu_\xi, \mu_\zeta) \tag{2.5.12}$$

and covariance matrix

$$\Sigma_P = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} & \sigma_{\theta\xi} & \sigma_{\theta\zeta} \\ \sigma_{\tau\theta} & \sigma_\tau^2 & \sigma_{\tau\xi} & \sigma_{\tau\zeta} \\ \sigma_{\xi\theta} & \sigma_{\xi\tau} & \sigma_\xi^2 & \sigma_{\xi\zeta} \\ \sigma_{\zeta\theta} & \sigma_{\zeta\tau} & \sigma_{\zeta\xi} & \sigma_\zeta^2 \end{pmatrix}. \tag{2.5.13}$$

Lastly, the likelihood function is defined as following:

$$L = \prod_{j=1}^{N}\prod_{i=1}^{I} P\left(Y_{ij}|\theta_j, a_i, b_i\right)^{1-D_{ij}} f\left(T_{ij}|\tau_j, \beta_i, \alpha_i\right)^{1-D_{ij}} \\ P(D_{ij}|\xi_j, \nu_i)^{D_{ij}} f(S_{ij}|\zeta_j, \delta_i, \omega_i)^{D_{ij}}\, g\left(\theta_j, \tau_j, \xi_j, \zeta_j|\mu_P, \Sigma_P\right). \tag{2.5.14}$$

In summary, the SA model can fit item responses and RTs simultaneously, but separately through a hierarchical structure: on the lower level, measurement models are specified for item response and RTs, while the dependencies between the item and person parameters are modeled on the higher level (van der Linden, 2007). Taking this advantage, the SA+O model extends the SA model by including a process model for nonresponse behavior, and it also allows examinees to operate on different speed levels for generating responses and omitted responses (Ulitzsch et al., 2020). Further, Ulitzsch et al. (2020) suggest a possibility of the SA+O model to account for

not-reached items, since this model controls for general working speed as well as nonresponse speed (Pohl et al., 2019).

In this research, the SA+O model (Ulitzsch et al., 2020) is used to account for omitted responses as well as corresponding NRTs. Further details of simulation conditions are presented in Chapter 3.

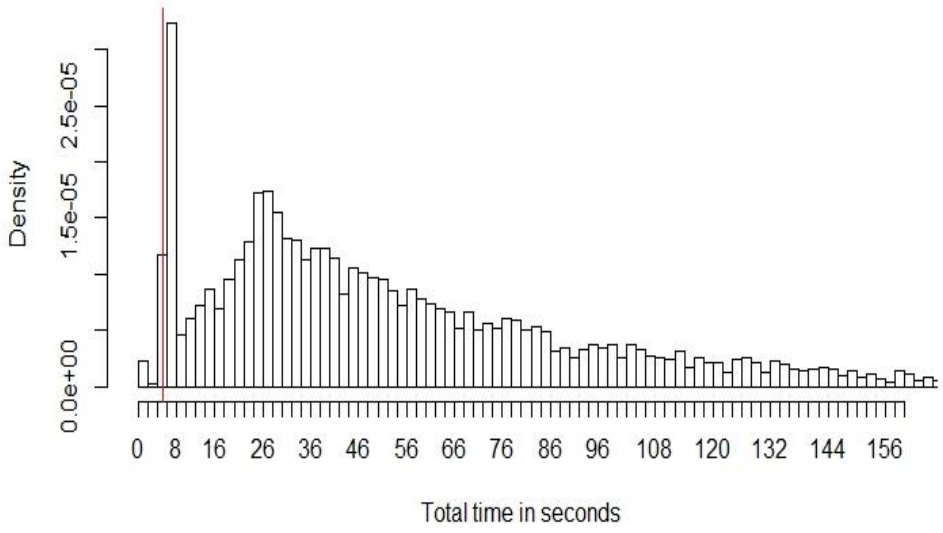


Figure 1: Response Time Distribution for a Single PIAAC Literacy Item

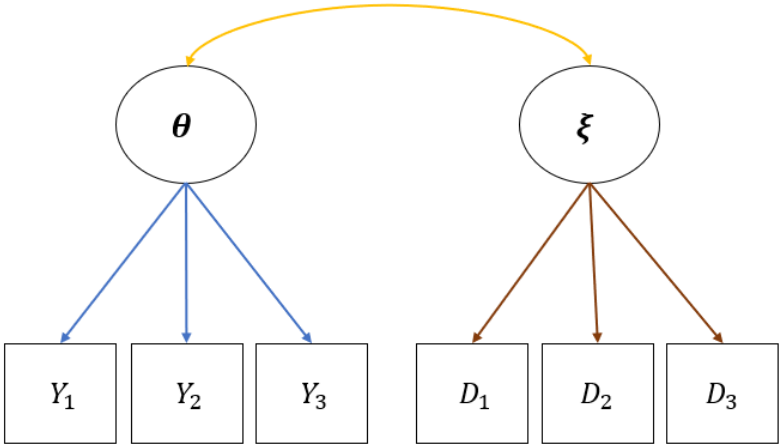


Figure 2: Latent approach: a multidimensional IRT model

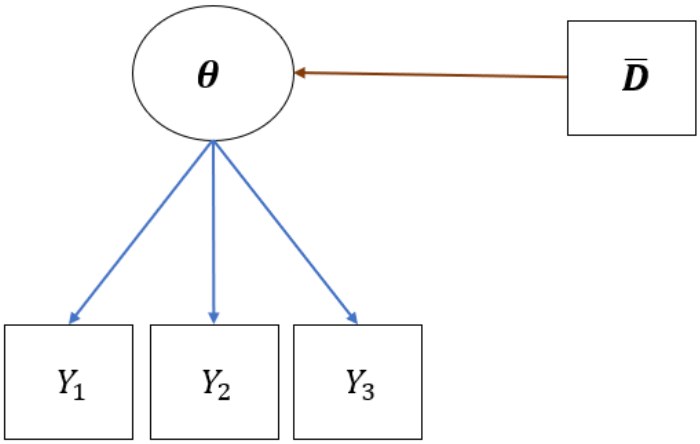


Figure 3: Manifest approach for modeling missing responses

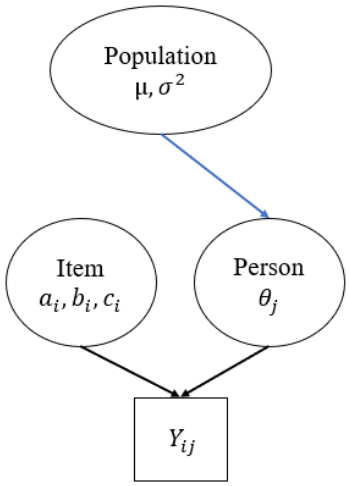


Figure 4: A Unidimensional Framework for IRT model

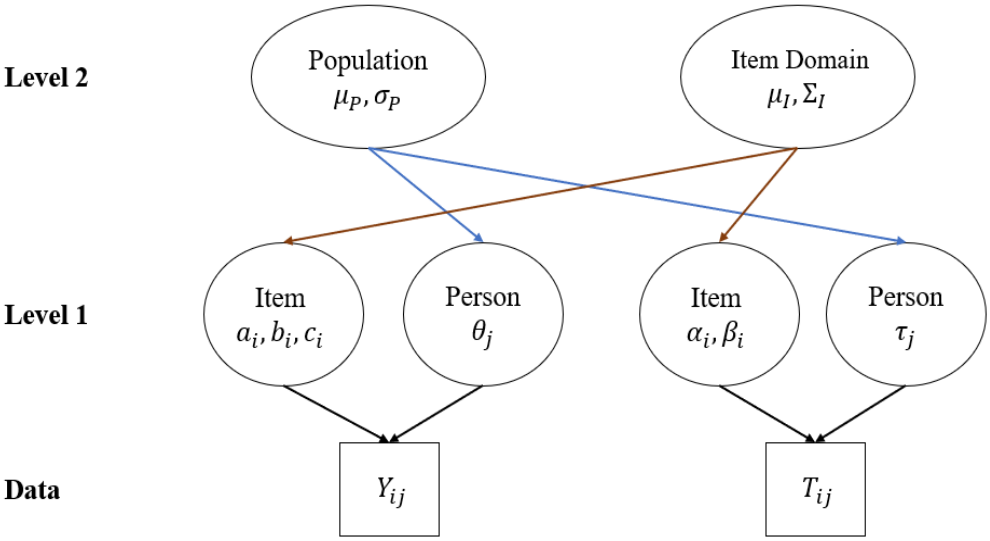


Figure 5: A Hierarchical Framework for SA model (van der Linden, 2007)

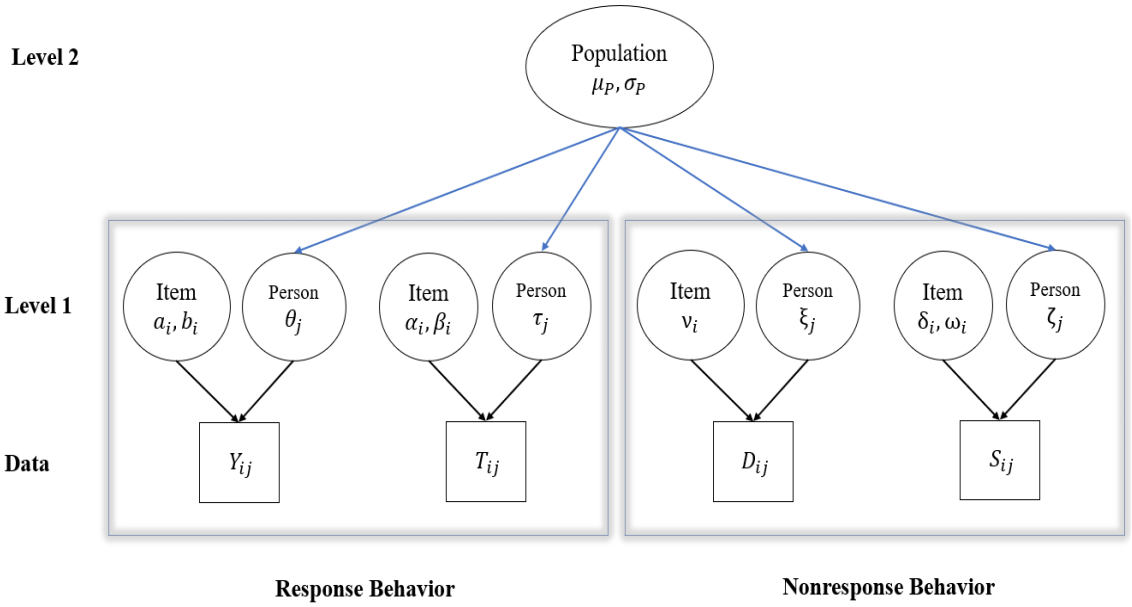


Figure 6: A Hierarchical SA+O Framework (Ulitzsch et al., 2020)

CHAPTER 3

METHODS

A Monte Carlo simulation study was designed to investigate the impact of omitted responses on item and person parameter estimates with the ad hoc and the model-based approaches under realistic conditions. The number of design factors and corresponding levels were chosen to address the research questions. However, more simulation studies can be conducted to provide insights for educational implications.

3.1 Simulation Design

To evaluate the impact of omitted responses on item and person parameter estimates with the ad hoc and the model-based approaches, not only the realistic, but also the challenging conditions were considered. The design factors were based on PIAAC 2012 Chile data to represent realistic conditions. Further, to investigate the performance of the SA+O model, unfavorable conditions (e.g., zero correlation structure between omission speed and ability, speed, and omission propensity) were also considered. The current study varied four different design factors with two levels each fully-crossed:

(1) Sample sizes (e.g., N = 375 and 750) which represent small and moderate sample sizes, respectively.

(2) Number of items (e.g., I = 10 and 30), resembling balanced incomplete block designs with planned missingness[1].

(3) Omission rates (e.g., O = 5% and 17%), which represent small and high proportion of missingness per item.

[1] For incomplete block designs, only a fraction of items is administered to each examinee. As a result, while the overall sample size is large, the number of examinees, assigned to each item can be small (Gonzalez & Rutkowski, 2010).

(4) Correlation structure between omission speed and proficiency, speed, and omission propensity, which is referred to as $\rho_{.\eta}$. The $\rho_{.\eta} \neq 0$ condition represents a realistic condition under which omission speed is correlated with proficiency, speed, and omission propensity, while $\rho_{.\eta} = 0$ condition represents the extremely unfavorable condition under which the correlation between omission speed and proficiency, speed, and omission propensity is equal to 0.

Regardless of whether $\rho_{.\eta} \neq 0$ or $\rho_{.\eta} = 0$, only the SA+O model is correctly specified. In the case where $\rho_{.\eta} = 0$, the SA+O model is more complex than the data-generating model, and it estimates additional parameters, and we can expect the parameter estimates to be relatively inefficient.

In total, there were 16 conditions. For each condition, 100 Monte Carlo replications were attempted. Each simulated dataset was calibrated twice: once using MML and the standard 2PL IRT model, and once using Bayesian estimation and the SA+O model, again with a 2PL IRT model.

3.2 Data Generation

For data generation, parameters that typically represent large scale assessments were chosen. To evaluate the performance of ad hoc and model-based approaches, realistic conditions which reflect empirical data (e.g., PIAAC) were considered as well as extremely unfavorable conditions to challenge the estimation (Ulitzsch et al., 2020). The data-generating model was the SA+O model.

Person parameters. For each condition, person parameters were randomly drawn from a multivariate normal distribution with mean vector and covariance matrix. The definitions of

generating densities are given in Table 1. For model identification, the expectations of all person parameters are fixed to zero and the variance of proficiency (e.g., $var(\theta)$) is fixed to unity.

Item parameters. The generating item parameters were fixed across replications. To generate response and omission indicators, 2PL and Rasch models were used respectively. Table 2 presents the item parameter values for all conditions. For all item parameter types, five different values were considered (Ulitzsch et al., 2020), and for test lengths of 10 items and 30 items, each item sequence was repeated two and six times, respectively.

3.3 ML Estimation Procedures and Ad Hoc Approaches

Ad hoc approaches are commonly used to deal with omitted responses in LSAs. As discussed in Chapter 2, omitted responses are scored as incorrect (e.g., TIMSS, PIRLS, and PISA), and ignored or incorrect based on a five-seconds rule (PIAAC) for item parameter estimation. For person parameter estimation, omitted responses are scored as incorrect (TIMSS and PIRLS) or ignored (PISA and PIAAC). For NAEP, the omitted responses are scored as partially correct throughout the analysis.

More recently, the model-based approach for handling omitted items, SA+O model (Ulitzsch et al., 2020) has been proposed with the evidence of reliable item and person parameter recovery. However, a practical difference between ad hoc and model-based approaches to handling omitted responses has not been evaluated. Thus, a comparison of ad hoc and model-based approaches for handling omitted responses in terms of item and person parameter estimation in IRT is of primary interest in simulation study.

The procedure for investigating the impact of omitted responses on item and person parameter estimates with the ad hoc and the model-based approaches is as follows:

(1) Under the two-parameter logistic (2PL) model and Rasch model, simulate item responses and omission indicators, respectively, using the known item parameters.

(2) Obtain item parameter estimates with the ad hoc approaches (e.g., incorrect, ignored, five seconds rule)[2] by fitting 2PL model, denoted as $\widehat{K}_a$.

(3) Obtain EAP score estimates, $\tilde{\theta}_a$, using corresponding item parameter estimates from each of ad hoc approaches.

(4) Obtain item parameter and person parameter estimates with the SA+O model, denoted as $\widehat{K}_m$ and $\tilde{\theta}_m$, respectively.

(5) Compare the bias and root mean square error (RMSE) for item parameter estimates.

(6) Compare the bias and mean absolute error (MAE) of the posterior distribution for proficiency estimates.

Because for all conditions, $cov(\theta, \xi) = -1.66$, we can anticipate certain biases in the ad hoc approaches. The *ignored* approach is always going to have negative bias in the difficulty parameter estimates, because it is wrongly treating omission as unrelated to proficiency. The *incorrect* approach is always going to have positive bias in the difficulty parameter estimates, because it is wrongly treating omissions as always due to low proficiency. On the other hand, it is harder to anticipate what will occur with the 5-second rule. Further, it is also harder to anticipate how estimation of the discrimination parameter may be affected.

3.4 Bayesian Estimation Procedures for SA+O Model

3.4.1 Prior Specification

[2] For incorrect and ignored approach, missing values are scored as incorrect and ignored, respectively. For five seconds rule, if examinee spends more than or equal to five seconds on an item, the missing value is scored as incorrect, whereas it is scored as ignored.

The SA+O model assumes a multivariate normal distribution for person parameters. For this study, we follow the prior specifications used in Ulitzsch et al. (2020). To estimate variance-covaraince matrix Σ_p, Bayesian estimation requires a prior setting. In Bayesian statistics, noninformative priors are a common choice (Gelman & Hill, 2007; Fox, 2010), especially, when there is the lack of interpretability of the parameters and the data to be posterior dominant for the quantity of interest. Further, the conjugate prior is a popular choice because due to the conjugacy properties, the posterior distribution follows the same distribution as the prior - the conjugate prior for the multivariate normal distribution is the inverse Wishart (IW) distribution (Barnard, McCulloch, & Meng, 2000).

However, there are some disadvantages for using the IW prior for variance-covariance matrix. First, the marginal distribution for the variances has low density when it is close to zero (Gelman, 2006). Second, the uncertainty for all variances is controlled by a single degree of freedom parameter (Gelman et al., 2013). Lastly, there is a priori dependence between correlations and variances (Tokuda et al., 2011). All these can affect posterior inferences about the variance-covariance matrix. Further, the IW prior tends to be informative about variances, especially, when the sample size is small (Alvarez, Niemi, & Simpson, 2014).

To deal with these disadvantages, several covariance matrix priors haven been proposed as an alternative to the IW prior, including the scaled IW (O'Malley, & Zaslavsky, 2008), a hierarchical IW (Huang, & Wand, 2013), and a separation strategy (Bernard et al., 2000). These prior choices have yielded unbiased variance-covariance estimates even with a small sample size, and a separation strategy showed the most flexibility on a priori dependence between the correlations and the variances (Alvarez et al., 2014).

In a separation strategy, the standard deviations and correlations are modeled independently (Bernard et al., 2000). As a result, variances are less dependent on correlations, and this results in modeling flexibility and desirable inference properties. The person parameter variance-covariance matrix can be decomposed as following:

$$\Sigma_p = \Lambda_p \Omega_p \Lambda_p \tag{3.4.1}$$

where Λ_p is a diagonal matrix of person parameter standard deviations and Ω_p is a person parameter correlation matrix[3].

One disadvantage of a separation strategy is its computational complexity. Fortunately, with a development of the Hamiltonian Monte Carlo (HMC)[4] sampler, it is possible to use a separation strategy. In the Stan manual (Stan Development Team, 2019), it is recommended to use a separation strategy along with the LKJ prior (Lewandowski, Kurowicka, & Joe, 2009) which can be used to control the expected amount of correlation among the parameters. With shape 1 for the correlation matrix Ω_p, it suggests a uniform distribution on the correlation parameters and half Cauchy priors with location 0 and scale 25 (Gelman & Hill, 2007) for each freely estimated element of Λ_p, including standard deviations of speed τ, omission propensity ξ, and omission speed ζ.

To estimate item parameters, the noninformative priors are applied based on the previous studies. For the item difficulty *b*, time intensity β, omission difficulty ν and omission time intensity δ parameters, noninformative normal priors with mean zero and standard deviation 100 are employed (Fox, 2010), while half normal prior with mean 0 and standard deviation 100 is

[3] The correlation matrix is constructed from an IW distribution. Let Q ~ $IW(\nu, I)$, then $\Omega = \Delta Q \Delta$ where Δ is a diagonal matrix with i^{th} diagonal element $Q_{ii}^{-1/2}$ (Bernard et al., 2000).
[4] HMC is efficient where parameters are correlated in the posterior (Stan Development Team, 2019).

employed for the item discrimination parameter *a* due to its restriction to be positive values (Fox & Marianti, 2016). For inverse time discrimination α and omission time discrimination ω parameters, diffuse half Cauchy priors with location 0 and scale 25 are applied (Gelman & Hill, 2007).

3.4.2 Implementation Details

For ad hoc approaches, the 2PL model was used for analyzing the datasets with the mirt package (Chalmers, 2012) via R version 3.6.0 (R Development Core Team, 2019). For item parameter estimation, the expectation-maximization (EM) algorithm (Bock & Aitkin, 1981) was employed by using fixed Gauss-Hermite quadrature which was appropriate for lower dimensional models. For person parameter estimation, EAP score estimates were obtained.

For model-based approach, more advanced Bayesian estimation was employed with the Stan (Carpenter et al., 2017) which uses the No-U-Turn sampler (NUTS; Hoffman & Gelman, 2014), an extension of MCMC algorithm. Advantages of NUTS are as follows: first, it does not require tuning parameters and efficiently samples from posterior distributions with correlated parameters. As a result, it performs better for more complex models with correlation dimensions, compared with a Metropolis-Hastings algorithm (Annis et al., 2017). Stan code for the SA+O model is provided by Ulitzsch et al. (2020).

