ITEM POSITION AND MOTIVATION EFFECTS IN LARGE-SCALE ASSESSMENTS

by

Nayeon Yoo

Dissertation Committee:

Professor Young-Sun Lee, Sponsor
Professor Bryan Keller

Approved by the Committee on the Degree of Doctor of Education

Date 12 February 2020

Submitted in partial fulfillment of the
Requirements for the Degree of Doctor of Education in
Teachers College, Columbia University

2020

ABSTRACT

ITEM POSITION AND MOTIVATION EFFECTS IN LARGE-SCALE ASSESSMENTS

Nayeon Yoo

The purpose of this study is to propose a model that includes student dynamics with the item position effect and student motivation as a test-related psychological factor. In addition, missing data mechanisms were incorporated into models to mimic the actual scene when taking tests. As a prerequisite of the study, the existence of item position effects was identified. Following the first study, SEM models that included student motivation with the item position effect were evaluated, and different missing dataset types were fitted to the SEM models. All analyses used TIMSS 2015 grade 8 mathematics data from the U.S. as exemplary large-scale assessment data. Thus, a significant item position effect was identified on mathematics achievements, accounting for the relevant covariates of student background, socioeconomic status, and psychological variables; the full model with both student motivation and the item position effect was revealed as the best with complete data; the MNAR missingness type was

found to have meaningful information that must be considered in test administration. This research should complement well-developed item position effect studies by focusing on modeling the effect with personal factors that show individual differences, thereby avoiding biased estimates in large-scale assessments.

ACKNOWLEDGMENTS

First, I would like to thank my mom and dad for all their love and cheers. From the day they welcomed me into the world, they have always been beside me. Sometimes taking a step back, smiling and waving whenever I looked their way and always being there, which I believe is the foundation of my strength and courage to step further. I feel lucky to be their cherished little one and best friend. I would also like to thank my parents-in-law and the rest of my family for their love, trust, and support.

I want to express my appreciation to my advisor, Professor Young-Sun Lee, for her guidance and support in both this dissertation and my life. As a supportive advisor, wise teacher, and sometimes friendly senior, she has motivated and encouraged me throughout my studies. I fully enjoyed this stressful yet favorable journey thanks to her.

I would also like to thank my committee members, Professor Bryan Keller, Professor Ye Wang, and Professor Judith Scott-Clayton for their insightful support.

I thank my colleagues, Yu Bai, Ummugul Bezirhan, Ngalula Fleurant, Huei-Yi Lai, Rui Lu, Tianyang Zhang, and Xiaoliang Zhou for their beyond-colleague friendship. I feel very lucky to have been with them not only while in school but also as a part of my life in which we spent the best and worst times together tightly supporting each other.

Most importantly, I dedicate this thesis to my husband Jonghee Kang. He has been always with me from the first day of my studies as my loving one, best friend, inspiring colleague, and supportive family. He turned my fearful journey into a cheerful one at every moment even though there were hard times. At the end of this journey, I am even more certain that we can happily work through whatever comes in our life together. I thank Jonghee for teaming up with me to make an unbeatable pair!

N. Y.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I – INTRODUCTION

With the widespread use of large-scale assessments in reading, mathematics, and science both internationally and nationally, educators are becoming concerned about possible threats to item parameter invariance because the major considerations in educational assessments—referring to the standards—were limited in terms of construct-related factors such as content domains, knowledge, skills, and abilities (AERA, APA, & NCME, 1999). Known sources of item parameter drift that often occur in large-scale assessments include context effects, item position effects, instructional effects, and variable sample sizes (Kingston & Dorans, 1984; Meyers, Miller, & Way, 2008). When considering the less-frequently addressed question of why these effects occur (Debeer & Janssen, 2013), expanding limited construct-related factors to construct-irrelevant factors such as psychological factors in large-scale assessments would better explain item parameter drift occurrences.

Large-scale assessments are designed to measure student progress and achievement for a given content domain at the state, national, or international level by assessing a large number of items. Each assessment has its own content domain and objectives; for example, Trends in International Mathematics and Science Study (TIMSS) includes the mathematics areas of numbers and operations, algebra, geometry, and data and probability, and the science areas of biology, chemistry, and earth science (Martin, Mullis, & Foy, 2015). Progress in International Reading Literacy Study (PIRLS) assesses fourth-grade reading skills including online information (Mullis, Martin, & Sainsbury, 2016), and Programme for International Student Assessment (PISA) is a worldwide test of 15-year-old students' reading, mathematics, and science capabilities (OECD, 2006).

While these are the most representative international large-scale assessments, the National Assessment of Educational Progress (NAEP), which aims to assess the mathematics, science, reading, technology, and geography capabilities of fourth-, eighth-, and twelfth-grade students, is the largest national-level large-scale assessment in the United States. Other large-scale assessments include school admissions tests such as the SAT, ACT, and GRE and independent statewide tests.