For parameter estimation, two MCMC chains were used with 10,000 iterations for each and the first 5,000 iterations were discarded as burn-in. Convergence was evaluated by the potential scale reduction factor (PSRF). Specifically, PSRF values below 1.10 were considered as acceptable (Gelman & Shirley, 2011). In other words, if a replication generates a PSRF values exceeding 1.10, it was regarded as not converged and that replication is not considered in further analyses (Ulitzsch et al., 2020).

3.5 Collected Statistics

Only converged replications were considered for further analysis. The accuracy and efficiency of item and person parameter estimates were examined by the median and 90% ranges of differences between estimated and true parameter estimates, bias and root mean square error (RMSE) or mean absolute error (MAE) of item and person parameter estimates.

The bias describes the discrepancies between the population and sample estimates. For a population parameter ζ and a corresponding sample estimate $\hat{\zeta}$, the bias is defined as

$$\text{bias}(\hat{\zeta}) = R^{-1} \sum_{r=1}^{R} (\hat{\zeta}_r - \zeta) \tag{3.5.1}$$

where R is the number of Monte Carlo replications.

The RMSE serves to aggregate information on errors and variabilities into a single measure. RMSE is the square root of the average of squared errors and is defined as

$$\text{RMSE}(\hat{\zeta}) = \sqrt{R^{-1} \sum_{r=1}^{R} (\hat{\zeta}_r - \zeta)^2} \tag{3.5.2}$$

RMSE value of 0 indicate a perfect fit to the data; thus, small RMSE values represent the estimates do not vary substantially across replications.

Similarly, MAE was calculated for person proficiency estimates. MAE is a mean absolute vertical or horizontal distance between each point and is defined as

$$\text{MAE}(\hat{\zeta}) = R^{-1} \sum_{r=1}^{R} |\hat{\zeta}_r - \zeta| \tag{3.5.3}$$

In other words, MAE is simply the mean absolute difference between sample estimates and true values. MAE is easier to interpret than RMSE. In the simulation study, bias and RMSE/MAE were computed for all item and person parameter estimates.

Table 1: Generating Densities

p	μ	$\Sigma\ (\rho_{\cdot\eta} \neq 0)$	$\Sigma\ (\rho_{\cdot\eta} = 0)$
4	$\begin{pmatrix}0\\0\\0\\0\end{pmatrix}$	$\begin{pmatrix}\theta & \tau & \xi & \eta\\ 1.00 & & & \\ -0.16 & 0.20 & & \\ -1.66 & 0.46 & 6.50 & \\ -0.49 & 0.32 & 2.37 & 1.20\end{pmatrix}$	$\begin{pmatrix}\theta & \tau & \xi & \eta\\ 1.00 & & & \\ -0.16 & 0.20 & & \\ -1.66 & 0.46 & 6.50 & \\ 0.00 & 0.00 & 0.00 & 1.20\end{pmatrix}$
		$\begin{pmatrix}\theta & \tau & \xi & \eta\\ 1.00 & & & \\ -0.35 & 1.00 & & \\ -0.65 & 0.40 & 1.00 & \\ -0.45 & 0.65 & 0.85 & 1.00\end{pmatrix}$	$\begin{pmatrix}\theta & \tau & \xi & \eta\\ 1.00 & & & \\ -0.35 & 1.00 & & \\ -0.65 & 0.40 & 1.00 & \\ 0.00 & 0.00 & 0.00 & 1.00\end{pmatrix}$

Note. p = dimensionality of density; μ = mean; Σ = variance; θ = person proficiency; τ = examinee speed; ξ = omission propensity; η = omission speed. $\rho_{\cdot\eta}$ denotes the correlation structure between omission speed and proficiency, speed, and omission propensity.

Table 2: Generating Item Parameter Values

Parameter Types	Omission rate	Item				
		1	2	3	4	5
Item discrimination (a)		0.75	1.125	1.5	1.875	2.25
Item difficulty (b)		-1	-0.5	0	0.5	1
Time intensity (β)		3.5	3.75	4	4.25	4.5
Omission time intensity (δ)		3.5	3.75	4	4.25	4.5
Omission difficulty (ν)	17%	2	2.5	3	3.5	4.
	5%	4.25	4.75	5.25	5.75	6.25
Time discrimination (α)		2	2.75	3.5	4.25	5
Omission time discrimination (ω)		1	1.625	2.25	2.875	3.5

Note. For omission difficulty, two sets of item parameter values are provided to generate small (5%) and moderate (17%) proportion of item nonresponses, respectively.

CHAPTER 4

RESULTS

The descriptive details on the generated data and convergence rate of the SA+O model will be presented first. Next, the results from the ad hoc approaches and the model-based approach will be compared in terms of medians, 90% ranges of differences between estimated and true parameter estimates, bias and RMSE/MAE of item (e.g., discrimination and difficulty) and person proficiency estimates. Lastly, the item and person parameter estimates from the SA+O model will also be presented to verify the implementation.

4.1 Descriptive Details on the Simulated Data

In the simulated data, the SA+O model generates RT and NRT distributions across all items. Figure 7 shows one example of RT and NRT distributions for a single item from a simulated data set with sample size of 750, test length of 30 items, omission rate of 17% and $\rho_{\cdot\eta} \neq 0$. On the left plot, overall RT distribution illustrates a bimodal distribution. This bimodal trend seems more prominent on the right plot where the dashed line represents NRT (i.e., RTs for omitted items) and solid line represents RT distribution. Across all conditions and replications, it is assumed that NRT distribution is located to the left and more peaked than RT distribution. In other words, examinees who omit an item moves to the next item much faster than those who exhibits solution behavior throughout the test.

The omission rates varied by items. For instance, item-level omission rate ranges from 1.5% to 11.2% under the conditions with low omission rate. On the other hand, under the conditions with high omission rate, item-level omission rate ranges from 7.6% to 29.6%. This item-level omission rate resembles the empirical findings. For instance, in 2012 PIAAC, the rate of omitted responses ranges from 2% for the numeracy domain in South Korea to 25.9% for the

literacy domain in Chile (OECD, 2013). Further, in PISA 2006, the proportion of omitted responses varies substantially from 1% in Netherlands to 16% in Kyrgyzstan and 0.3% in Azerbaijan to 13% in Colombia, respectively (OECD, 2009, p.220).

The results in Table 3 summarize the proportions of convergence for SA+O model out of 100 replications. The convergence rates for ad hoc approaches were 100% throughout all conditions. Overall, when omission rate is high, convergence rate for the SA+O model was at least 98%. However, convergence rate can be as low as 75%, when sample size is small (N=375), test length is short (I=10), omission rate is low (5%), and omission speed is correlated with other person parameters ($\rho_{.\eta} \neq 0$). Further, the average convergence rate under the conditions with a small omission rate was 87%, ranging from 75% to 96%, whereas that under the conditions with a large omission rate (17%) was 99%, ranging from 98% to 100%. In general, convergence rates under the conditions with omission speed uncorrelated with other person parameters ($\rho_{.\eta} = 0$) are higher than that under the conditions with $\rho_{.\eta} \neq 0$, because PSRF values which exceed 1.10 were oftentimes person parameter covariance estimates between omission propensity and omission speed. This result is in line with the previous study (Ulitzsch et al., 2020).

4.2 Bias and RMSE of Item Parameter Estimates

Table 4 through Table 6 present the bias, RMSE, SD and 95% coverage intervals for the item discrimination parameter estimates. Since there is no meaningful difference across $\rho_{.\eta}$ conditions for bias or RMSE results for the item discrimination or difficulty parameter estimates, only $\rho_{.\eta} \neq 0$ results are presented. The ignored approach tends to show downwards bias, especially under the conditions with high omission rates, while other approaches show upwards bias. And the RMSE under the ignored approach follows closely to that under the SA+O model

throughout all conditions. The bias and RMSE for item discrimination parameter estimates increase as item discrimination parameter value increases with the SA+O model. An interesting pattern is shown for the incorrect approach and 5-second rule. For instance, the bias decrease as item discrimination parameter value increases; however, under the small omission rate conditions (5%), the RMSE increases, while under the high omission rate (17%) conditions, the RMSE decreases as item discrimination parameter value increases. Following that, the difference on bias and RMSE between the SA+O model and incorrect approach and 5-second rule is the highest for the smallest item discrimination value (0.75) under the conditions with a high omission rate. For instance, the difference on RMSE between the SA+O model and incorrect approach and 5-second rule under the large sample size, longer test length and high omission rate condition is 0.34 and 0.23, respectively.

Table 6 presents that the average posterior SD for item discrimination parameter estimates increases as item discrimination parameter value increases for both ad hoc and model-based approaches throughout all conditions. Under the small sample size and short test length conditions, the SD for SA+O model is larger than that for ad hoc approaches and this difference is most prominent for the largest item discrimination parameter value. Further, as sample size increases, the SD decreases substantially. For instance, the difference of SD between the small sample (e.g., N=375, I=10, O=5) and large sample (e.g., N=750, I=10, O=5) is at most 0.15. This result is expected because when the sample size is smaller the sample size for that item is also smaller and the corresponding standard error will likely be larger.

Figures 8 depicts the medians and 90% ranges of item discrimination parameter estimates. $\rho_{.\eta} \neq 0$ represents the simulation conditions that omission speed is correlated with proficiency, speed, and omission propensity. N denotes number of examinees, I, number of

items, and O, omission rate, respectively. The x-axis indicates true item parameters and y-axis indicates the bias. Solid line denotes SA+O model, dashed line, incorrect, dotted line, ignored, and dotdash line, 5-second rule. Lastly, the grey dashed horizontal line indicates unbiased estimation.

First, when omission rate is high (17%), incorrect approach and 5-second rule show upwards bias, while ignored approach shows downwards bias. This trend is most noticeable for small item discrimination parameters but subsides as item discrimination parameter value increases (2.25). Further, the medians of 5-second rule are smaller than that for the incorrect approach; however, this trend again subsides as item discrimination parameter value increases. Lastly, as sample size and test length increase, the 90% ranges of item discrimination parameter estimates decrease for both SA+O model and ad hoc approaches.

Table 7 through Table 9 present the bias, RMSE, SD and 95% coverage intervals for the item difficulty parameter estimates. Throughout all conditions, the bias under SA+O model is close to zero. Taking account of the timing data, the SA+O model is simultaneously modeling response as well as nonresponse behavior. As a result, more variability is explained by the SA+O model, compared to the ad hoc approaches, and this leads to the lack of bias in the item difficulty parameter estimates for the SA+O model. On the other hand, ignored approach shows substantial downwards bias, while incorrect approach and 5-second rule show substantial upwards bias under a high omission rate. The RMSEs for incorrect approach and 5-second rule decrease as item difficulty values increase, especially under the conditions with a high omission rate. However, the RMSEs for ignored approach and SA+O model show the U-shaped pattern-that is, the RMSE is the highest for the smallest and largest item difficulty values (e.g., -1.0 and 1.0), but the RMSE is the lowest for the mid item difficulty value (0.0). Further, the difference on RMSEs

between the SA+O model and incorrect approach and 5-second rule is substantial under the conditions with a high omission rate. For instance, that difference ranges from 0.46 to 0.17 under the large sample size, longer test length and high omission rate condition.

Table 9 presents that the average posterior SD for item difficulty parameter estimates shows the U-shaped pattern-that is, the SD is the largest for the smallest and largest item difficulty values (e.g., -1.0 and 1.0), but the SD is the lowest for the mid item difficulty value (0.0) for both ad hoc and model-based approaches throughout all conditions. As it is shown in Table 6, as sample size increases, the SD decreases substantially. Further, the SD for SA+O model is smaller than that for ad hoc approaches when omission rate is high. In particular, the difference between ad hoc and model-based approach is largest for the smallest and largest item difficulty values and lowest for the mid item difficulty value. For instance, the difference between incorrect approach and SA+O model is at most 0.13 under the condition of N=750, I=30, O=17.

Figures 9 depicts the medians and 90% ranges of item difficulty parameter estimates. Item difficulty parameter *b* values were well recovered throughout all conditions with SA+O model without systematic bias and as sample size increases (N=750), the bias is reduced. Under the conditions with a high omission rate, incorrect approach and 5-second rule show upwards bias, while ignored approach shows downwards bias. In particular, the largest difference is shown for the smallest item difficulty parameter *b* value (-1) and as item difficulty value increases, the difference decrease. This pattern is consistent throughout the conditions with a high omission rate. Lastly, the medians for a 5-second rule are smaller than that for an incorrect approach throughout all conditions.

In summary, given the SA+O model is the data generating model, the SA+O model shows the smallest bias and lowest RMSE throughout all conditions, as it is correctly specified. The Ignored approach seems to outperform the incorrect and 5-second rule approaches for the item difficulty parameter, in terms of both bias and RMSE. The same seems to be true for the discrimination parameter. The incorrect and 5-second rule seem to perform similarly. According to the results, the SA+O model is recommended for use for item calibration when omission rate is high; however, when omission rate is small and test length is long, there is not much to be gained in terms of both bias and RMSE from using the SA+O model. Lastly, the Appendix contains tables and figures that present the bias and RMSE for the omission difficulty, time intensity, omission time intensity, time discrimination and omission time discrimination parameter estimates.

4.3 Bias and RMSE/MAE of Person Parameter Estimates

Table 10 and Table 11 present the bias and mean absolute error (MAE) for person proficiency estimates, respectively. The person proficiency estimates are divided into three groups, based on the magnitude of the data generating (i.e., true) proficiency values θ. EAPs estimated from ad hoc approaches and proficiency estimates from the SA+O model substantially underestimate positive θ, but overestimate negative θ values under the conditions with a short test length (I=10). This result may reflect the lack of examinees' response patterns at the proficiency extremes. Further, the ignored approach continuously underestimates central θ values to some degree under conditions with a high omission rate (17%).

MAE values indicate that the variability of the proficiency estimates is moderately similar for all fittings. There is almost no difference on MAE values among ad hoc approaches and the SA+O model with high performers ($\theta \geq 1$); however, that difference becomes prominent

with average ($-1 \leq \theta \leq 1$) and low performers ($\theta \leq -1$) under the conditions with a high omission rate. For instance, the range of that difference on MAE between ad hoc approaches and the SA+O model is from 0.02 to 0.09. In particular, the ignored approach shows the highest MAE values throughout all conditions with low performers, while MAE values for incorrect approach and 5-second rule follow close to each other.

Table 12 presents the average posterior SD and 95% coverage intervals for person proficiency estimates. The SD for positive and negative person proficiency (e.g., $\theta \geq 1$ and $\theta \leq -1$) is smaller than that for the central person proficiency (e.g., $-1 \leq \theta \leq 1$) for both ad hoc and model-based approaches. Again, this result may reflect the lack of examinees' response patterns at the proficiency extremes. As test length increases, the intervals for positive, negative and the central person proficiency become more balanced. Interestingly, under the condition with high omission rate, the SD for SA+O is smallest, especially for the negative person proficiency.

Table 13 and Table 14 present the bias and mean absolute error (MAE) for person proficiency estimates conditioning on true omission propensity, respectively. The true omission propensity values are divided into three groups: high ($\xi > 2.5$), central ($-2.5 \leq \xi \leq 2.5$) and low ($\xi < -2.5$) omission propensity. Under the high ξ conditions, EAPs estimated from incorrect and 5-second rule noticeably underestimate person proficiency estimates θ, while that from ignored approach overestimates θ, when test length is long and omission rate is high. Similarly, MAE values also indicate that the difference among ad hoc approaches and the SA+O model becomes prominent with high omission propensity performers ($\xi > 2.5$) under the conditions with a longer test length and high omission rate. For instance, the range of that difference on MAE between ad hoc approaches and the SA+O model is from 0.06 to 0.11. In short, when omitted responses are ignored under the conditions with a high ξ, longer test length, and high omission rate, incorrect

and 5-second rule can substantially underestimate person proficiency estimates, while ignored approach can overestimate θ.

Figure 10 shows the bias as a function of the number of item omissions and true omission propensity. One set of parameter estimates from a single replication for the condition with a large sample size (N=750), longer test length (I=30), high omission rate (17%) and $\rho_{.\eta} \neq 0$ is used. The color of the points indicates the number of item omissions for each examinee. For instance, blue denotes zero omissions, while red denotes 30 omissions. As it is shown in Ulitzsch et al. (2020), the bias fluctuates around zero for those whose omission propensity are low; however, the bias increases when the number of omission as well as omission propensity increase. On the other hand, incorrect approach shows slightly downwards bias, while ignored approach shows upwards bias as the number of omission as well as omission propensity increase. Lastly, the results from the 5 seconds method follow closely to that from the SA+O model.

Figure 11 shows the difference between proficiency estimates retrieved from the ad hoc approaches and the SA+O model as a function of omission propensity estimates retrieved from the SA+O model. Since the data-generating model is SA+O model, proficiency is negatively correlated with omission propensity and omission speed. In other words, missing values are not MAR (missing at random). When omissions are ignored in proficiency estimation (i.e., ad hoc approaches), incorrect approach shows the downwards difference, while ignored approach shows the upwards difference as the number of omission as well as omission propensity increase. And the 5-second rule shows the trend in between incorrect and ignored approach.

Overall, the person parameter variance and correlation estimates show less variability as sample size, test length, and omission rate increases. For instance, the 90% range of omission propensity variance estimates ranges from 5 to 10 under the least favorable condition, while that

under the most favorable condition (e.g., N=750, I=30, O=17%) ranges from 5.5 to 7.5. Further, the person parameter variance and correlation estimates involved in omission behavior (i.e., propensity ξ, omission speed η) show more variability, especially under the conditions with a low omission rate.

In summary, it appears again that incorrect and 5-second approach perform very similarly across all conditions. Ignored approach performs well except with low proficiency and high omission propensity, even though this provides information that the test-takers with low proficiency tend to omit an item. Further, the SA+O model does not perform better than incorrect or 5-second approach in terms of the bias and MAE for the person parameter estimates. Overall, test-takers with low proficiency are most affected by the different approaches and the SA+O model does not perform better than ad hoc approaches even with the conditions under high omission rate. Lastly, the Appendix contains the tables (e.g., A.11 through A.14) that present the bias and RMSE for the person parameter variance and correlation estimates of the SA+O model.

Table 3: Proportions of Convergence

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Converged
$\rho_{\cdot\eta} = 0$	375	10	5	0.83
			17	0.99
		30	5	0.84
			17	1.00
	750	10	5	0.93
			17	1.00
		30	5	0.94
			17	1.00
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.75
			17	0.98
		30	5	0.84
			17	1.00
	750	10	5	0.86
			17	1.00
		30	5	0.95
			17	0.98

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

Table 4: Bias for Item Discrimination Parameters

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter 0.750	1.125	1.500	1.875	2.250
$\rho_{\cdot\eta} \neq 0$	375	10	5	Incorrect	0.15	0.12	0.10	0.09	0.06
				Ignored	-0.01	0.00	0.01	0.02	0.02
				5” rule	0.08	0.08	0.07	0.07	0.06
				SA+O	0.04	0.06	0.10	0.17	0.15
			17	Incorrect	0.41	0.34	0.26	0.19	0.06
				Ignored	-0.04	-0.02	-0.03	-0.04	-0.06
				5” rule	0.30	0.27	0.22	0.15	0.06
				SA+O	0.04	0.10	0.12	0.16	0.23
		30	5	Incorrect	0.16	0.11	0.09	0.06	0.08
				Ignored	0.00	0.00	0.01	0.01	0.05
				5” rule	0.10	0.07	0.06	0.05	0.07
				SA+O	0.07	0.09	0.14	0.17	0.25
			17	Incorrect	0.45	0.34	0.26	0.17	0.11
				Ignored	-0.01	-0.03	-0.03	-0.04	0.00
				5” rule	0.34	0.28	0.22	0.15	0.10
				SA+O	0.08	0.11	0.15	0.19	0.27
	750	10	5	Incorrect	0.18	0.11	0.07	0.07	0.00
				Ignored	-0.02	-0.03	-0.03	-0.01	-0.06
				5” rule	0.10	0.07	0.05	0.05	-0.01
				SA+O	0.02	0.03	0.04	0.08	0.07
			17	Incorrect	0.42	0.32	0.23	0.14	0.01
				Ignored	-0.04	-0.06	-0.07	-0.07	-0.13
				5” rule	0.31	0.25	0.19	0.11	0.00
				SA+O	0.02	0.04	0.04	0.07	0.05
		30	5	Incorrect	0.16	0.11	0.07	0.06	0.04
				Ignored	-0.01	-0.01	-0.02	-0.01	-0.01
				5” rule	0.10	0.07	0.05	0.04	0.03
				SA+O	0.03	0.04	0.06	0.08	0.10
			17	Incorrect	0.44	0.34	0.24	0.17	0.07
				Ignored	-0.03	-0.04	-0.05	-0.05	-0.06
				5” rule	0.33	0.27	0.20	0.14	0.06
				SA+O	0.03	0.05	0.06	0.09	0.11

Note. $\rho_{\cdot\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity. N = number of examinees; Item = number of items; Omitted = omission rate.

Table 5: RMSE for Item Discrimination Parameters

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter 0.750	1.125	1.500	1.875	2.250
$\rho_{\cdot\eta} \neq 0$	375	10	5	Incorrect	0.23	0.23	0.27	0.29	0.36
				Ignored	0.17	0.19	0.25	0.29	0.37
				5" rule	0.19	0.21	0.26	0.29	0.36
				SA+O	0.20	0.21	0.28	0.36	0.43
			17	Incorrect	0.46	0.41	0.39	0.33	0.36
				Ignored	0.19	0.22	0.27	0.31	0.41
				5" rule	0.36	0.35	0.35	0.33	0.37
				SA+O	0.21	0.26	0.34	0.38	0.60
		30	5	Incorrect	0.23	0.21	0.22	0.25	0.30
				Ignored	0.15	0.18	0.20	0.24	0.30
				5" rule	0.19	0.20	0.21	0.25	0.30
				SA+O	0.18	0.22	0.25	0.31	0.41
			17	Incorrect	0.49	0.40	0.34	0.31	0.31
				Ignored	0.17	0.20	0.22	0.25	0.31
				5" rule	0.39	0.34	0.31	0.30	0.31
				SA+O	0.21	0.24	0.29	0.34	0.44
	750	10	5	Incorrect	0.21	0.18	0.19	0.23	0.25
				Ignored	0.13	0.15	0.18	0.23	0.26
				5" rule	0.16	0.16	0.18	0.23	0.25
				SA+O	0.12	0.14	0.18	0.23	0.26
			17	Incorrect	0.45	0.35	0.29	0.25	0.23
				Ignored	0.14	0.15	0.19	0.23	0.28
				5" rule	0.35	0.29	0.26	0.23	0.23
				SA+O	0.14	0.15	0.19	0.25	0.28
		30	5	Incorrect	0.20	0.16	0.16	0.18	0.23
				Ignored	0.11	0.12	0.14	0.17	0.22
				5" rule	0.15	0.14	0.15	0.18	0.22
				SA+O	0.11	0.13	0.16	0.19	0.25
			17	Incorrect	0.47	0.36	0.29	0.24	0.24
				Ignored	0.12	0.14	0.16	0.19	0.23
				5" rule	0.36	0.29	0.25	0.23	0.23
				SA+O	0.13	0.15	0.17	0.21	0.27

Note. $\rho_{\cdot\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

Table 6: SD and 95% Coverage Intervals for Item Discrimination Parameters

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					0.750	1.125	1.500	1.875	2.250
$\rho_{\cdot\eta} \neq 0$	375	10	5	Incorrect	0.17	0.19	0.25	0.28	0.36
					[0.58, 1.22]	[0.92, 1.64]	[1.18, 2.17]	[1.54, 2.50]	[1.67, 3.07]
				Ignored	0.17	0.19	0.25	0.29	0.37
					[0.45, 1.08]	[0.78, 1.53]	[1.10, 2.07]	[1.40, 2.53]	[1.61, 3.07]
				5” rule	0.17	0.19	0.25	0.28	0.35
					[0.52, 1.14]	[0.89, 1.61]	[1.16, 2.14]	[1.53, 2.52]	[1.69, 3.04]
				SA+O	0.19	0.20	0.26	0.32	0.40
					[0.46, 1.14]	[0.85, 1.60]	[1.15, 2.14]	[1.51, 2.72]	[1.67, 3.30]
			17	Incorrect	0.21	0.22	0.28	0.28	0.36
					[0.80, 1.61]	[1.08, 1.95]	[1.27, 2.38]	[1.51, 2.66]	[1.75, 3.04]
				Ignored	0.19	0.22	0.27	0.31	0.40
					[0.34, 1.08]	[0.68, 1.59]	[1.01, 2.1]	[1.31, 2.51]	[1.51, 2.96]
				5” rule	0.20	0.22	0.28	0.29	0.36
					[0.68, 1.47]	[1.03, 1.89]	[1.24, 2.29]	[1.49, 2.63]	[1.74, 3.06]
				SA+O	0.21	0.24	0.31	0.34	0.56
					[0.38, 1.20]	[0.77, 1.78]	[1.11, 2.30]	[1.44, 2.75]	[1.72, 3.64]
		30	5	Incorrect	0.16	0.18	0.20	0.24	0.29
					[0.61, 1.24]	[0.94, 1.60]	[1.23, 1.99]	[1.56, 2.49]	[1.80, 2.97]
				Ignored	0.15	0.18	0.20	0.24	0.30
					[0.47, 1.05]	[0.82, 1.49]	[1.13, 1.92]	[1.49, 2.43]	[1.77, 2.97]
				5” rule	0.16	0.18	0.20	0.24	0.29
					[0.55, 1.18]	[0.90, 1.54]	[1.19, 1.96]	[1.53, 2.46]	[1.79, 2.96]
				SA+O	0.16	0.20	0.22	0.26	0.33
					[0.50, 1.16]	[0.89, 1.63]	[1.23, 2.06]	[1.60, 2.62]	[1.89, 3.22]

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					0.750	1.125	1.500	1.875	2.250
			17	Incorrect	0.19	0.21	0.23	0.26	0.29
					[0.84, 1.58]	[1.10, 1.89]	[1.31, 2.20]	[1.66, 2.63]	[1.83, 3.02]
				Ignored	0.17	0.19	0.22	0.25	0.31
					[0.42, 1.10]	[0.77, 1.49]	[1.06, 1.92]	[1.42, 2.39]	[1.72, 2.95]
				5” rule	0.18	0.20	0.22	0.26	0.29
					[0.76, 1.48]	[1.05, 1.81]	[1.28, 2.17]	[1.63, 2.61]	[1.82, 2.99]
				SA+O	0.19	0.22	0.25	0.29	0.35
					[0.47, 1.26]	[0.86, 1.67]	[1.20, 2.14]	[1.61, 2.70]	[1.91, 3.36]
	750	10	5	Incorrect	0.13	0.14	0.16	0.20	0.23
					[0.62, 1.15]	[0.97, 1.55]	[1.26, 1.85]	[1.57, 2.40]	[1.83, 2.76]
				Ignored	0.12	0.13	0.17	0.21	0.23
					[0.49, 0.95]	[0.87, 1.45]	[1.16, 1.79]	[1.48, 2.29]	[1.75, 2.72]
				5” rule	0.12	0.14	0.16	0.21	0.23
					[0.57, 1.05]	[0.92, 1.49]	[1.24, 1.85]	[1.55, 2.38]	[1.82, 2.76]
				SA+O	0.12	0.14	0.17	0.22	0.25
					[0.52, 0.99]	[0.91, 1.46]	[1.22, 1.85]	[1.53, 2.43]	[1.81, 2.86]
			17	Incorrect	0.18	0.14	0.18	0.21	0.23
					[0.80, 1.49]	[1.16, 1.70]	[1.39, 2.07]	[1.60, 2.43]	[1.86, 2.72]
				Ignored	0.13	0.14	0.18	0.22	0.25
					[0.42, 0.96]	[0.83, 1.35]	[1.05, 1.80]	[1.39, 2.26]	[1.72, 2.67]
				5” rule	0.16	0.14	0.17	0.20	0.23
					[0.72, 1.35]	[1.09, 1.61]	[1.34, 2.01]	[1.58, 2.39]	[1.85, 2.71]
				SA+O	0.14	0.15	0.19	0.24	0.27
					[0.48, 1.03]	[0.89, 1.47]	[1.14, 1.89]	[1.45, 2.47]	[1.85, 2.92]
		30	5	Incorrect	0.12	0.12	0.14	0.17	0.22
					[0.67, 1.13]	[1.00, 1.47]	[1.31, 1.86]	[1.63, 2.40]	[1.89, 2.77]

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					0.750	1.125	1.500	1.875	2.250
				Ignored	0.11	0.12	0.14	0.17	0.22
					[0.53, 0.94]	[0.89, 1.34]	[1.23, 1.76]	[1.59, 2.24]	[1.87, 2.73]
				5” rule	0.11	0.12	0.14	0.17	0.22
					[0.63, 1.05]	[0.97, 1.44]	[1.29, 1.82]	[1.63, 2.26]	[1.89, 2.76]
				SA+O	0.11	0.13	0.15	0.18	0.23
					[0.56, 0.99]	[0.94, 1.41]	[1.28, 1.84]	[1.67, 2.35]	[1.95, 2.85]
			17	Incorrect	0.16	0.14	0.15	0.18	0.22
					[0.86, 1.47]	[1.22, 1.74]	[1.45, 2.06]	[1.72, 2.41]	[1.93, 2.81]
				Ignored	0.12	0.13	0.15	0.18	0.22
					[0.48, 0.95]	[0.83, 1.35]	[1.17, 1.76]	[1.52, 2.23]	[1.79, 2.72]
				5” rule	0.14	0.13	0.15	0.18	0.22
					[0.79, 1.35]	[1.14, 1.67]	[1.39, 2.01]	[1.70, 2.40]	[1.91, 2.81]
				SA+O	0.13	0.14	0.16	0.19	0.24
					[0.52, 1.03]	[0.90, 1.45]	[1.26, 1.90]	[1.63, 2.39]	[1.93, 2.93]

Note. $\rho_{\cdot\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

Table 7: Bias for Item Difficulty Parameters

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					-1.0	-0.5	0.0	0.5	1.0
$\rho_{\cdot\eta}$	375	10	5	Incorrect	0.13	0.08	0.08	0.07	0.07
$\neq 0$				Ignored	-0.02	-0.03	-0.02	-0.02	-0.02
				5" rule	0.09	0.06	0.06	0.05	0.05
				SA+O	-0.01	-0.02	0.01	0.01	0.03
			17	Incorrect	0.51	0.38	0.33	0.28	0.26
				Ignored	-0.09	-0.11	-0.10	-0.11	-0.10
				5" rule	0.51	0.34	0.29	0.25	0.21
				SA+O	-0.03	-0.01	0.02	0.01	0.02
		30	5	Incorrect	0.13	0.08	0.07	0.07	0.08
				Ignored	-0.02	-0.02	-0.02	-0.01	-0.01
				5" rule	0.09	0.06	0.05	0.05	0.06
				SA+O	-0.02	-0.02	-0.01	0.00	0.02
			17	Incorrect	0.59	0.43	0.37	0.30	0.30
				Ignored	-0.07	-0.07	-0.07	-0.06	-0.05
				5" rule	0.45	0.35	0.30	0.26	0.25
				SA+O	-0.02	-0.01	0.00	0.02	0.03
	750	10	5	Incorrect	0.16	0.09	0.08	0.08	0.09
				Ignored	-0.02	-0.05	-0.03	-0.02	-0.01
				5" rule	0.12	0.06	0.06	0.06	0.07
				SA+O	-0.01	-0.01	0.00	0.02	0.02
			17	Incorrect	0.51	0.36	0.31	0.29	0.28
				Ignored	-0.09	-0.13	-0.11	-0.10	-0.09
				5" rule	0.44	0.32	0.28	0.25	0.23
				SA+O	0.00	-0.02	0.00	0.02	0.03
		30	5	Incorrect	0.14	0.10	0.08	0.08	0.08
				Ignored	-0.02	-0.01	-0.01	-0.01	-0.01
				5" rule	0.10	0.07	0.06	0.06	0.05
				SA+O	-0.01	0.00	0.00	0.01	0.01
			17	Incorrect	0.54	0.41	0.33	0.31	0.30
				Ignored	-0.06	-0.06	-0.07	-0.06	-0.06
				5" rule	0.47	0.36	0.30	0.27	0.25
				SA+O	-0.01	0.00	0.00	0.01	0.01

Note. $\rho_{\cdot\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

Table 8: RMSE for Item Difficulty Parameters

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					-1.0	-0.5	0.0	0.5	1.0
$\rho_{\cdot\eta} \neq 0$	375	10	5	Incorrect	0.25	0.18	0.17	0.18	0.20
				Ignored	0.20	0.17	0.14	0.17	0.19
				5” rule	0.22	0.17	0.16	0.17	0.19
				SA+O	0.19	0.17	0.15	0.17	0.19
			17	Incorrect	0.58	0.42	0.36	0.34	0.35
				Ignored	0.22	0.21	0.18	0.21	0.22
				5” rule	0.51	0.39	0.33	0.31	0.30
				SA+O	0.25	0.17	0.16	0.19	0.23
		30	5	Incorrect	0.24	0.19	0.18	0.18	0.21
				Ignored	0.19	0.17	0.16	0.17	0.18
				5” rule	0.21	0.18	0.17	0.17	0.19
				SA+O	0.19	0.17	0.17	0.17	0.19
			17	Incorrect	0.59	0.43	0.37	0.36	0.38
				Ignored	0.21	0.19	0.18	0.18	0.20
				5” rule	0.52	0.40	0.34	0.32	0.33
				SA+O	0.20	0.18	0.17	0.18	0.20
	750	10	5	Incorrect	0.21	0.16	0.14	0.14	0.17
				Ignored	0.14	0.14	0.12	0.12	0.14
				5” rule	0.18	0.15	0.13	0.13	0.16
				SA+O	0.13	0.12	0.11	0.11	0.14
			17	Incorrect	0.57	0.40	0.33	0.32	0.34
				Ignored	0.17	0.18	0.16	0.15	0.16
				5” rule	0.49	0.36	0.30	0.28	0.29
				SA+O	0.13	0.13	0.12	0.12	0.14
		30	5	Incorrect	0.20	0.15	0.14	0.14	0.15
				Ignored	0.13	0.12	0.11	0.11	0.13
				5” rule	0.17	0.14	0.12	0.13	0.14
				SA+O	0.13	0.12	0.11	0.11	0.12
			17	Incorrect	0.59	0.43	0.35	0.34	0.36
				Ignored	0.15	0.14	0.13	0.13	0.14
				5” rule	0.51	0.39	0.32	0.30	0.30
				SA+O	0.13	0.12	0.11	0.12	0.13

Note. $\rho_{\cdot\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

Table 9: SD and 95% Coverage Intervals for Item Difficulty Parameters

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					-1.0	-0.5	0.0	0.5	1.0
$\rho_{\cdot\eta} \neq 0$	375	10	5	Incorrect	0.21	0.17	0.15	0.17	0.19
					[-1.33, -0.53]	[-0.79, -0.12]	[-0.17, 0.38]	[0.22, 0.86]	[0.65, 1.47]
				Ignored	0.20	0.16	0.14	0.17	0.19
					[-1.47, -0.66]	[-0.92, -0.23]	[-0.28, 0.27]	[0.14, 0.78]	[0.59, 1.43]
				5” rule	0.20	0.16	0.14	0.17	0.18
					[-1.36, -0.60]	[-0.81, -0.14]	[-0.20, 0.35]	[0.21, 0.85]	[0.64, 1.47]
				SA+O	0.19	0.17	0.15	0.17	0.19
					[-1.48, -0.67]	[-0.92, -0.18]	[-0.26, 0.33]	[0.12, 0.79]	[0.62, 1.38]
			17	Incorrect	0.29	0.19	0.16	0.20	0.24
					[-1.09, -0.08]	[-0.53, 0.22]	[0.05, 0.65]	[0.36, 1.12]	[0.73, 1.68]
				Ignored	0.20	0.17	0.15	0.17	0.20
					[-1.57, -0.76]	[-1.00, -0.29]	[-0.37, 0.19]	[0.06, 0.66]	[0.52, 1.28]
				5” rule	0.26	0.19	0.16	0.19	0.22
					[-1.15, -0.19]	[-0.55, 0.16]	[0.02, 0.62]	[0.33, 1.08]	[0.73, 1.61]
				SA+O	0.25	0.17	0.16	0.19	0.23
					[-1.55, -0.60]	[-0.88, -0.21]	[-0.25, 0.36]	[0.18, 0.85]	[0.63, 1.47]
		30	5	Incorrect	0.20	0.17	0.16	0.17	0.19
					[-1.29, -0.53]	[-0.77, -0.07]	[-0.26, 0.36]	[0.23, 0.90]	[0.69, 1.43]
				Ignored	0.19	0.17	0.16	0.17	0.18
					[-1.39, -0.69]	[-0.86, -0.17]	[-0.35, 0.29]	[0.16, 0.82]	[0.62, 1.36]
				5” rule	0.19	0.17	0.16	0.17	0.19
					[-1.31, -0.56]	[-0.78, -0.09]	[-0.28, 0.35]	[0.21, 0.88]	[0.67, 1.41]
				SA+O	0.19	0.17	0.17	0.17	0.19
					[-1.39, -0.66]	[-0.86, -0.15]	[-0.35, 0.30]	[0.17, 0.84]	[0.64, 1.41]

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					-1.0	-0.5	0.0	0.5	1.0
			17	Incorrect	0.29	0.19	0.17	0.19	0.24
					[-1.03, -0.01]	[-0.51, 0.26]	[-0.02, 0.66]	[0.39, 1.19]	[0.83, 1.70]
				Ignored	0.20	0.17	0.17	0.17	0.20
					[-1.49, -0.72]	[-0.90, -0.22]	[-0.42, 0.24]	[0.10, 0.77]	[0.57, 1.31]
				5” rule	0.26	0.18	0.17	0.19	0.22
					[-1.07, -0.11]	[-0.54, 0.21]	[-0.04, 0.61]	[0.37, 1.15]	[0.81, 1.64]
				SA+O	0.20	0.18	0.17	0.18	0.20
					[-1.42, -0.65]	[-0.85, -0.15]	[-0.33, 0.32]	[0.17, 0.87]	[0.62, 1.43]
	750	10	5	Incorrect	0.14	0.13	0.11	0.11	0.14
					[-1.22, -0.65]	[-0.71, -0.20]	[-0.13, 0.30]	[0.31, 0.78]	[0.82, 1.34]
				Ignored	0.13	0.12	0.11	0.11	0.13
					[-1.33, -0.76]	[-0.78, -0.32]	[-0.21, 0.21]	[0.25, 0.67]	[0.76, 1.24]
				5” rule	0.13	0.12	0.11	0.11	0.13
					[-1.23, -0.68]	[-0.72, -0.23]	[-0.14, 0.28]	[0.30, 0.76]	[0.82, 1.31]
				SA+O	0.13	0.12	0.11	0.11	0.13
					[-1.30, -0.78]	[-0.79, -0.30]	[-0.19, 0.23]	[0.27, 0.70]	[0.78, 1.29]
			17	Incorrect	0.26	0.16	0.12	0.14	0.20
					[-0.95, -0.14]	[-0.46, 0.13]	[0.10, 0.56]	[0.51, 1.06]	[0.89, 1.59]
				Ignored	0.14	0.13	0.12	0.11	0.14
					[-1.42, -0.83]	[-0.94, -0.38]	[-0.30, 0.12]	[0.17, 0.61]	[0.63, 1.22]
				5” rule	0.21	0.15	0.12	0.13	0.17
					[-0.98, -0.24]	[-0.48, 0.08]	[0.07, 0.53]	[0.49, 1.00]	[0.87, 1.51]
				SA+O	0.13	0.13	0.12	0.12	0.14
					[-1.31, -0.74]	[-0.79, -0.27]	[-0.19, 0.27]	[0.28, 0.72]	[0.77, 1.35]
		30	5	Incorrect	0.15	0.12	0.11	0.11	0.13
					[-1.18, -0.61]	[-0.64, -0.19]	[-0.14, 0.28]	[0.36, 0.79]	[0.81, 1.33]

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	Omission Treatment	True Parameter				
					-1.0	-0.5	0.0	0.5	1.0
				Ignored	0.13	0.11	0.11	0.11	0.12
					[-1.29, -0.76]	[-0.74, -0.29]	[-0.23, 0.20]	[0.29, 0.70]	[0.76, 1.23]
				5” rule	0.14	0.12	0.11	0.11	0.12
					[-1.21, -0.66]	[-0.66, -0.22]	[-0.15, 0.26]	[0.34, 0.77]	[0.81, 1.29]
				SA+O	0.13	0.12	0.11	0.11	0.12
					[-1.27, -0.75]	[-0.72, -0.28]	[-0.22, 0.21]	[0.31, 0.72]	[0.78, 1.26]
			17	Incorrect	0.25	0.15	0.12	0.14	0.19
					[-0.95, -0.10]	[-0.40, 0.15]	[0.08, 0.56]	[0.53, 1.07]	[0.93, 1.63]
				Ignored	0.13	0.12	0.11	0.11	0.13
					[-1.33, -0.82]	[-0.81, -0.34]	[-0.28, 0.15]	[0.23, 0.67]	[0.70, 1.22]
				5” rule	0.22	0.14	0.11	0.13	0.17
					[-0.98, -0.21]	[-0.43, 0.11]	[0.06, 0.53]	[0.50, 1.04]	[0.91, 1.56]
				SA+O	0.13	0.12	0.11	0.12	0.13
					[-1.28, -0.76]	[-0.75, -0.27]	[-0.23, 0.21]	[0.30, 0.75]	[0.76, 1.29]

Note. $\rho_{\cdot\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

Table 10: Bias of Person Proficiency Parameters

				Positive θ (θ > 1)				Central θ (-1≤ θ ≤ 1)				Negative θ (θ < -1)			
$\rho_{.\eta}$	N	Item	Omitted (%)	SA+O	Inc	Ign	5 sec	SA+ O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec
$\rho_{.\eta} \neq 0$	375	10	5	-0.36	-0.35	-0.37	-0.36	0.01	0.01	-0.01	0.00	0.36	0.35	0.40	0.36
			17	-0.36	-0.34	-0.40	-0.34	-0.01	0.00	-0.04	-0.01	0.42	0.37	0.47	0.38
		30	5	-0.21	-0.16	-0.14	-0.15	0.00	0.01	-0.01	0.00	0.20	0.12	0.17	0.13
			17	-0.21	-0.19	-0.15	-0.18	-0.01	0.01	-0.03	0.00	0.24	0.16	0.24	0.16
	750	10	5	-0.34	-0.34	-0.36	-0.34	0.00	0.00	-0.01	0.00	0.35	0.35	0.41	0.37
			17	-0.34	-0.33	-0.39	-0.33	-0.01	0.00	-0.04	-0.01	0.41	0.38	0.48	0.39
		30	5	-0.16	-0.15	-0.14	-0.14	0.00	0.01	-0.01	0.00	0.18	0.13	0.17	0.14
			17	-0.17	-0.18	-0.14	-0.17	-0.01	0.01	-0.03	0.00	0.22	0.16	0.24	0.17

Note. $\rho_{.\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity. N = number of examinees; Item = number of items; Omitted = omission rate; θ = examinee proficiency; Inc = incorrect; Ign = ignored; 5 sec = 5-second rule.