While the objectives of these assessments are to assess students' content and skills knowledge and performance, the individual assessments contain far more items than one person could complete in a single sitting; rather than giving all students in a test population the same test form, individual students are administered different combinations of test items (Frey, Hartig, & Rupp, 2009), which also helps prevent cheating and enhances test security (Debeer & Janssen, 2013). Students will typically receive tests with subsets of all questions from a given test's content domain so that students can be adequately assessed with minimal testing burden (Weirich, Hecht, Penk, Roppelt, & Böhme, 2017).

Administering subsets of test items is defined as matrix sampling design; this is used often in large-scale assessments to cover a wide range of content domains. Specifically, students are administered subsets of items in booklets. In a matrix sampling design, available items are grouped into blocks—each of which contains a certain number of items—and each booklet consists of several blocks; the booklets are randomly assigned to each student.

In an experimental design context, a matrix sampling design could potentially have confounding variables that cannot be controlled due to the inconsistent use of booklets. To maximize control for unwanted sources of variation that can affect item

parameters, conditions have been developed for the use of booklets such as each block appearing only once in a booklet, all booklets being of identical length with equal numbers of blocks, and booklets being carefully matched between item distributions within each form and for the distribution across the overall item pool (Frey et al., 2009; Martin et al., 2015). Although analytic procedures to date have relied on the assumption that common items can be treated equally because there are no consequential differences in the effects for these large-scale assessment testing strategies, there may still be contextual effects that can create threats to item parameter invariance.

The item position effect is a predominant source of biased item parameter estimates from the matrix sampling design. It is defined as the impact of an item's position within a test from estimates of the item parameters and the latent trait (Debeer & Janssen, 2013); for example, students may experience different item difficulties depending on an item's position. As stated earlier, while item position effects are often assumed either equal or negligible for all examinees and items (Hahne, 2008), researchers have addressed position effects in various operational testing applications. Their findings have been inconclusive; some authors found no significant item position effect (i.e. no difference in item difficulty across positions; Klein & Bolus, 1983; Zwick, 1991) while some found the significant effect that items became more difficult toward the end of a test (Le, 2007; Meyers et al., 2008). Le (2007) found the item position effect that items were more difficult when located toward the end of a test upon examining PISA 2006 science data, which suggests the appropriateness of considering different item locations when administering different test forms. Meyers et al. (2008) examined item position effects and determined that changing item position significantly impacted item difficulty; specifically, difficult items became more difficult when shifted to the end of a test and

easier items were easier when moved toward the beginning of a test. Therefore, item difficulty was influenced by both the item's original difficulty level and its position.

Despite conflicting results, prior studies have often discussed common psychological effects of fatigue and practice in relation to item position effects (Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011; Kingston & Dorans, 1984); specifically, the common suggestion is that fatigue effects can make items more difficult in later positions on tests but practice effects can make items in later positions on tests easier. In terms of student performance, performance can decrease toward the end of a test due to fatigue but can increase toward the end of a test due to practice.

Researchers have interpreted their findings on item position effects in the contexts of both fatigue and practice effects; they suggested context significantly affects results and canceling between fatigue and practice has a non-significant effect on findings. One study conducted in reverse measured test fatigue effects on students using item difficulty and item position; the author found a close mutual relationship between fatigue and item position effects (Davis, 2005). To better understand the dynamics between context and item position effects, Debeer and Janssen (2013) suggested examining "the why question" by including person predictor variables or psychological factor variables such as test-taking efforts or motivation in models. Making this adjustment would more fundamentally explain the increased item difficulty from fatigue effects by addressing decreased test-taking efforts or motivation. A recent study examined the moderation between item position effect and test-taking efforts and found that both significantly affected student performance; examinees' test-taking efforts decreased and item difficulty increased as the test progressed (Weirich et al., 2017). They concluded that the item position effect depended on the individual and that changes in test-taking efforts could

moderate the position effect. Therefore, there is a requirement to consider construct-irrelevant yet test-related factor effects to avoid biased estimates in large-scale assessments.

Missing data is an intrinsic issue that should be considered with these effects because disregarding information-rich missing data could create biased estimates (Bradlow & Thomas, 1998). Large-scale assessments often encounter missing item responses or missing background data (Weirich, Haag, et al., 2014), excluding planned missing values caused by matrix sampling design (Frey et al., 2009). Rubin (1976) introduced three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Among these, MNAR may directly reflect the effects of item position and psychological factors. Specifically, examinee non-response is a major source of missing data and is often caused by time limitations or a lack of motivation; such missing data that are not independent of item position and psychological factors could be regarded as MNAR (Happ & Förster, 2018). Thus, incorporating missing data mechanisms into models could help better understand what student dynamics are associated with item position and person predictor effects, thereby improving test validity and fairness in large-scale assessments. For example, one of the most researched large-scale assessments—TIMSS—would fit such a model in the current study.