Table 11: Mean Absolute Error (MAE) of Person Proficiency Parameters

				Positive θ (θ > 1)				Central θ (-1≤ θ ≤ 1)				Negative θ (θ < -1)			
$\rho_{.\eta}$	N	Item	Omitted (%)	SA+O	Inc	Ign	5 sec	SA+ O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec
$\rho_{.\eta} \neq 0$	375	10	5	0.46	0.45	0.46	0.45	0.33	0.34	0.35	0.34	0.45	0.45	0.49	0.46
			17	0.46	0.45	0.48	0.45	0.34	0.36	0.36	0.36	0.49	0.47	0.56	0.48
		30	5	0.33	0.31	0.31	0.31	0.21	0.22	0.22	0.22	0.33	0.32	0.34	0.32
			17	0.33	0.33	0.32	0.32	0.22	0.27	0.24	0.26	0.36	0.38	0.42	0.37
	750	10	5	0.45	0.45	0.45	0.45	0.33	0.33	0.34	0.34	0.45	0.46	0.50	0.46
			17	0.44	0.45	0.47	0.45	0.34	0.36	0.36	0.36	0.48	0.48	0.57	0.48
		30	5	0.31	0.31	0.31	0.31	0.21	0.21	0.22	0.21	0.32	0.33	0.35	0.33
			17	0.31	0.32	0.31	0.32	0.22	0.26	0.24	0.25	0.36	0.39	0.42	0.38

Note. $\rho_{.\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity. N = number of examinees; Item = number of items; Omitted = omission rate; θ = examinee proficiency; Inc = incorrect; Ign = ignored; 5 sec = 5-second rule.

Table 12: Mean SD and 95% Coverage Intervals of Person Proficiency Parameters

				Positive θ (θ > 1)				Central θ (-1≤ θ ≤ 1)				Negative θ (θ < -1)			
$\rho_{\cdot\eta}$	N	Item	Omit	SA+O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.40 [0.23, 1.77]	0.39 [0.23, 1.67]	0.38 [0.21, 1.64]	0.39 [0.23, 1.66]	0.64 [-1.28, 1.21]	0.66 [-1.35, 1.25]	0.67 [-1.35, 1.26]	0.66 [-1.34, 1.26]	0.40 [-1.95, -0.44]	0.39 [-1.60, -0.44]	0.38 [-1.66, -0.40]	0.38 [-1.61 - 0.44]
			17	0.41 [0.19, 1.79]	0.41 [0.11, 1.75]	0.39 [0.15, 1.61]	0.41 [0.12, 1.74]	0.65 [-1.34, 1.20]	0.68 [-1.45, 1.24]	0.67 [-1.35, 1.24]	0.68 [-1.45, 1.24]	0.35 [-1.73, -0.41]	0.37 [-1.46, -0.41]	0.41 [-1.69, -0.28]	0.36 [-1.49, -0.42]
		30	5	0.42 [0.51, 2.10]	0.41 [0.53, 2.12]	0.42 [0.54, 2.13]	0.41 [0.55, 2.13]	0.57 [-1.12, 1.02]	0.60 [-1.19, 1.08]	0.61 [-1.20, 1.09]	0.60 [-1.20, 1.08]	0.41 [-2.25, -0.67]	0.44 [-2.19, -0.68]	0.40 [-2.19, -0.70]	0.42 [-2.20, -0.69]
			17	0.42 [0.49, 2.10]	0.43 [0.28, 2.10]	0.42 [0.51, 2.12]	0.43 [0.32, 2.12]	0.58 [-1.20, 1.02]	0.63 [-1.43, 1.07]	0.63 [-1.27, 1.10]	0.63 [-1.38, 1.07]	0.37 [-2.00, -0.63]	0.46 [-1.95, -0.59]	0.42 [-2.12, -0.55]	0.43 [-2.00, -0.62]
	750	10	5	0.40 [0.29, 1.81]	0.39 [0.29, 1.68]	0.39 [0.27, 1.65]	0.39 [0.29, 1.67]	0.64 [-1.27, 1.25]	0.66 [-1.31, 1.28]	0.67 [-1.34, 1.29]	0.66 [-1.32, 1.29]	0.41 [-1.94, -0.40]	0.39 [-1.60, -0.39]	0.39 [-1.67, -0.35]	0.38 [-1.62, -0.38]
			17	0.41 [0.27, 1.82]	0.42 [0.22, 1.76]	0.39 [0.22, 1.62]	0.42 [0.23, 1.75]	0.66 [-1.34, 1.24]	0.68 [-1.47, 1.26]	0.67 [-1.35, 1.27]	0.68 [-1.47, 1.26]	0.36 [-1.68, -0.37]	0.38 [-1.47, -0.35]	0.41 [-1.70, -0.18]	0.37 [-1.49, -0.36]
		30	5	0.41 [0.59, 2.16]	0.41 [0.60, 2.16]	0.41 [0.60, 2.17]	0.41 [0.60, 2.17]	0.58 [-1.10, 1.08]	0.60 [-1.14, 1.10]	0.61 [-1.15, 1.12]	0.60 [-1.14, 1.11]	0.41 [-2.21, -0.66]	0.45 [-2.20, -0.65]	0.40 [-2.17, -0.66]	0.42 [-2.22, -0.67]
			17	0.42 [0.57, 2.17]	0.42 [0.47, 2.15]	0.42 [0.57, 2.17]	0.42 [0.49, 2.17]	0.60 [-1.19, 1.07]	0.63 [-1.39, 1.09]	0.62 [-1.23, 1.12]	0.63 [-1.34, 1.10]	0.37 [-1.98, -0.62]	0.46 [-1.95, -0.55]	0.41 [-2.10, -0.50]	0.43 [-2.00, -0.58]

Note. $\rho_{\cdot\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity. N = number of examinees; Item = number of items; Omitted = omission rate; θ = examinee proficiency; Inc = incorrect; Ign = ignored; 5 sec = 5-second rule.

Table 13: Bias of Person Proficiency Parameter Conditioning on True Omission Propensity

				High ξ ($\xi > 2.5$)				Central ξ ($-2.5 \leq \xi \leq 2.5$)				Low ξ ($\xi < -2.5$)			
$\rho_{.\eta}$	N	Item	Omitted (%)	SA+O	Inc	Ign	5 sec	SA+ O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec
$\rho_{.\eta} \neq 0$	375	10	5	0.09	0.11	0.27	0.15	0.03	0.03	-0.01	0.02	-0.18	-0.20	-0.22	-0.20
			17	0.07	-0.07	0.36	-0.02	0.02	0.05	-0.05	0.04	-0.15	-0.11	-0.26	-0.12
		30	5	0.05	-0.05	0.12	0.00	0.01	0.03	-0.01	0.02	-0.12	-0.09	-0.10	-0.09
			17	0.04	-0.27	0.20	-0.22	0.01	0.07	-0.03	0.05	-0.10	-0.02	-0.12	-0.02
	750	10	5	0.08	0.11	0.27	0.15	0.03	0.03	-0.01	0.02	-0.18	-0.20	-0.23	-0.20
			17	0.04	-0.08	0.35	-0.03	0.03	0.06	-0.04	0.04	-0.14	-0.12	-0.26	-0.13
		30	5	0.04	-0.04	0.13	0.01	0.01	0.03	-0.01	0.02	-0.08	-0.07	-0.09	-0.07
			17	0.03	-0.26	0.22	-0.20	0.01	0.07	-0.04	0.05	-0.06	-0.01	-0.10	-0.01

Note. $\rho_{.\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity. N = number of examinees; Item = number of items; Omitted = omission rate; θ = examinee proficiency; Inc = incorrect; Ign = ignored; 5 sec = 5-second rule.

Table 14: Mean Absolute Error (MAE) of Person Proficiency Parameter Conditioning on True Omission Propensity

				High ξ ($\xi > 2.5$)				Central ξ ($-2.5 \leq \xi \leq 2.5$)				Low ξ ($\xi < -2.5$)			
$\rho_{.\eta}$	N	Item	Omitted (%)	SA+O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec
$\rho_{.\eta} \neq 0$	375	10	5	0.41	0.42	0.48	0.43	0.35	0.35	0.36	0.35	0.41	0.42	0.43	0.42
			17	0.48	0.50	0.57	0.50	0.35	0.37	0.37	0.36	0.40	0.40	0.44	0.40
		30	5	0.29	0.31	0.31	0.30	0.23	0.23	0.23	0.23	0.28	0.28	0.28	0.28
			17	0.36	0.47	0.42	0.44	0.23	0.26	0.24	0.25	0.28	0.28	0.29	0.27
	750	10	5	0.41	0.42	0.47	0.43	0.35	0.35	0.35	0.35	0.41	0.41	0.42	0.41
			17	0.48	0.50	0.57	0.50	0.35	0.36	0.37	0.36	0.39	0.40	0.44	0.40
		30	5	0.29	0.31	0.32	0.30	0.22	0.23	0.23	0.23	0.27	0.27	0.27	0.27
			17	0.36	0.47	0.43	0.45	0.23	0.25	0.24	0.25	0.27	0.27	0.28	0.27

Note. $\rho_{.\eta} \neq 0$ denote omission speed is correlated with proficiency, speed, and omission propensity. N = number of examinees; Item = number of items; Omitted = omission rate; θ = examinee proficiency; Inc = incorrect; Ign = ignored; 5 sec = 5-second rule.

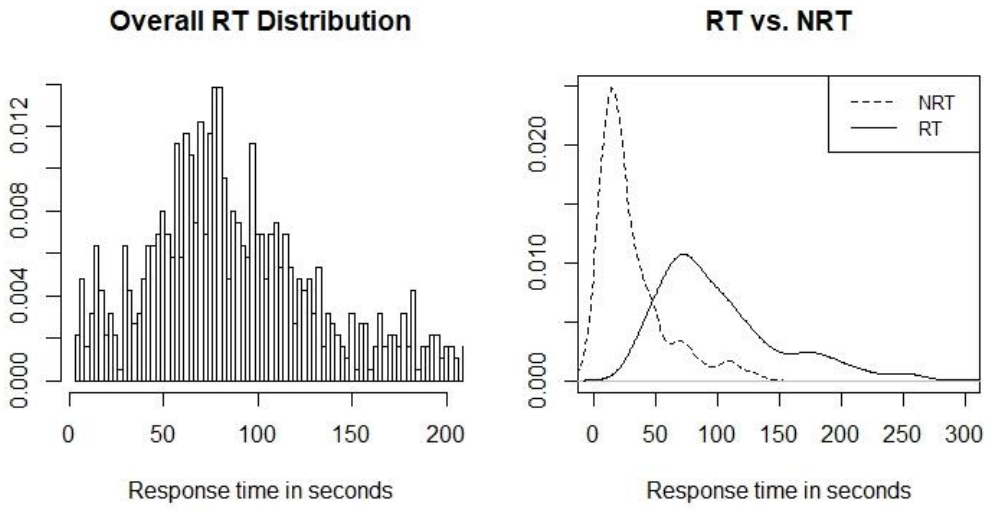


Figure 7: Item Response Time Distribution

$\rho_{.\eta} \neq 0$

a

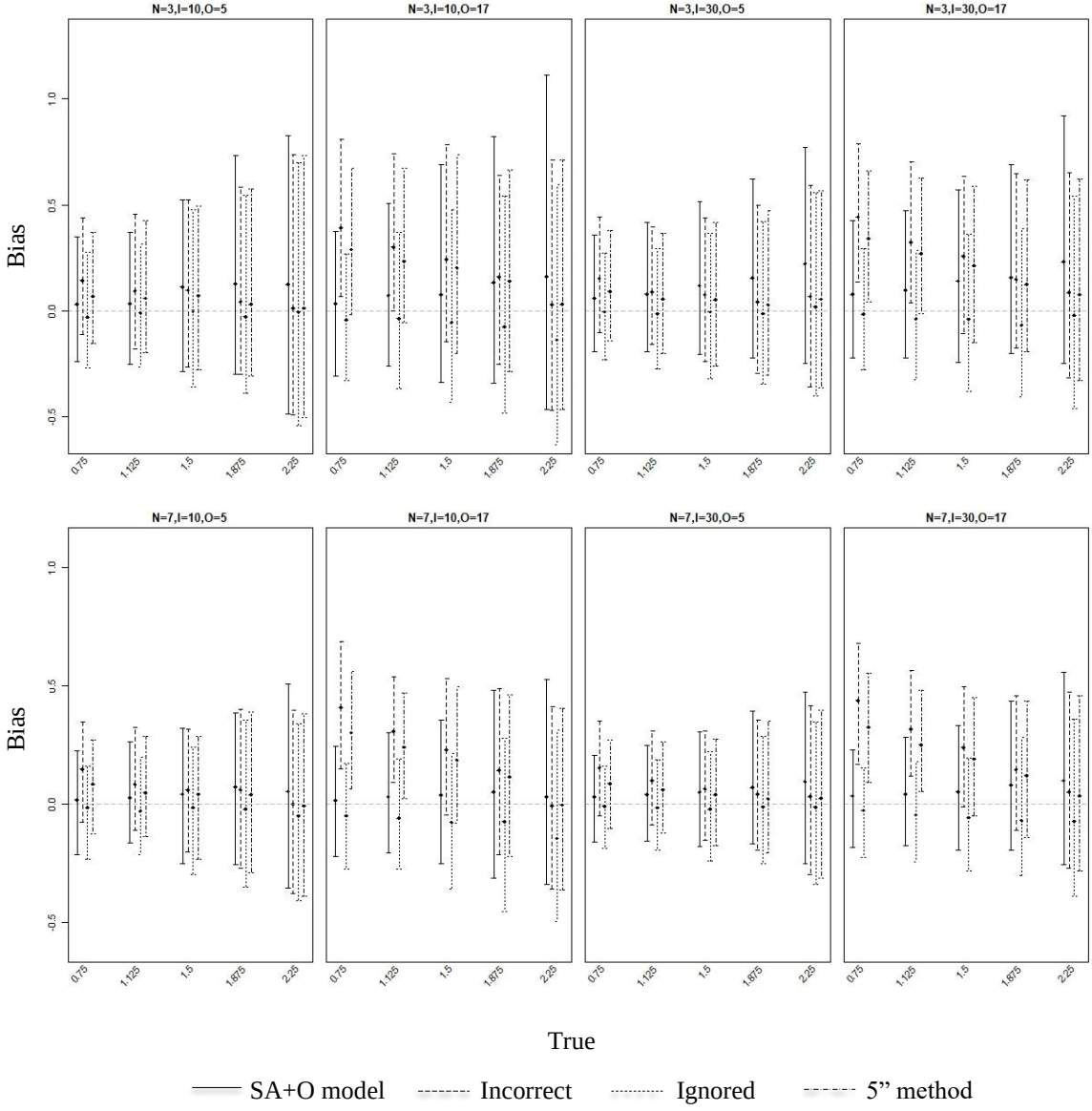


Figure 8: Medians and 90% Ranges of Differences between Estimated and True Item Discrimination Parameters under the Condition of $\rho_{.\eta} \neq 0$, Plotted against the True Parameters

$\rho_{.\eta} \neq 0$

b

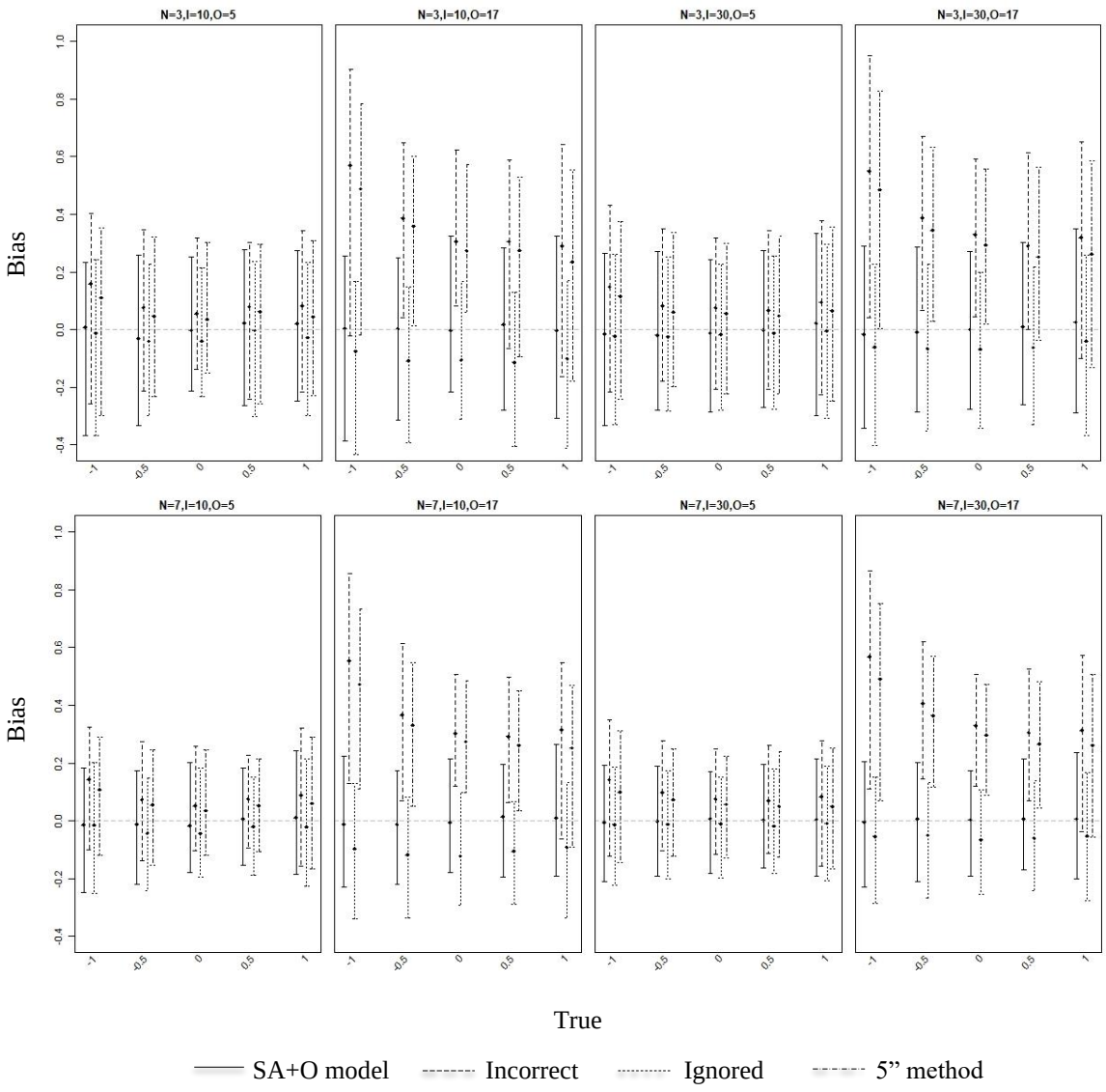
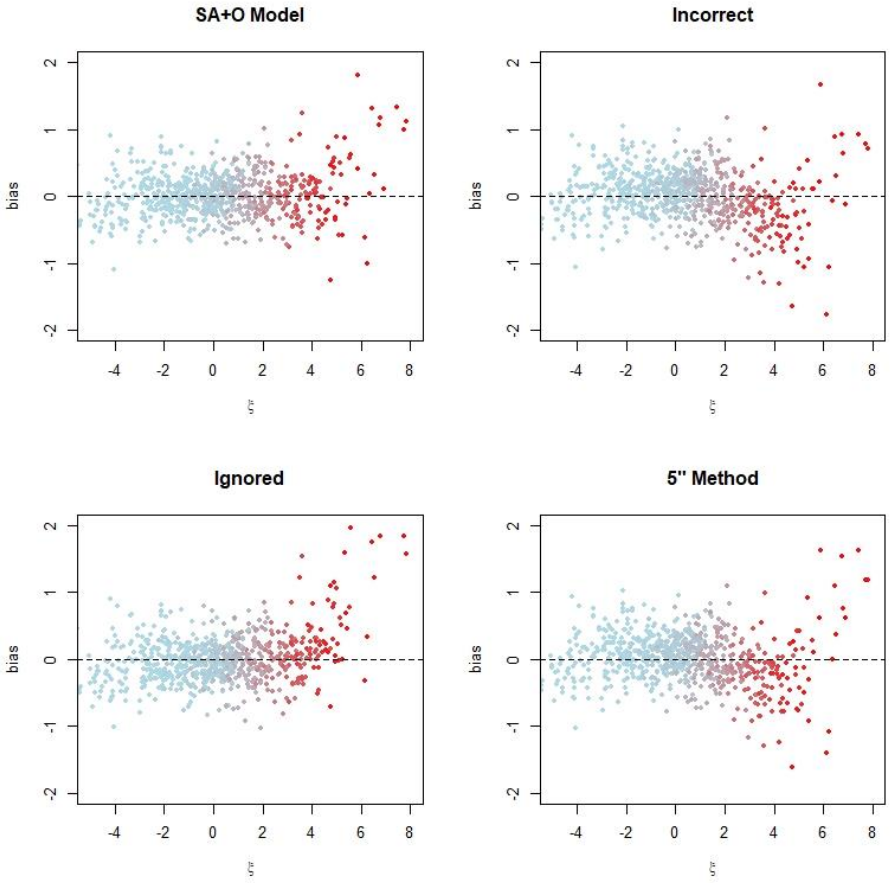
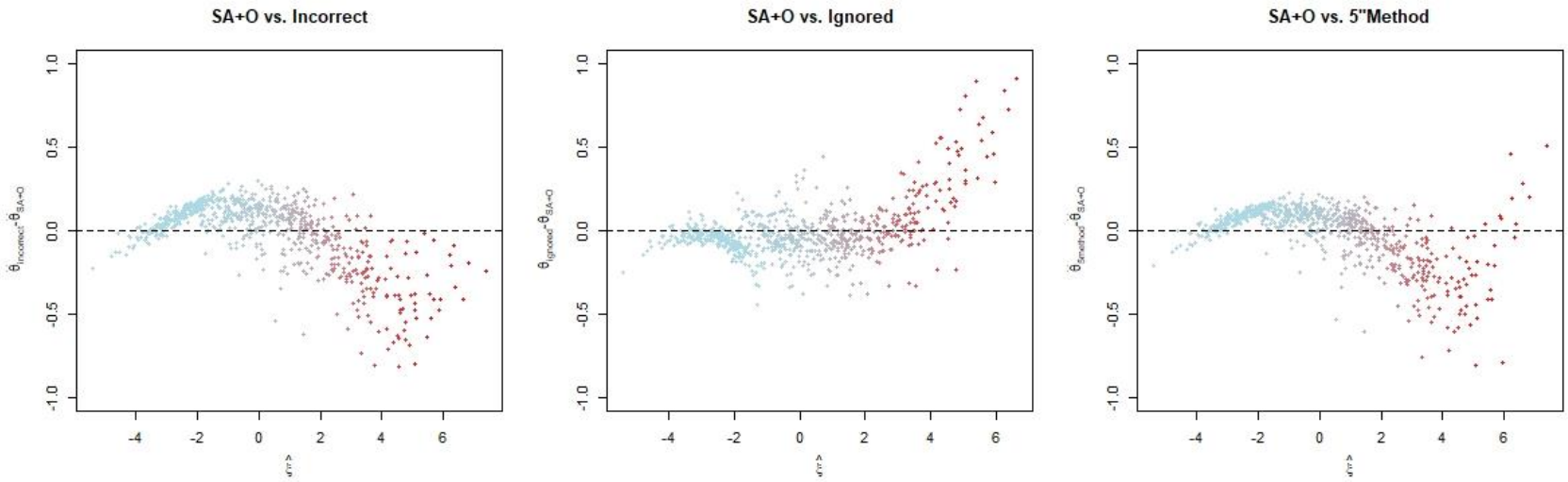


Figure 9: Medians and 90% Ranges of Differences between Estimated and True Item Difficulty Parameters under the Condition of $\rho_{.\eta} \neq 0$, Plotted against the True Parameters



Note. Blue dots denote zero omissions. Red dots denote 30 omissions.

Figure 10: Bias in Proficiency Estimates Retrieved from the SA+O model and Ad Hoc Approaches, Plotted Against True Omission Propensity



Note. Blue dots denote zero omissions. Red dots denote 30 omissions.

Figure 11: Difference in Proficiency Estimates Retrieved from the SA+O model and Ad Hoc Approaches, Plotted Against Omission Propensity Estimates Retrieved from the SA+O model

CHAPTER 5

EMPIRICAL APPLICATION

5.1 Purpose and Description of Dataset

In educational measurement, there has been a great interest in analyzing cognitive process as well as behavioral process while taking a test. The rise of CBTs and emerging technologies enables collection of process data, including RT information, eye tracking, keystroke data, and collaboration processes (Bergner & von Davier, 2019). In addition, innovative item types or technology-enhanced items are implemented in the LSA to explore new constructs such as NAEP digitally based assessment, PISA collaborative problem solving assessment, and PIAAC problem solving in technology-rich environments items. In particular, RT information have been used for analyzing test engagement in LSAs (Lee & Jia, 2014; Weeks et al., 2016).

Here, we re-analyze the empirical dataset from Ulitzsch et al. (2019a) (e.g., Chilean sample of PIAAC 2012), with the goal of comparing item and person parameter estimates. From over 40 participating countries' samples, the Chilean sample was selected due to its highest omission rate (e.g., 16.5%) on the numeracy items in PIAAC 2012 (OECD, 2013; Ulitzsch et al., 2020). The Main Study CBA consists of two modules and within a module, there are three domains: literacy, numeracy, and problem-solving (OECD, 2013). For each module, an examinee responds a total of 20 items: 9 items in Stage 1 and 11 items in Stage 2 for literacy and numeracy, while problem-solving module is designed to take an average of 30 minutes (Figure 13). For instance, an examinee is randomly assigned to a literacy domain in the fist cognitive assessment module, followed by a numeracy domain in the second module. More specifically, PIAAC implements a 2-2 multistage adaptive testing

design for literacy and numeracy. As shown in Figure 14, the literacy and numeracy modules each consists of two stages and each stage contains testlets varying in difficulty: three levels of difficulty in Stage 1 and four levels of difficulty in Stage 2. For instance, a leve of testlet difficulty in Stage 1 is determined by the background variables collected in the background questionnaire, along with the score received on the cognitive screener, while that in Stage 2 is based on the background variables, cognitive screener, and examinee's performance on the set of items administered in Stage 1 (Kirsch, & Lennon, 2017).

Of the 5212 Chilean examinees, 392 examinees were sampled from a numeracy testlet of medium difficulty in Stage 1 of the second cognitive assessment module. Nine examinees who didn't attempt the test were excluded from further analyses. Thus, the data set consisted of 383 eaminees and 9 items in total. The overall omission rate was 17.6%, ranging from 7.8% to 27.9% per item. 52.2% of examinees omitted an item at least once, while 36% of them showed omission behavior only once. For ad hoc approaches, omitted responses were scored as incorrect, ignored, or using 5-second rule. For a 5-second rule, if an examinee spends more than or equal to five seconds on an item, the omitted response is scored as incorrect, while examinee spends less than 5 seconds, the omitted response is ignored. For the SA+O model, the original RTs were log-transformed.

5.2 Estimation

For the ad hoc and model-based approaches, the IRT parameter estimation was used as in the simulation study.

5.3 Results

5.3.1 SA+O Model Estimates

The SA+O model converged properly and the results are presented in Table 15-16. Overall, the SA+O model results for item and person parameter estimates correspond to those in Ulitzsch et al. (2020). Table 15 presents the person parameter variances, correlations and credible intervals. First, the results show that examinee working speed τ and omission speed η are different constructs. For instance, the variability of omission speed between individuals ($var(\eta) = 6.65$) is much higher than that of working speed ($var(\tau) = 0.19$). And these speed components show a high correlation ($cor(\tau, \eta) = 0.66$), but differences in their correlations with other person parameters, suggesting that examinees operate on different speed levels for generating item responses or nonresponses. The negative correlation between examinee proficiency and omission propensity ($cor(\theta, \xi)$) suggests that a low-performing group of examinees tends to omit more items. In addition, the positive correlation between examinee speed and omission propensity ($cor(\tau, \xi)$) indicates that examinees with a fast responding rate tends to omit more items and generate this nonresponse behavior much faster ($cor(\tau, \eta)$). Further, examinee's tendency to omit an item is highly associated with examinee's omission speed η ($cor(\xi, \eta) = 0.82$; in other words, examinees who tend to omit an item do it with fast pace. Lastly, correlations between omission propensity and other person parameters: proficiency, speed, and omission speed are non-zero. In other words, parameters related to item responses, RTs, and NRTs are not distinct from the missingness mechanism, which indicates nonignorability of the missing responses (Ulitzsch et al., 2020).

Table 16 presents the item parameter estimates from the SA+O model. The easier items (e.g., item 1 and 2) tend to show lower time intensity, while harder items (e.g., item 7 and 8) tend to show higher time intensity β and lower omission difficulty ν, suggesting that examinees tend to exhibit omission behavior on harder and more time intense items. Time intensity and omission time intensity δ parameters are highly correlated ($cor(\beta, \delta) = 0.83$); however, it is not a perfect linear relationship, indicating that examinees operate different speed on generating a response or nonresponse, given different item characteristics.

5.3.2 Comparing Ad Hoc and SA+O Estimates

Item parameter estimates and 95% confidence intervals retrieved from ad hoc approaches and model-based approach are listed in Table 17-18. For the SA+O model, the credible intervals which are analogous to confidence intervals were estimated. The credible interval differs from the confidence interval that the credible intervals treat their bounds as fixed and the estimated parameter as a random variable, while the confidence intervals treat their bounds as random variables and the parameter as a fixed value. Item discrimination parameter *a* and item difficulty parameter *b* estimates are ordered by magnitude of the SA+O model, respectively. The item discrimination parameter estimates for the incorrect approach tend to be larger than those for the SA+O model, ranging from 0 to 0.52. On the other hand, the *a*s from the ignored approach tend to be smaller than that from the SA+O model, ranging from 0.07 to 0.40, except the item 9 (i.e., the least discriminating item). As it is shown in the simulation study, as item discrimination parameter values increase, the CI for SA+O model becomes wider than ad hoc

approaches. No pattern is shown for the 5-second rule; the difference on *a*s between the SA+O model and 5-second rule is either positive or negative.

In general, the CI for item difficulty parameter estimates for the SA+O model is narrower than that from the ad hoc approaches. Compared with the SA+O model, the item difficulty parameter estimates for incorrect approach are greater, while that for the ignored approach are smaller-this trend is in line with the results from the simulation study. For instance, the largest differences (e.g., 0.55) occur for estimates further from zero. Overall, the *b*s from the 5-second rule follow closely to that from the incorrect approach, while the values are smaller throughout all items.

Summary statistics for the person proficiency parameters are given in Table 19. The means and standard deviations (SD) of examinees' proficiency estimates are very close across different approaches. Compared with the SD from the SA+O model, the proficiency estimates from the incorrect and 5-second rule approach are more spread out from the mean, while that from the ignored approach are more concentrated.

Table 20 and Table 21 present the mean and standard deviation for person proficiency estimates conditioning on raw scores and number of omissions, respectively. The raw scores are divided into three groups: high performers ($\gamma_{raw} \geq 7$), average performers ($3 \leq \gamma_{raw} \leq 6$) and low performers ($\gamma_{raw} \leq 2$). With high performers, there is no noticeable difference on person proficiency estimates between ad hoc vs. model-based approaches. However, with average and low performers, SD for ignored approach is substantially higher than the other methods and the difference on SD is about 0.09 and 0.16, respectively. Further, θ estimates for average performers were less than that for the

other methods; however, θ estimates for low performers were greater than that for the other methods.

Similarly, the number of omissions are divided into four categories: zero omission (O_{num}= 0), low omissions ($1 \leq O_{num} \leq 3$), medium omissions ($4 \leq O_{num} \leq 6$), and high omissions ($7 \leq O_{num} \leq 9$). The sample size for each category varies substantially such as 183, 139, 35, 26 (17 for ignored approach, since if examinees generate all omitted responses, that examinee is excluded). Under the zero omission condition, the SD for the SA+O model are the smallest, while that for the ignored approach is the largest and mean for the ignored approach is considerably smaller than the other methods. When the number of omission increases, in particular larger than 4, there is a substantial difference in mean and SD between ad hoc vs. model based approach. For instance, the mean and SD for SA+O model are considerably larger than that for incorrect and 5-second rule (e.g., the difference is 0.18 and 0.21, respectively). Further, when omission rate is high ($7 \leq O_{num} \leq 9$), the θ estimates for ignored approach are substantially lower than that for the other methods and its variability is high (0.61).

In summary, the item parameter estimates for the incorrect approach and 5-second rule tend to be greater than that for the SA+O model, whereas that for the ignored approach tend to be smaller than that for the SA+O mode. This is in line with the results from the simulation study. In terms of person parameter estimates, the ignored approach shows a substantially high variability with average and low performers, compared with the other methods. Further, as the number of omission increases, there is a substantial difference in mean and SD between ad hoc approaches and SA+O model.

Table 15: Person Parameter Variances, Correlations and Credible Intervals

	θ	τ	ξ	η
θ	1.00			
τ	-0.33	0.19		
	[-0.45, -0.21]	[0.16, 0.22]		
ξ	-0.65	0.38	6.65	
	[-0.76, -0.53]	[0.25, 0.49]	[5.06, 8.63]	
η	-0.43	0.66	0.82	1.29
	[-0.58, -0.28]	[0.55, 0.75]	[0.73, 0.88]	[1.01, 1.65]

Note. 95% credible intervals are provided in square brackets. θ = proficiency; τ = speed; ξ = omission propensity; η = omission speed.

Table 16: Item Parameter Estimates and Credible Intervals from SA+O Model

Item	Item Parameter Estimates						
	a	b	β	ν	δ	α	ω
1	1.49 [1.04, 2.02]	-0.71 [-1.03, -0.42]	3.86 [3.79, 3.92]	4.20 [3.62, 4.84]	4.41 [4.08, 4.74]	2.25 [2.45, 2.07]	1.35 [1.80, 1.02]
2	2.65 [1.83, 3.78]	-0.69 [-1.14, -0.27]	4.36 [4.31, 4.42]	2.87 [2.41, 3.37]	4.41 [4.19, 4.66]	2.73 [3.02, 2.48]	2.50 [3.35, 1.89]
3	1.56 [1.03, 2.17]	-0.39 [-0.71, -0.07]	4.80 [4.74, 4.85]	2.24 [1.80, 2.71]	4.79 [4.58, 5.01]	3.65 [4.14, 3.25]	2.75 [4.55, 1.88]
4	1.61 [1.10, 2.23]	-1.09 [-1.45, -0.74]	5.14 [5.08, 5.20]	2.41 [1.97, 2.89]	5.11 [4.90, 5.33]	2.56 [2.82, 2.32]	3.12 [5.03, 2.13]
5	1.31 [0.92, 1.77]	-0.05 [-0.33, 0.23]	3.84 [3.77, 3.90]	3.99 [3.43, 4.61]	4.08 [3.75, 4.39]	2.17 [2.35, 2.00]	1.34 [1.76, 1.02]
6	1.81 [1.27, 2.50]	0.30 [-0.04, 0.67]	4.14 [4.08, 4.20]	2.74 [2.29, 3.22]	4.01 [3.74, 4.28]	2.49 [2.74, 2.27]	1.29 [1.61, 1.04]
7	1.75 [1.17, 2.47]	1.50 [1.05, 2.05]	4.69 [4.63, 4.75]	2.06 [1.64, 2.52]	4.67 [4.43, 4.91]	3.04 [3.39, 2.75]	1.46 [1.84, 1.18]
8	2.34 [1.61, 3.32]	0.66 [0.23, 1.16]	4.44 [4.36, 4.51]	1.75 [1.32, 2.20]	4.02 [3.80, 4.26]	1.86 [2.03, 1.71]	1.39 [1.43, 0.78]
9	0.77 [0.47, 1.09]	0.18 [-0.06, 0.42]	3.73 [3.66, 3.80]	4.44 [3.86, 5.08]	3.92 [3.52, 4.32]	2.01 [2.18, 1.85]	1.05 [1.67, 1.16]

Note. 95% credible intervals are provided in square brackets. a = item discrimination; b = item difficulty; β = time intensity; ν = omission difficulty; δ = omission time intensity; α = time discrimination; ω = omission time discrimination.

Table 17: Item Discrimination Estimates and 95% CI

Item Discrimination Parameter (a)				
Item	SA+O	Incorrect	Ignored	5” Rule
9	0.77 [0.47, 1.09]	0.78 [0.49, 1.06]	0.82 [0.50, 1.14]	0.71 [0.43, 0.99]
5	1.31 [0.92, 1.77]	1.36 [0.98, 1.75]	1.17 [0.78, 1.56]	1.31 [0.93, 1.68]
1	1.49 [1.04, 2.02]	1.41 [1.01, 1.82]	1.42 [0.95, 1.90]	1.35 [0.95, 1.75]
3	1.56 [1.03, 2.17]	1.94 [1.41, 2.47]	1.41 [0.91, 1.91]	1.93 [1.39, 2.47]
4	1.61 [1.10, 2.23]	2.13 [1.53, 2.72]	1.36 [0.87, 1.86]	2.07 [1.48, 2.66]
7	1.75 [1.17, 2.47]	1.81 [1.25, 2.36]	1.57 [0.99, 2.15]	1.76 [1.20, 2.31]
6	1.81 [1.27, 2.50]	1.81 [1.32, 2.31]	1.54 [1.02, 2.05]	1.69 [1.21, 2.18]
8	2.34 [1.61, 3.32]	2.18 [1.56, 2.80]	2.25 [1.56, 2.80]	2.09 [1.48, 2.70]
2	2.65 [1.83, 3.78]	2.90 [2.00, 3.80]	2.25 [1.41, 3.08]	2.85 [1.94, 3.76]

Table 18: Item Difficulty Estimates and 95% CI

Item Difficulty Parameter (b)				
Item	SA+O	Incorrect	Ignored	5” Rule
4	-1.09 [-1.45, -0.74]	-0.54 [-0.90, -0.18]	-1.23 [-1.59, -0.87]	-0.54 [-0.89, -0.18]
1	-0.71 [-1.03, -0.42]	-0.52 [-0.80, -0.23]	-0.83 [-1.15, -0.51]	-0.54 [-0.82, -0.26]
2	-0.69 [-1.14, -0.27]	-0.24 [-0.67, -0.19]	-0.90 [-1.34, -0.47]	-0.30 [-0.72, 0.13]
3	-0.39 [-0.71, -0.07]	0.12 [-0.20, 0.45]	-0.56 [-0.88, -0.24]	0.10 [-0.23, 0.42]
5	-0.05 [-0.33, 0.23]	0.08 [-0.16, 0.35]	-0.16 [-0.43, 0.11]	0.04 [-0.23, 0.31]
9	0.18 [-0.06, 0.42]	0.29 [0.06, 0.52]	0.13 [-0.11, 0.37]	0.25 [0.02, 0.48]
6	0.30 [-0.04, 0.67]	0.55 [0.22, 0.87]	0.09 [-0.22, 0.41]	0.48 [0.16, 0.79]
8	0.66 [0.23, 1.16]	1.07 [0.67, 1.47]	0.41 [-0.02, 0.83]	1.00 [0.61, 1.40]
7	1.50 [1.05, 2.05]	1.78 [1.33, 2.23]	1.26 [0.84, 1.68]	1.74 [1.30, 2.19]

Table 19: Summary Statistics for Person Proficiency Estimates

		SA+O	Incorrect	Ignored	5" Rule
θ	Mean	0.00	0.00	0.00	0.00
	SD	0.86	0.89	0.84	0.88

Table 20: Summary Statistics for Person Proficiency Estimates Conditioning on Raw Scores

	High Performers ($\gamma_{raw} \geq 7$)				Average Performers ($3 \leq \gamma_{raw} \leq 6$)				Low Performers ($\gamma_{raw} \leq 2$)			
	SA+O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec	SA+O	Inc	Ign	5 sec
Mean	1.11	1.17	1.05	1.17	0.13	0.13	0.06	0.12	-0.93	-0.96	-0.85	-0.95
SD	0.34	0.35	0.36	0.35	0.33	0.34	0.42	0.35	0.36	0.35	0.50	0.34

Note. γ_{raw} = raw scores

Table 21: Summary Statistics for Person Proficiency Estimates Conditioning on Number of Omissions

	O_{num}= 0			$1 \leq O_{num} \leq 3$			$4 \leq O_{num} \leq 6$			$7 \leq O_{num} \leq 9$		
	Mean	SD	N_{samp}	Mean	SD	N_{samp}	Mean	SD	N_{samp}	Mean	SD	N_{samp}
SA+O	0.50	0.73	183	-0.27	0.63	139	-0.65	0.64	35	-1.21	0.44	26
Inc	0.56	0.75	183	-0.29	0.64	139	-0.83	0.43	35	-1.27	0.25	26
Ign	0.35	0.81	183	-0.33	0.75	139	-0.42	0.69	35	-0.15	0.61	17
5 sec	0.55	0.76	183	-0.31	0.65	139	-0.82	0.45	35	-1.13	0.30	26

Note. O_{num} = number of omissions; N_{samp} = sample size

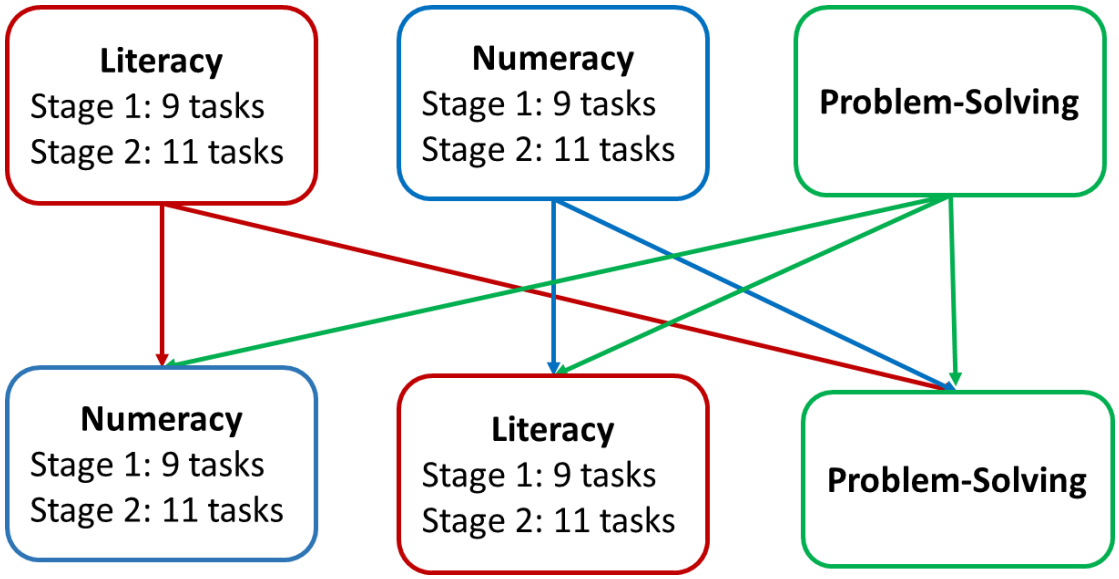


Figure 12: PIAAC Main Study Assessment Design

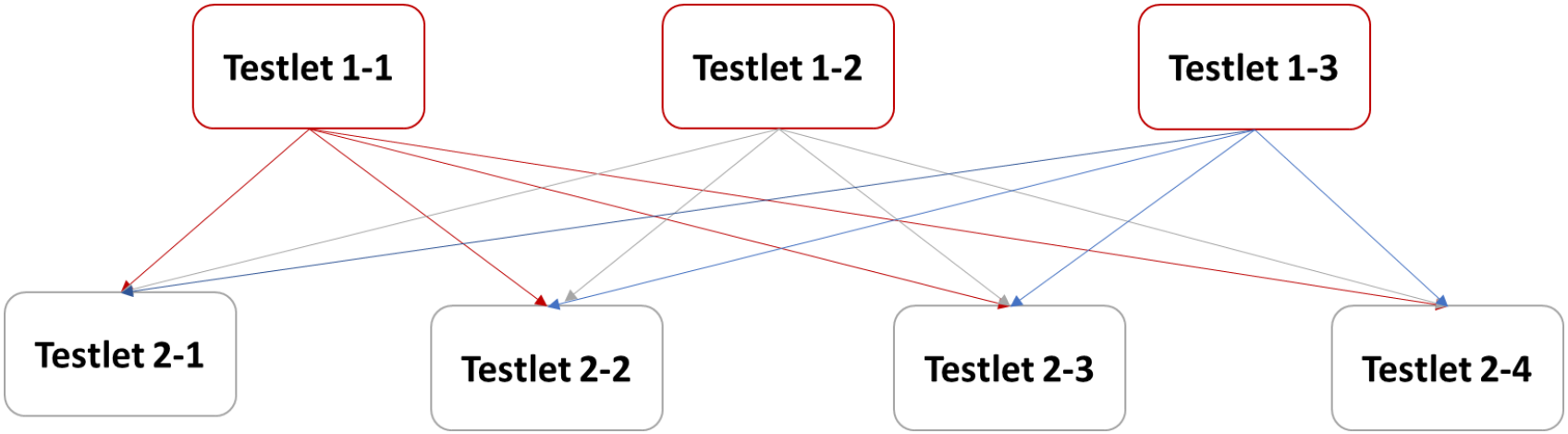


Figure 13: PIAAC Multistage Adaptive Testing Design for Literacy and Numeracy

CHAPTER 6

DISCUSSION

In this chapter, the main ideas from the previous chapters are summarized and discussed in greater detail. The chapter consists of three sections. First, a summary of the main findings from the simulation study and empirical application. Next, the implications of this study for practitioners will be presented. Following that, the limitations of this study design and the possible directions for future research will be discussed.

6.1 Summary

The purpose of the present research is to investigate the impact of how omitted responses are handled on item and person parameter estimates with the ad hoc and the model-based approaches. Although the simultaneous modeling of responses and RT as well as nonresponses and NRT has been proposed (e.g., SA+O model; Ulitzsch et al., 2020), no studies to date have compared this model-based approach with the current treatments on omitted items in LSA. This study specifically focuses on the performance of the ad hoc approaches vs. the model-based approach with the presence of omitted items. To that end, this study addresses the following research questions: In the context of LSA, how do the ad hoc and the model-based approaches for handling omitted responses compare in terms of item and person parameter estimation in IRT? And in real data analyses, is there a practical difference between the ad hoc and the model-based approaches to handling omitted responses?

Simulation studies were carried out to verify the implementation and compare the performance of the ad hoc vs. the model-based approach. A simulation study varied by four different design factors with two levels each fully-crossed: sample sizes, number of

items, omission rates, and correlation structure between omission speed vs. proficiency, speed, and omission propensity. The estimates were compared in terms of bias and efficiency. All estimation for the ad hoc approaches was performed using the EM algorithm, while that for the SA+O model was performed using the Bayesian estimation. Notably, the convergence rate of the SA+O model was only 75% out of 100 replications under the most challenging condition with a small sample size, a short test length, and a low omission rate. Further, the convergence rates were under 90%, when an omission rate was low. Throughout all simulation conditions, item and person parameter recovery followed closely to that from the previous study (Ulitzsch et al., 2020).

First, the simulation results for item discrimination estimates from the ad hoc and the model-based approach were relatively similar across conditions, except for the conditions with a high omission rate. When the omission rate was high, the incorrect approach and 5-second rule showed upwards bias, while the ignored approach showed downwards bias. The difference in variability between the ad hoc vs. the model-based approach was largest for the smallest item discrimination parameter value, but that difference decreased as the item discrimination parameter value increased.

The item difficulty parameter values were well recovered throughout all conditions with the SA+O model without systematic bias. On the other hand, the incorrect approach and 5-second rule showed upwards bias, while the ignored approach showed downwards bias under the conditions with a high omission rate. Interestingly, as item difficulty values increased, the bias decreased for the incorrect approach and 5-second rule. Further, the SD for the SA+O model was smaller than that for the ad hoc approaches when the omission rate was high. In particular, the difference in the ad hoc

vs. the model-based approach was largest for the smallest and largest item difficulty values and lowest for the mid item difficulty value.

The impact of bias in item discrimination and difficulty parameter on the proficiency estimates can be a more important concern. Overall, proficiency estimates for both the model-based and the ad hoc approaches were underestimated for a high-performing group (i.e., $\theta \geq 1$) and were overestimated for a low-performing group (i.e., $\theta \leq -1$). This phenomenon is caused by the EAP estimator and is known as shrinkage (Tong & Kolen, 2010). There was almost no difference on MAE values among the ad hoc approaches and the SA+O model with high performers; however, that difference became prominent with average and low performers under the conditions with a high omission rate. Especially, the ignored approach showed the highest MAE values throughout all conditions with low performers. Similarly, the SD for person proficiency estimates decreased as test length increased for both the ad hoc and the model-based approaches. In particular, under the conditions with a longer test length, the SD showed the U-shaped pattern-that is, the SD was relatively large for high and low-performing groups, compared to that for an average performing group (e.g., $-1 \leq \theta \leq 1$) for both the ad hoc and the model-based approaches. Again, this result may reflect the lack of examinees' response patterns at the proficiency extremes.

Further, conditioning on omission propensity, the EAPs for the incorrect and the 5-second rule noticeably underestimate person proficiency estimates with high omission propensity performers ($\xi > 2.5$), while that for an ignored approach overestimates θ, under the conditions with a longer test length and high omission rates. In short, when omitted responses were ignored under the conditions with a longer test length, high

omission rate, and high omission propensity performers, the incorrect and the 5-second rule can substantially underestimate person proficiency estimates, while the ignored approach can overestimate θ.

Finally, an empirical data analysis of the Chilean sample of PIAAC 2012 demonstrated that the choice of handling omitted responses affects both item and person parameter estimates. In general, the item discrimination parameter estimates from the incorrect approach tend to be larger than that from the SA+O model, while the *a*s from the ignored approach tend to be smaller. Overall, the CI for item difficulty parameter estimates from the SA+O model is narrower than that from the ad hoc approaches. Compared with the SA+O model, the incorrect approach tends to overestimate the *b*s, while the ignored approach tends to underestimate the *b*s. The overall trend of item parameter estimates is in line with the results from the simulation study.

In terms of the person parameter estimates, the mean and SD for person parameter estimates conditioning on raw scores and number of omissions show interesting differences. With average and low performers, SD for the ignored approach is substantially higher than for the other methods. In addition, the ignored approach underestimates θ for average performers, while overestimating θ for low performers, compared with the other methods. Further, when the number of omissions increases, in particular larger than 4, there is a substantial difference in mean and SD of θ between the ad hoc vs. the model based approach.

In summary, if there is a high percentage of omits and these are related to proficiency, the SA+O model performs better than the ad hoc approaches in terms of item parameter estimation; however, when the omission rate is small and the test length is

long, there is not much to be gained in terms of both bias and RMSE from using the SA+O model. In terms of the person parameter estimates, test-takers with low proficiency are most affected by the different approaches, and the SA+O model doesn't perform better than the ad hoc approaches even with the conditions under a high omission rate.

6.2 Implications of Findings

Comparing the current approaches (i.e., ad hoc approaches) on handling omitted responses in LSA with the recently proposed model-based approach has important implications for educational measurement practice. Omitted responses, prevalent in low-stakes tests, need to be handled appropriately. Otherwise, inappropriate treatment of omitted responses might lead to biased item and person parameter estimates, group statistics, and further result in difference country rankings. There are several prominent findings that can help practitioners choose how to handle omitted responses.

First, the item discrimination and difficulty parameter estimates from the ad hoc approaches are considerably biased under the conditions with a high omission rate. More specifically, the incorrect approach and 5-second rule showed upwards bias, while the ignored approach showed downwards bias. It is interesting to note that the difference in variability between the ad hoc vs. the model-based approach was largest for the smallest item discrimination parameter value, but that difference decreased as the item discrimination parameter value increased. However, for the item difficulty parameter estimates, the variability of the ignored approach follows closely to that of SA+O model, while that of the incorrect and 5-second rule shows largest variability at the extreme values (e.g., b = -1 or 1). This leads to the second implication.

Under the high omission rate, the bias across all items for item discrimination parameters under the incorrect approach is 0.25, while that for item difficulty is 0.35. And this can lead to biased proficiency estimates. There was almost no difference on MAE values among the ad hoc approaches and the SA+O model with high performers; however, that difference became prominent with average and low performers. Further, when omitted responses were ignored under the conditions with a longer test length, high omission rate, and high omission propensity performers, the incorrect and 5-second rule can substantially underestimate person proficiency estimates, while the ignored approach can overestimate θ.

In summary, the SA+O model is preferable especially under the conditions with a high omission rate (at least 17%) and when proficiency is related to omission propensity. However, the convergence rate of the SA+O model is not optimal for the conditions with a small sample size ($N < 750$), short test length ($I < 30$) and low omission rate ($O < 17\%$). The computation time for the SA+O model is also considerably longer than that for the ad hoc approaches. When researchers lack expertise in Bayesian estimation, the ad hoc approaches might be a reasonable choice. Further, most of the advantages gained by the SA+O model is due to the magnitude of the (negative) correlation between proficiency and omission propensity. This factor was not manipulated in the simulation study. To the extent that this correlation tends towards zero, the SA+O model is going to be less useful. Given the limitations of SA+O model, it is up to the practitioners to decide which approach to use, depending on the available resources at hand. Thus, this study can provide valuable information on the advantages and disadvantages of each approach in terms of item and person parameter estimates to the practitioners.

6.3 Future Directions

Recently, the SA+O model has been introduced for dealing with omitted responses by simultaneously modeling item responses and RTs as well as item nonresponses and NRTs. However, no study has examined the performance of SA+O model with the current ad hoc approaches handling with omitted items in LSA. To fill this gap, this study investigated the impact of omitted responses on item and person parameter estimates with the ad hoc and the model-based approaches. Although the scope of this study is broad, there is great room for improvement.

First, future efforts can focus on when the lognormal model for RT and NRT distributions violate the normality assumption. Previous research demonstrated that the RT distribution differed substantially across items (Ranger & Kuhn, 2012) and proposed the linear transformation model to deal with this problem (Wang, Chang & Douglas, 2013). Since the RT and NRT distributions might vary across items, including gamma, exponential, and Weibull, the performance of the SA+O model and the ad hoc approaches can be compared under such conditions.

Second, a more realistic condition can also be considered where omitted and not-reached items occur at the same time. This is a practical issue, since empirical data set would be more likely to contain both omitted and not-reached items, and LSA treat omitted and not-reached items differentially. Further, Pohl at al. (2019) argued that the SA+O model should be able to account for not-reached items as well because it controls for both general working speed and omission speed. Thus, it would be interesting to compare the performance of the SA+O model with the ad hoc approaches on the data set, which contains both omitted and not-reached items.

Third, the violation of constant working speed and homogeneous omission strategies across examinees can be explored. The SA model assumes that the working speed of an examinee is constant across the items; however, given a limited time, it is reasonable to assume that examinees vary their working speed to finish the test in time. In addition, examinees can change omission strategies across the items or different groups of examinees employ distinctive omission strategies. To deal with each case, the variable working speed model (Fox & Marianti, 2016) and mixture modeling (Molenaar et al., 2016) was proposed, respectively.

Other potential research topics include comparisons of country-level proficiency estimates using SA+O model as this one of the ultimate estimates in LSAs, the NAEP approach (i.e., partially correct), and other simulation conditions, not studied in Ulitzsch et al. (2019). Research on these topics would provide valuable insights to practitioners and policymakers.

APPENDIX

Tables, A1 through A14 present the bias and RMSE for the omission difficulty, time intensity, omission time intensity, time discrimination and omission time discrimination parameter estimates. In general, the item parameter estimates from the SA+O model follow closely to those in Ulitzsch et al. (2020). For instance, under the condition with a small sample size, a short test length, a low omission rate and $\rho_{.\eta} \neq 0$, omission time intensity parameters δ and omission time discrimination parameters ω show downwards bias and the corresponding RMSEs are substantial. Further, the RMSE for omission difficulty parameters ν and time discrimination parameters α increase as true parameter values increase. However, the RMSE rapidly decreases with an increasing omission rate, test length, and sample size.

Figures such as A15 through A19 present depict the medians and 90% ranges of the omission difficulty, time intensity, omission time intensity, time discrimination and omission time discrimination parameter estimates. N denotes number of examinees, and I, number of items, respectively. The x-axis indicates true item parameters and y-axis indicates the bias. Solid lines represent $\rho_{.\eta} \neq 0$ conditions, while dashed lines represent $\rho_{.\eta} = 0$ conditions. Grey lines represent a 5% omission rate, while black lines represent 17% omission rate. Lastly, the grey dashed horizontal line indicates unbiased estimation.

The variability of omission difficulty ν parameters increase as true values increase under the conditions with a small omission rate. A19 shows the substantial difference on the variability of omission time intensity δ parameters between a small and large omission rate under the conditions with a small sample size and a short test length. Further, omission time discrimination estimates ω with larger true values show substantial downwards bias under the conditions with a

small sample size. Throughout all simulation conditions, the variability of item parameter estimates decreases with an increasing sample size and test length.

Figures such as A.20 and A.21 show medians and 90% ranges of person parameter variance and correlation estimates-that is, the posterior distribution mean of each parameter across all conditions and replications. The black solid and dashed line indicates the omission rate of 5% and 17%, respectively. The grey dashed horizontal line indicates the true parameter value. All the median person parameter variance estimates are close to zero (i.e., unbiased), except an omission propensity ξ illustrating upwards bias especially under the conditions with a small sample size (N=375) and a low omission rate (5%). Further, 90% ranges of omission propensity ξ and omission speed η are substantially wide under the least favorable condition (e.g., N=375, I=10, O=5%). In addition, 90% ranges of omission propensity ξ and omission speed η under the conditions with a high omission rate are much narrower than that with a low omission rate. Throughout all conditions, as sample size, number of items, and omission rate increase, the bias of person parameter variance estimates decreases rapidly.

Largely, all the median person parameter correlation estimates are close to zero (i.e., unbiased) under conditions with a high omission rate, except for that of $cor(\xi, \eta)$. However, under the most challenging condition (e.g., N=375, I=10, O=5%) with $\rho_{.\eta} \neq 0$, correlations with omission speed η (e.g., $cor(\theta, \eta), cor(\tau, \eta), cor(\xi, \eta)$) show substantial differences between estimated and true variance estimates. For instance, the person parameter correlation estimates between omission propensity and speed ranges from 0.60 to 0.9 where the true correlation value is 0.85.

Tables A.11 through A.14 present the bias and RMSE for the person parameter variance and correlation estimates. The bias for speed parameter variance estimate is zero throughout all

conditions and its RMSE is also close to zero. However, the bias and RMSE for omission propensity and omission speed variance estimates are relatively larger under the conditions with $\rho_{.\eta} = 0$, which is an extremely unfavorable condition than that under $\rho_{.\eta} \neq 0$ where omission speed is related to proficiency, speed, and omission propensity. Further, under the conditions with a high omission rate, large sample size, and longer test length, variance estimates have less bias and their RMSEs are smaller than that with a low omission rate, small sample size, and shorter test length. All the bias values are acceptably small, except for the bias of correlation estimate between omission propensity ξ and omission speed η. Further, all the RMSE values for cor(ξ, η) are noticeably larger under the conditions with $\rho_{.\eta} = 0$. However, all the RMSE values for the person parameter correlation estimates are acceptably small under the conditions with $\rho_{.\eta} \neq 0$, except for the condition with small sample size (N=375), short test length (I=10), and low omission rate (5%).

A.1: Bias for Omission Difficulty Parameters

				True Parameter									
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	2.00	2.50	3.00	3.50	4	4.25	4.75	5.25	5.75	6.25
$\rho_{\cdot\eta} = 0$	375	10	5						0.21	0.26	0.27	0.38	0.38
			17	0.05	0.04	0.04	0.04	0.06					
		30	5						0.12	0.14	0.15	0.20	0.25
			17	0.05	0.04	0.04	0.05	0.08					
	750	10	5						0.05	0.04	0.03	0.07	0.05
			17	0.01	0.02	0.03	0.03	0.03					
		30	5						0.08	0.08	0.09	0.12	0.13
			17	0.01	0.01	0.02	0.02	0.02					
$\rho_{\cdot\eta} \neq 0$	375	10	5						0.21	0.27	0.27	0.37	0.40
			17	0.04	0.03	0.03	0.03	0.05					
		30	5						0.12	0.14	0.16	0.18	0.26
			17	0.04	0.03	0.03	0.04	0.07					
	750	10	5						0.03	0.01	0.01	0.03	0.04
			17	0.00	0.02	0.02	0.02	0.03					
		30	5						0.05	0.04	0.05	0.07	0.10
			17	0.01	0.01	0.02	0.02	0.02					

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.2: RMSE for Omission Difficulty Parameters

				True Parameter									
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	2.00	2.50	3.00	3.50	4	4.25	4.75	5.25	5.75	6.25
$\rho_{\cdot\eta} = 0$	375	10	5						0.45	0.49	0.54	0.65	0.80
			17	0.23	0.26	0.27	0.28	0.30					
		30	5						0.34	0.39	0.43	0.50	0.58
			17	0.22	0.23	0.24	0.26	0.28					
	750	10	5						0.27	0.30	0.33	0.35	0.39
			17	0.16	0.17	0.19	0.19	0.21					
		30	5						0.24	0.25	0.28	0.34	0.38
			17	0.16	0.16	0.17	0.18	0.20					
$\rho_{\cdot\eta} \neq 0$	375	10	5						0.44	0.50	0.54	0.64	0.77
			17	0.22	0.26	0.27	0.28	0.30					
		30	5						0.35	0.39	0.43	0.46	0.57
			17	0.22	0.23	0.24	0.26	0.28					
	750	10	5						0.26	0.29	0.33	0.34	0.40
			17	0.16	0.17	0.19	0.19	0.21					
		30	5						0.22	0.24	0.27	0.32	0.36
			17	0.16	0.16	0.17	0.18	0.20					

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.3: Bias for Time Intensity Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	3.50	3.75	4.00	4.25	4.50
$\rho_{\cdot\eta} = 0$	375	10	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00
		30	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00
	750	10	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00
		30	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.00	0.01	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00
		30	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00
	750	10	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00
		30	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	0.00	0.00	0.00

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.4: RMSE for Time Intensity Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	3.50	3.75	4.00	4.25	4.50
$\rho_{\cdot\eta} = 0$	375	10	5	0.04	0.03	0.03	0.03	0.03
			17	0.04	0.03	0.03	0.03	0.02
		30	5	0.04	0.03	0.03	0.03	0.03
			17	0.04	0.03	0.03	0.03	0.02
	750	10	5	0.03	0.02	0.02	0.02	0.02
			17	0.03	0.03	0.02	0.02	0.02
		30	5	0.03	0.02	0.02	0.02	0.02
			17	0.03	0.02	0.02	0.02	0.02
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.04	0.03	0.03	0.02	0.02
			17	0.04	0.03	0.03	0.03	0.02
		30	5	0.04	0.03	0.03	0.03	0.03
			17	0.04	0.03	0.03	0.03	0.03
	750	10	5	0.03	0.02	0.02	0.02	0.02
			17	0.03	0.02	0.02	0.02	0.02
		30	5	0.03	0.02	0.02	0.02	0.02
			17	0.03	0.02	0.02	0.02	0.02

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.5: Bias for Omission Time Intensity Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	3.50	3.75	4.00	4.25	4.50
$\rho_{\cdot\eta} = 0$	375	10	5	-0.03	0.00	-0.01	-0.01	-0.02
			17	0.00	0.01	0.01	0.02	0.02
		30	5	-0.02	-0.03	-0.02	-0.02	-0.01
			17	0.01	0.00	0.00	0.01	0.01
	750	10	5	0.03	0.03	0.02	0.02	0.02
			17	0.01	0.00	0.00	0.00	0.00
		30	5	0.01	0.00	0.02	0.02	0.02
			17	0.00	0.00	0.00	0.00	0.00
$\rho_{\cdot\eta} \neq 0$	375	10	5	-0.09	-0.08	-0.07	-0.08	-0.09
			17	-0.03	-0.01	-0.02	-0.02	-0.02
		30	5	-0.03	-0.04	-0.03	-0.03	-0.02
			17	0.01	0.00	0.01	0.01	0.01
	750	10	5	-0.04	-0.06	-0.07	-0.06	-0.07
			17	0.00	-0.01	-0.01	-0.01	-0.01
		30	5	-0.02	-0.02	-0.01	-0.01	-0.01
			17	0.00	0.00	0.00	0.00	0.00

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.6: RMSE for Omission Time Intensity Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	3.50	3.75	4.00	4.25	4.50
$\rho_{\cdot\eta} = 0$	375	10	5	0.35	0.35	0.34	0.37	0.40
			17	0.15	0.15	0.15	0.15	0.15
		30	5	0.27	0.23	0.22	0.22	0.22
			17	0.14	0.12	0.11	0.11	0.11
	750	10	5	0.26	0.26	0.26	0.25	0.25
			17	0.11	0.10	0.10	0.09	0.10
		30	5	0.18	0.16	0.16	0.16	0.16
			17	0.10	0.09	0.09	0.08	0.08
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.29	0.28	0.29	0.31	0.29
			17	0.14	0.13	0.12	0.13	0.12
		30	5	0.24	0.20	0.19	0.19	0.20
			17	0.12	0.10	0.10	0.09	0.09
	750	10	5	0.20	0.20	0.20	0.20	0.21
			17	0.10	0.09	0.08	0.08	0.09
		30	5	0.17	0.14	0.14	0.14	0.13
			17	0.09	0.07	0.07	0.07	0.07

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.7: Bias for Time Discrimination Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	2.00	2.75	3.50	4.25	5.00
$\rho_{\cdot\eta} = 0$	375	10	5	0.00	0.00	0.00	0.00	0.03
			17	0.00	0.00	0.00	0.00	0.02
		30	5	0.00	0.00	0.00	0.00	0.01
			17	0.00	0.00	0.00	0.00	0.01
	750	10	5	-0.01	-0.01	0.00	-0.01	0.03
			17	-0.01	-0.01	0.01	-0.01	0.03
		30	5	0.00	0.00	0.00	0.00	-0.01
			17	0.00	0.00	-0.01	0.00	0.00
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.00	0.00	0.00	0.01	0.01
			17	0.00	-0.01	0.00	0.00	0.02
		30	5	0.00	0.00	0.00	0.00	0.02
			17	0.00	0.00	0.00	0.00	0.01
	750	10	5	-0.01	-0.01	0.01	-0.01	0.02
			17	-0.01	-0.01	0.01	-0.01	0.03
		30	5	0.00	0.00	0.00	0.00	0.00
			17	0.00	0.00	-0.01	0.00	0.00

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.8: RMSE for Time Discrimination Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	2.00	2.75	3.50	4.25	5.00
$\rho_{\cdot\eta} = 0$	375	10	5	0.08	0.11	0.16	0.18	0.23
			17	0.08	0.12	0.17	0.19	0.25
		30	5	0.08	0.11	0.13	0.16	0.20
			17	0.09	0.12	0.15	0.18	0.21
	750	10	5	0.06	0.08	0.10	0.12	0.16
			17	0.06	0.09	0.11	0.12	0.18
		30	5	0.05	0.07	0.10	0.12	0.14
			17	0.06	0.08	0.10	0.13	0.15
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.08	0.11	0.16	0.18	0.22
			17	0.08	0.12	0.17	0.19	0.26
		30	5	0.08	0.11	0.14	0.16	0.20
			17	0.09	0.12	0.15	0.18	0.21
	750	10	5	0.05	0.08	0.09	0.12	0.16
			17	0.06	0.09	0.11	0.12	0.18
		30	5	0.05	0.07	0.09	0.12	0.14
			17	0.06	0.08	0.10	0.13	0.15

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.9: Bias for Omission Time Discrimination Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	1.000	1.625	2.250	2.875	3.500
$\rho_{\cdot\eta} = 0$	375	10	5	-0.01	0.10	-0.04	-0.25	-0.64
			17	-0.01	0.00	-0.04	0.02	-0.04
		30	5	-0.01	-0.05	-0.06	-0.11	-0.33
			17	0.00	-0.02	-0.01	-0.04	-0.05
	750	10	5	-0.01	-0.01	-0.02	0.01	-0.12
			17	0.00	0.01	0.00	-0.03	-0.01
		30	5	0.00	-0.03	-0.01	-0.06	-0.07
			17	0.00	-0.01	-0.01	0.00	-0.03
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.00	0.03	-0.01	-0.12	-0.52
			17	-0.01	0.01	-0.05	0.02	-0.01
		30	5	-0.01	-0.04	-0.05	-0.12	-0.30
			17	0.00	-0.01	-0.02	-0.04	-0.05
	750	10	5	0.00	-0.01	0.01	0.01	-0.12
			17	0.00	0.00	0.00	-0.02	0.01
		30	5	0.00	-0.02	-0.02	-0.06	-0.06
			17	0.00	-0.01	-0.02	-0.01	-0.03

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.10: RMSE for Omission Time Discrimination Parameters

				True Parameter				
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	1.000	1.625	2.250	2.875	3.500
$\rho_{\cdot\eta} = 0$	375	10	5	0.17	0.60	0.72	0.99	1.55
			17	0.08	0.16	0.27	0.52	0.66
		30	5	0.16	0.32	0.72	1.14	1.80
			17	0.07	0.14	0.23	0.36	0.54
	750	10	5	0.12	0.27	0.52	1.01	1.26
			17	0.06	0.11	0.20	0.32	0.50
		30	5	0.10	0.20	0.41	0.69	1.20
			17	0.05	0.10	0.16	0.25	0.35
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.15	0.47	0.69	0.99	1.52
			17	0.08	0.15	0.26	0.48	0.66
		30	5	0.15	0.31	0.67	0.98	1.81
			17	0.08	0.14	0.22	0.35	0.54
	750	10	5	0.11	0.21	0.43	0.83	1.15
			17	0.05	0.10	0.18	0.30	0.52
		30	5	0.10	0.19	0.37	0.66	1.15
			17	0.05	0.10	0.16	0.24	0.35

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate.

A.11: Bias of Person Parameter Variances

$\rho_{.\eta}$	N	Item	Omitted (%)	var(τ)	var(ξ)	var(η)
$\rho_{.\eta} = 0$	375	10	5	0.00	1.26	0.12
			17	0.00	0.48	0.07
		30	5	0.00	0.75	0.08
			17	0.00	0.44	0.03
	750	10	5	0.00	0.31	0.08
			17	0.00	0.29	0.04
		30	5	0.00	0.48	0.04
			17	0.00	0.24	0.02
$\rho_{.\eta} \neq 0$	375	10	5	0.00	1.14	0.01
			17	0.00	0.36	-0.01
		30	5	0.00	0.60	0.00
			17	0.00	0.32	0.02
	750	10	5	0.00	0.21	-0.02
			17	0.00	0.21	0.00
		30	5	0.00	0.27	0.02
			17	0.00	0.18	0.02

Note. $\rho_{.\eta} = 0$ and $\rho_{.\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate; τ = speed; ξ = omission propensity; η = omission speed. Proficiency variance var(θ) was set to unity for model identification.

A.12: RMSE of Person Parameter Variances

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	var(τ)	var(ξ)	var(η)
$\rho_{\cdot\eta} = 0$	375	10	5	0.02	2.00	0.31
			17	0.02	1.09	0.17
		30	5	0.02	1.33	0.22
			17	0.02	0.90	0.12
	750	10	5	0.01	1.01	0.23
			17	0.01	0.72	0.12
		30	5	0.01	0.84	0.14
			17	0.01	0.53	0.09
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.02	1.88	0.30
			17	0.02	0.99	0.18
		30	5	0.02	1.20	0.21
			17	0.02	0.87	0.14
	750	10	5	0.01	1.04	0.20
			17	0.01	0.67	0.12
		30	5	0.01	0.79	0.15
			17	0.01	0.52	0.10

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate; τ = speed; ξ = omission propensity; η = omission speed. Proficiency variance var(θ) was set to unity for model identification.

A.13: Bias of Person Parameter Correlations

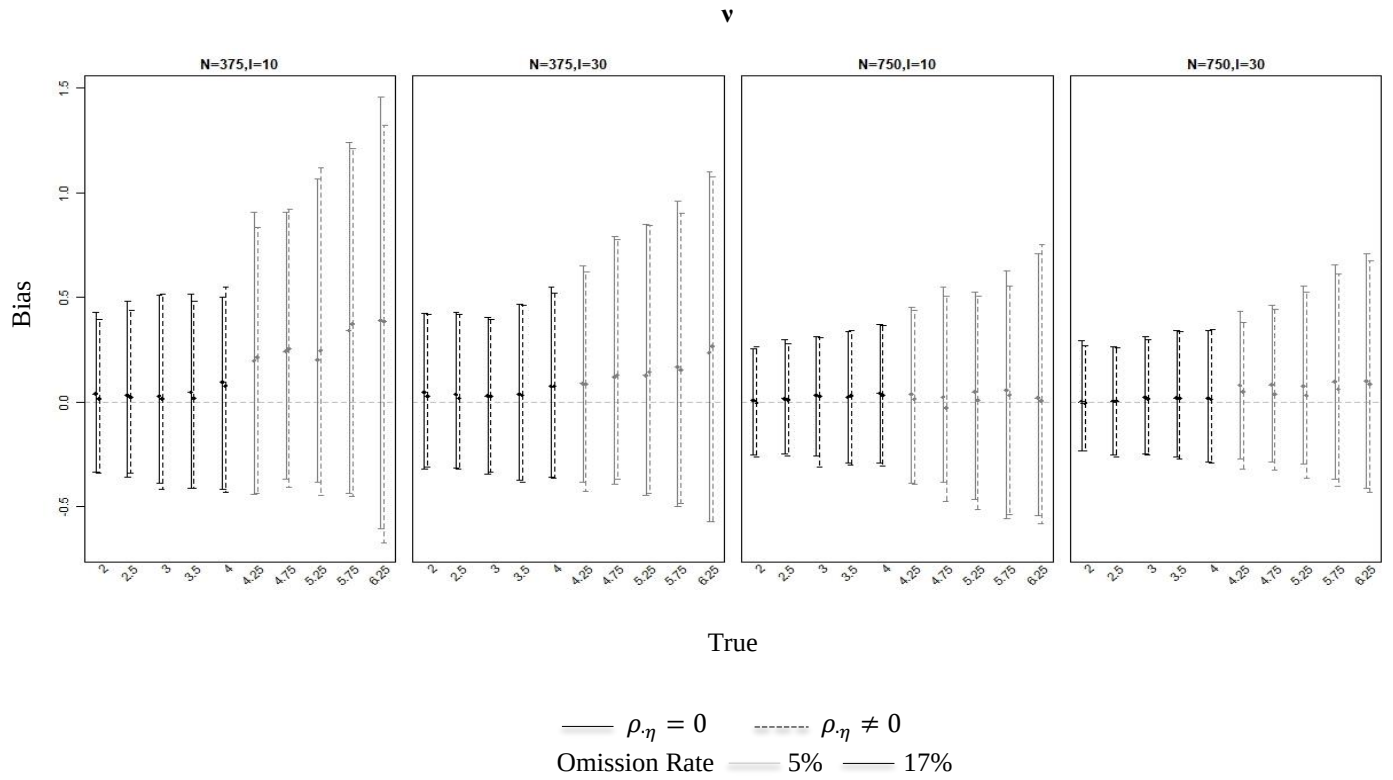
$\rho_{\cdot\eta}$	N	Item	Omitted (%)	cor(θ, τ)	cor(θ, ξ)	cor(θ, η)	cor(τ, ξ)	cor(τ, η)	cor(ξ, η)
$\rho_{\cdot\eta} = 0$	375	10	5	0.00	0.03	0.00	-0.01	0.01	-0.02
			17	0.00	0.00	0.00	0.00	0.00	0.00
		30	5	-0.01	0.00	0.02	-0.01	0.00	-0.02
			17	-0.01	0.00	0.00	0.00	0.01	0.00
	750	10	5	0.00	0.00	0.00	0.00	0.00	0.01
			17	0.00	0.00	0.00	0.00	-0.01	0.00
		30	5	0.00	0.00	0.00	0.00	-0.01	0.00
			17	0.00	-0.01	0.00	0.00	0.00	0.00
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.00	0.03	0.05	-0.02	-0.02	-0.08
			17	0.00	0.01	0.02	0.00	-0.01	-0.02
		30	5	0.00	0.00	0.02	-0.01	-0.02	-0.03
			17	0.00	0.00	-0.01	-0.01	-0.01	0.00
	750	10	5	0.00	0.00	0.02	0.00	-0.01	-0.03
			17	0.00	0.00	0.01	0.00	0.00	-0.01
		30	5	-0.01	0.00	0.00	0.00	0.00	-0.01
			17	-0.01	-0.01	0.00	0.00	0.00	0.00

Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate; τ = speed, ξ = omission propensity; η = omission speed.

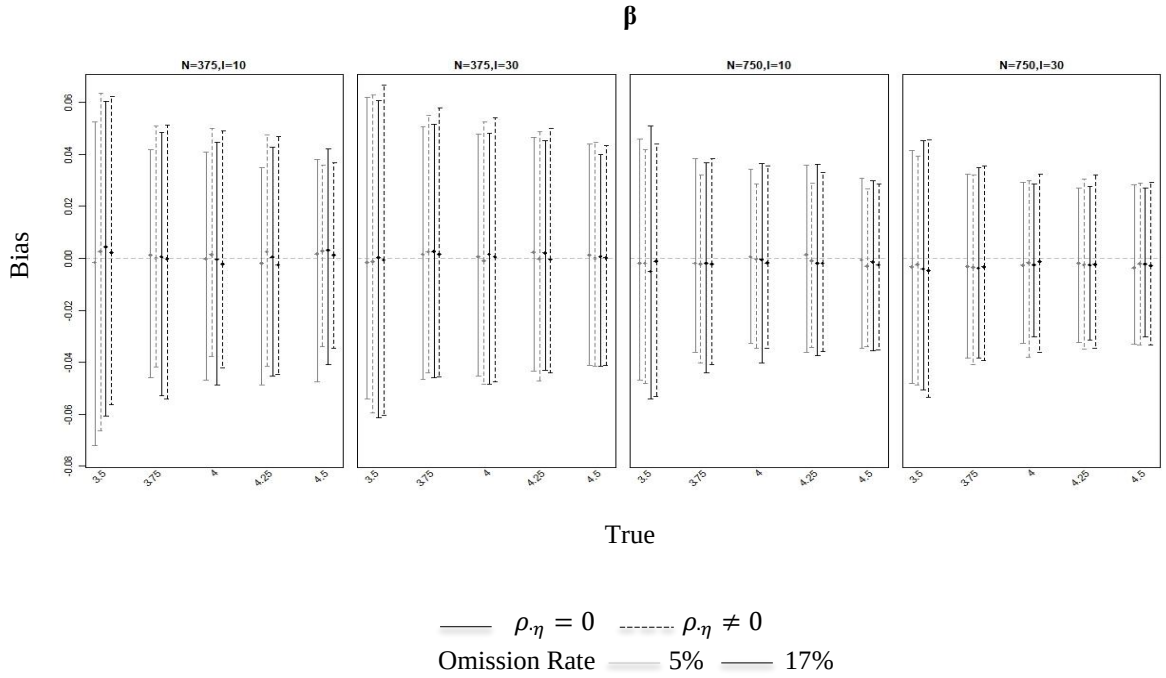
A.14: RMSE of Person Parameter Correlations

$\rho_{\cdot\eta}$	N	Item	Omitted (%)	cor(θ, τ)	cor(θ, ξ)	cor(θ, η)	cor(τ, ξ)	cor(τ, η)	cor(ξ, η)
$\rho_{\cdot\eta} = 0$	375	10	5	0.06	0.08	0.17	0.07	0.13	0.23
			17	0.06	0.07	0.11	0.05	0.08	0.11
		30	5	0.05	0.06	0.13	0.07	0.12	0.16
			17	0.05	0.04	0.08	0.06	0.08	0.09
	750	10	5	0.04	0.05	0.13	0.05	0.10	0.15
			17	0.04	0.04	0.08	0.04	0.07	0.09
		30	5	0.03	0.03	0.08	0.04	0.07	0.11
			17	0.03	0.03	0.07	0.03	0.05	0.08
$\rho_{\cdot\eta} \neq 0$	375	10	5	0.06	0.08	0.13	0.09	0.09	0.11
			17	0.06	0.06	0.08	0.07	0.05	0.04
		30	5	0.05	0.05	0.09	0.07	0.06	0.05
			17	0.05	0.05	0.06	0.06	0.04	0.02
	750	10	5	0.04	0.05	0.07	0.05	0.05	0.06
			17	0.04	0.04	0.05	0.04	0.03	0.03
		30	5	0.03	0.04	0.05	0.04	0.04	0.03
			17	0.03	0.03	0.04	0.04	0.03	0.02

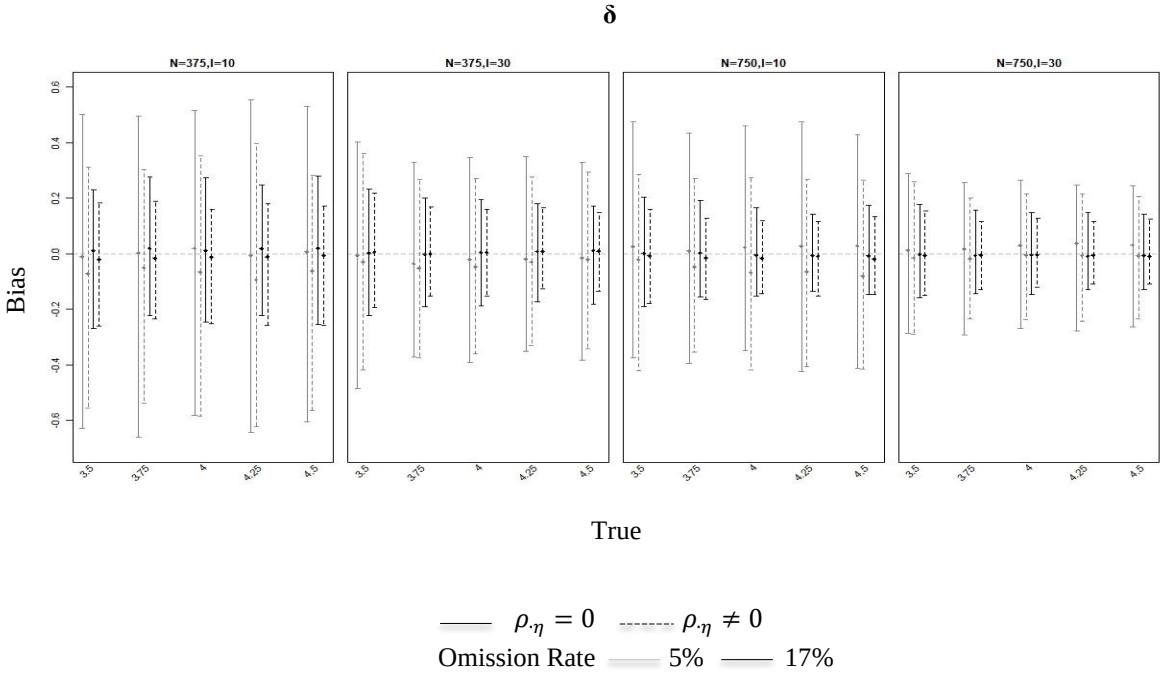
Note. $\rho_{\cdot\eta} = 0$ and $\rho_{\cdot\eta} \neq 0$ denote omission speed is uncorrelated or correlated with proficiency, speed, and omission propensity, respectively. N = number of examinees; Item = number of items; Omitted = omission rate; τ = speed, ξ = omission propensity; η = omission speed.



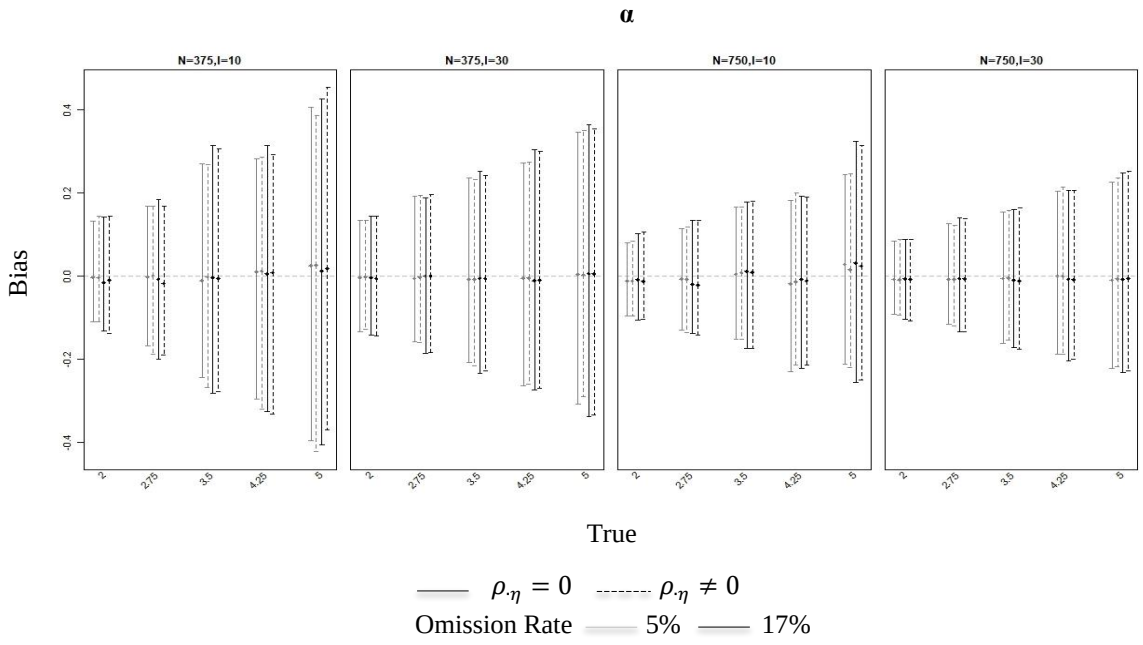
A.15: Medians and 90% Ranges of Differences between Estimated and True Omission Difficulty Parameters ν, Plotted against the True Parameters.



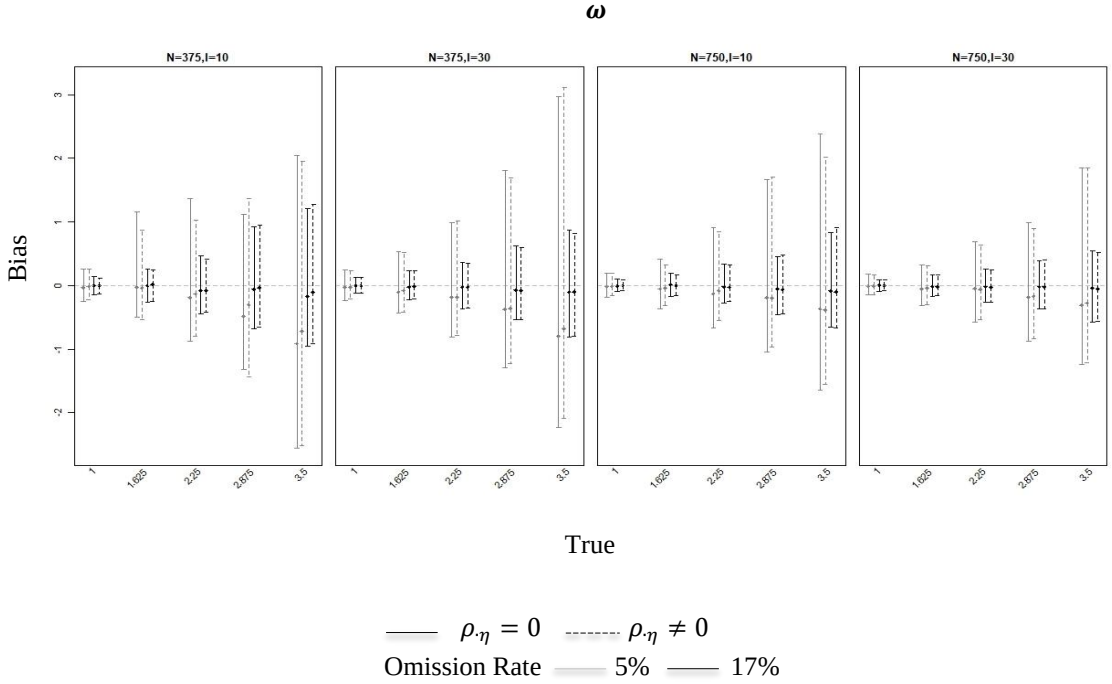
A.16: Medians and 90% Ranges of Differences between Estimated and True Time Intensity Parameters β, Plotted against the True Parameters.



A.17: Medians and 90% Ranges of Differences between Estimated and True Omission Time Intensity Parameters δ, Plotted against the True Parameters.

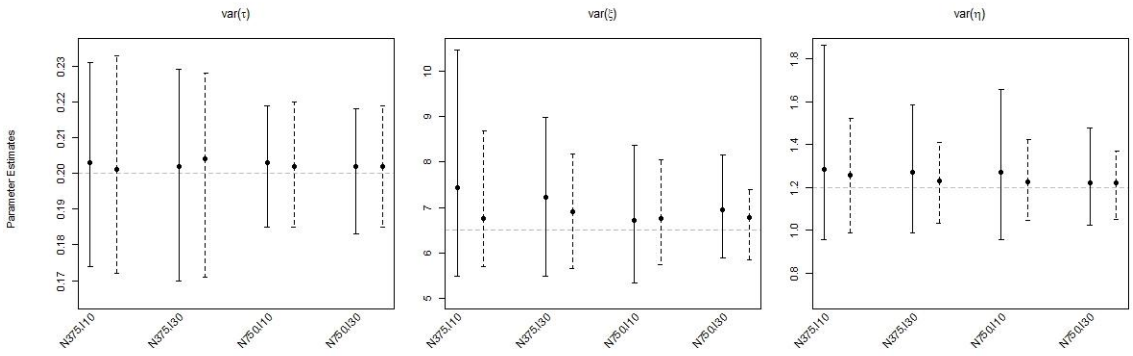


A.18: Medians and 90% Ranges of Differences between Estimated and True Time Discrimination Parameters α, Plotted against the True Parameters.

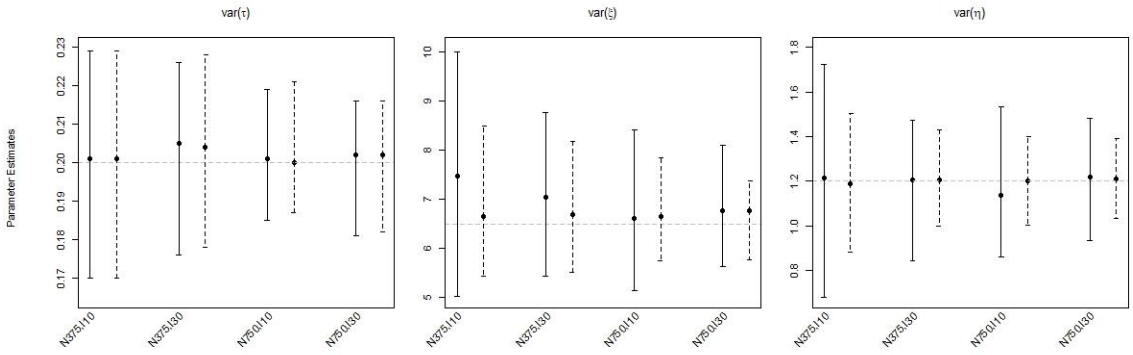


A.19: Medians and 90% Ranges of Differences between Estimated and True Omission Time Discrimination Parameters ω, Plotted against the True Parameters.

$\rho_{.\eta} = 0$



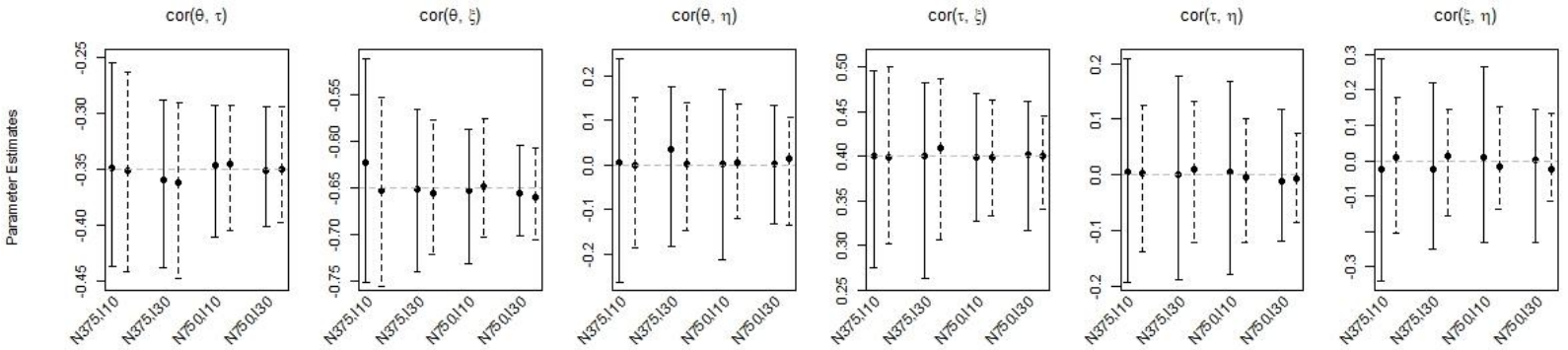
$\rho_{.\eta} \neq 0$



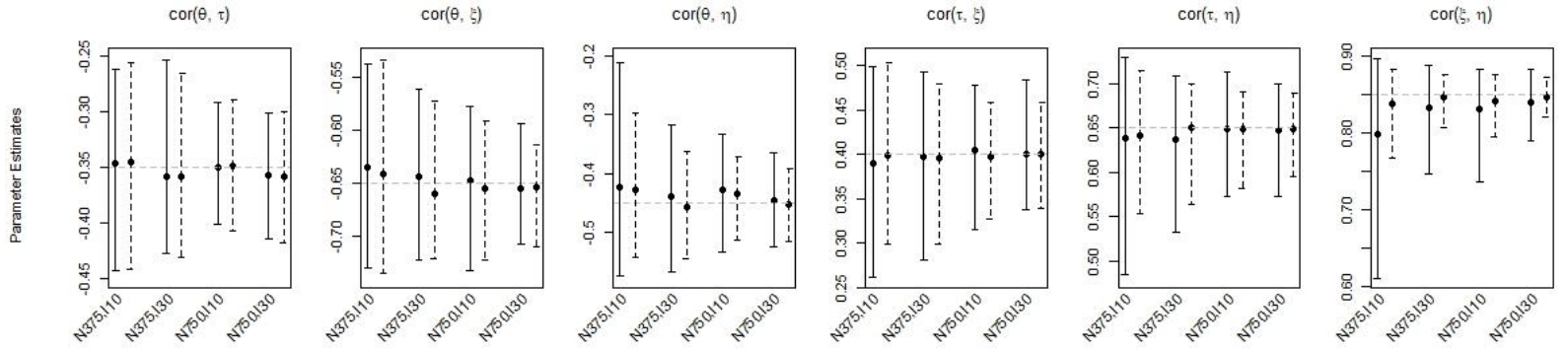
Omission Rate —— 5% ------ 17%

A.20: Medians and 90% Ranges of Person Parameter Variance Estimates

$\rho_{\cdot\eta} = 0$



$\rho_{\cdot\eta} \neq 0$



Omission Rate —— 5% ------ 17%

A.21: Medians and 90% Ranges of Person Parameter Correlation Estimates

REFERENCES

Albanese, M. T., & Knott, M. (1992). TWOMISS: a computer program for fitting a one-or-two-factor logit-probit latent variable model to binary data when observations may be missing. *Technical Report*. Statistics Department, London School of Economics and Political Science, London.

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv preprint arXiv*:1408.4050.

Annis, J., Miller, B. J., & Palmeri, T. J. (2017). Bayesian inference with Stan: A tutorial on adding custom distributions. *Behavior Research Methods, 49*(3), 863-886.

Barnard, J., McCulloch, R., & Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica, 10*, 1281-1311.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics, 44*(6), 706-732.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment. *Teachers College Record, 113*(11), 2309-2344.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician, 49*(4), 327-335.

Cosgrove, J. (2011). *Does student engagement explain performance on PISA? Comparisons of response patterns on the PISA tests across time.* Dublin, Ireland: Educational Research Centre. Retrieved from http://www.erc.ie/documents/engagement_and_performance_ over_time.pdf

Chudgar, A., & Luschei, T. F. (2016). The untapped promise of secondary data sets in international and comparative education policy research. *Education Policy Analysis Archives, 24*, 113.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*(3), 213-234.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*(2), 89-102.

Entink, R. K., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of examinees. *Psychometrika, 74*(1), 21-48.

Feldman, B. J., & Rabe-Hesketh, S. (2012). Modeling achievement trajectories when

attrition is informative. *Journal of Educational and Behavioral Statistics, 37*(6), 703-736.

Follmann, D., & Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics, 51,* 151-168.

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

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4), 540-553.

Foy, P., & Yin, L. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016*. International Association for the Evaluation of Educational Achievement. Retrieved from https://eric.ed.gov/?id=ED580352

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*(3), 515-534.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models*, Vol. *1*. New York, NY, USA: Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*, 3rd ed. Boca Raton, FL: CRC.

Glas, C. A., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling, 57*(4), 523.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute*

Monograph, 3, 125-156.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173-183.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers, No. 133). Paris, France: OECD Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement, 5*, 475-492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153-161.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, *15*(1), 1593-1623.

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732-746.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*(4), 1523-1543.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms

with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*(1), 1-17.

Huang, A., & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis, 8*(2), 439-452.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: a new design for a new era. *Large-scale assessments in education, 5*(1), 11.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015a). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling, 57*(4), 499-522.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015b). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement, 75*(5), 850-874.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress*. (CSE Tech. Rep. No. 357). Los Angeles: University of California.

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika, 45*(2), 527-563.

Lee, Y. H., & Jia, Y. (2012, April). *An investigation of response time in a NAEP computer-based assessment*. Paper presented at the annual conference of the American Educational Research Association, Vancouver, Canada.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation

matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989-2001.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88*(421), 125-134.

Little, R. (2009). Selection and pattern mixture models. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 409–431). Boca Raton, FL: Chapman & Hall.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. New York: Wiley.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*(290), 307-321.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*(2), 247-264.

Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*(4), 615-630.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Boston College, TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/publications/timss/2015-methods/T15-Methods-and-Procedures-TIMSS-2015.pdf

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and Procedures in PIRLS 2016.* Boston College, TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Mislevy, R. J. (2017). Missing responses in item response modeling. In WJ van der Linden (Eds.), *Handbook of Item Response Theory, Volume Two: Statistical Tools* (pp. 171-194). Boca Raton, FL: Taylor and Francis Group.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R. J., & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing*. (ETS Research Rep. no. RR-98-30-ONR). Princeton, NJ: Educational Testing Service.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(2), 371-388.

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies*. Chichester, West Sussex: John Wiley & Sons.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research, 51*(5), 606-626.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 163*(3), 445-459.

Mullis, I. V. S., Martin, M. O., & Diaconu, D. (2004). Item analysis and review. In M. O.

Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 225-252). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

National Center for Education Statistics (2018). *NAEP Technical Documentation*. Washington, D.C: NCES. Retrieved from https://nces.ed.gov/nationsreportcard/tdw

National Center for Education Statistics (2020a). *TIMSS participating countries*. Retrieved January 31, 2020, from https://nces.ed.gov/timss/countries.asp.

National Center for Education Statistics (2020b). *PIRLS Countries*. Retrieved January 31, 2020, from https://nces.ed.gov/surveys/pirls/countries.asp

National Center for Education Statistics (2020c). *Participation in PISA by Year*. Retrieved January 31, 2020, from https://nces.ed.gov/surveys/pisa/countries.asp

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo, 2*(11), 2.

O'Malley, A. J., & Zaslavsky, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association, 103*(484), 1405-1418.

O'muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 162*(2), 177-194.

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris: OECD. Retrieved from https://www.oecd.org/pisa/data/42025182.pdf

Organisation for Economic Co-operation and Development. (2013). *Technical report of the survey of adult skills (PIAAC)*. Paris: OECD. Retrieved from http://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

Organisation for Economic Co-operation and Development (2017). *PISA 2015 technical report*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423-452.

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika, 84*, 892-920.

R Development Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/

Ranger, J., & Kuhn, J. T. (2012). A flexible latent trait model for response times in tests. *Psychometrika, 77*(1), 31-47.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?. *International Journal of Testing*, *17*(1), 74-104.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing, 5*(2), 121-125.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement*

(Unpublished doctoral dissertation). Friedrich-Schiller-University of Jena, Germany.

Rose, N., Von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)*. (ETS Research Rep. no. RR-10-11). Princeton, NJ: Educational Testing Service.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*(3), 795-819.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. New York, NY: Cambridge university press.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142-151.

Sachse, K. A., Mahler, N., & Pohl, S. (2019). When nonresponse mechanisms change: Effects on trends and group comparisons in international large-scale assessments. *Educational and Psychological Measurement*, *79*(4), 699-726.

Schnipke, D. L. (1995). *Accessing speededness in computer-based tests using item response times*. (Unpublished doctoral dissertation). Johns Hopkins University, Baltimore, MD.

Schnipke, D. L. (1999). *The influence of speededness on item-parameter estimation*. Princeton, NJ: Law School Admission Council.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights

gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stan Development Team. (2019). Stan modeling language: User's guide and reference manual. Version 2.19.0.

Sterba, S. K., & Gottfredson, N. C. (2015). Diagnosing global case influence on MAR versus MNAR model comparisons. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 294-307.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 85-152). Mahwah, NJ: Lawrence Erlbaum Associates.

Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology, 9*, 964.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., & Tuerlinckx, F. (2011). Visualizing distributions of covariance matrices. *Columbia Univ., New York, USA, Tech. Rep*, 18-18.

Tong, Y., & Kolen, M. J. (2010). IRT proficiency estimators and their impact. In *annual conference of the National Council of Measurement in Education, Denver, CO*.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425-453.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for

inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*(1), 83-112.

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

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*(1), 5-20.

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

van der Linden, W. J. (2009b). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics, 34*(3), 378-394.

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

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 195-210.

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6), 671-705.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*(1), 144-168.

Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling, 58*(4), 671.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95-114.

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at annual conference of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52-61.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343-354.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee

motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (pp. 408-440). Paris, France: OECD. Retrieved from https://www.oecd.org/skills/piaac/_Technical%20Report _17OCT13.pdf

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education, 26*, 196-212.