TIMSS is suitable for real-world data analysis example because the test employs a matrix sampling approach (i.e., booklet design), and measures students' psychological factors. For eighth-grade mathematics assessment, TIMSS includes the domains of numbers, algebra, geometry, data, and chance. There are topic areas within each domain such as whole numbers; fractions, decimals, and integers; and ratios, proportions, and

percentages in the number domain. To ensure that this wide range of domains and topic areas is assessed reliably, TIMSS aims to maximize available items while considering the appropriate number of items that students can answer carefully without undue burden; the test employs a matrix sampling design that groups the entire pool of items into subsets of booklets, and each student completes one. Chapter III explains the design in further detail. In addition, TIMSS measures information about students' readiness to learn, motivation, self-concept, and characteristics while emphasizing the importance of student attitudes toward a subject in educational research. Specifically, TIMSS administrators explained that they included these student motivation- and confidence-related psychological factors because they are essential when assessing student performance. In addition, there has been recent discussion that educational goals should be reset to improve positive attitudes toward subjects rather than simply improve student achievement (Martin et al., 2015). This approach indicates that the approach to improving student performance is to gradually move from solely construct-related factors to different underlying construct-irrelevant factors.

This work is necessary because few researchers have explored item position effects in a large-scale assessment that incorporates construct-irrelevant yet test-related psychological factors. In extensive empirical research, student anxiety, motivation, confidence, and the threat of evaluation are often the most prominent and relatively stable test-related characteristics that influence students' psychological and behavioral responses in threatening situations such as during tests (Hancock, 2001; Hembree, 1988). Specifically, student motivation is an underlying foundation for other psychological factors given that it is flexible and frequently used as both a predictor and outcome in education research. This conclusion raises the need to carefully consider motivation in

student-assessment dynamics and the aim of this work is thus to expand the existing item position effect research by considering student motivation, which is an important test-related psychological factor. This effort should complement well-developed item position effect studies by focusing on modeling the effect with personal factors that show individual differences, thereby avoiding biased estimates in large-scale assessments.

Objectives

This study adds to the existing literature that only covers construct-related factors when examining the item position effect. This study's purpose is to explore how item position effect models are improved by student motivation effect to suggest the versatile application of standards in education assessments. Specifically, this dissertation consists of two studies that seek to answer three Research Questions. Research Question 1 is the fundamental question of whether there are significant item position effects when accounting for students' psychological factor variables; groups were compared, and propensity score matching analysis was conducted in addition to answer this. Research Question 2 seeks to understand the effect of motivation on the relationship between item position and math ability and Research Question 3 is an additional examination that asks whether the effects of student motivation and item position on math ability differ with different types of data missingness. Structural Equation Modeling (SEM) was used to answer Research Questions 2 and 3. In study 1, Research Question 1 was investigated using the real-world data of TIMSS 2015 grade 8 mathematics data from the United States. In Study 2, Research Questions 2 and 3 were examined using the same TIMSS data where missing data techniques had been applied—cases were deleted according to missing data types and imputed—to the original dataset.

This dissertation consists of five chapters. Chapter II contains a general overview of item position effects and current issues and discusses modeling approaches and relevant frameworks. Chapter III illustrates the methods employed in the study and Chapter IV presents the results of data analyses. Chapter V summarizes the results, discussions, and limitations and makes suggestions for future research.

Chapter II – LITERATURE REVIEW

Chapter II consists of two sections; the first introduces the general concept of item position effects and relevant issues and the second presents the existing modeling approaches of item position effects and relevant frameworks.

Item Position Effects

Overview of Item Position Effects

Context effects can occur when item positions change under circumstances such as developing tests, administering alternate forms, or computerized adaptive tests. Although placing the same items in a consistent position across various operations is preferable (Li, Cohen, & Shen, 2012), items' positions must be changed in testing practice for operational and security reasons. The item position effect is a context effect that can affect examinees' response behavior or item parameter estimates through items' different positions in a test (Bulut, 2015). In addition to parameter estimates, test-equating results were affected (Zwick, 1991).

International and national large-scale assessments were actively used to examine item position effects as they mostly used booklet designs for test administration; these have included the National Assessment of Educational Progress (NAEP; Zwick, 1991), the Programme for International Student Achievement (PISA; Le, 2007), the Graduate Record Examination (GRE; Albano, 2013; Kingston & Dorans, 1982), the mathematical competency test of the Austrian Educational Standards (Hohensinn et al., 2011), and the German nationwide low-stakes large-scale assessment (Weirich et al., 2017). Among such studies, some of their results support the hypothesis that items administered at the

end of a test show different difficulty levels compared to the same items administered at the beginning of a test. Significant item position effects were detected for the PISA, GRE, NAEP, and German nationwide low-stakes large-scale assessment; in the PISA 2006 science assessment that used a booklet design, Le (2007) found that items became more difficult when located at the end of the test for all 53 participating countries and raised issues of fatigue or speed effects in test design. In addition, the results revealed that the item difficulty estimates from each position group were highly correlated with each other. The administration of the GRE, for which the items appeared in various different positions within each test, also showed a significant item position effect. Student performance was higher for identical items shown at the beginning compared to those at the end for both quantitative and verbal content (Albano, 2013), which was consistent with the results from another study with the GRE (Kingston & Dorans, 1984). For the NAEP reading, Zwick (1991) found that inconsistent scaling in 1984–1986 was due to changes in item positions. The results from a study of the German nationwide low-stakes large-scale assessment presented a linear item position effect in which item difficulty increased while taking the test (Weirich et al., 2017). In addition, they found that the effect varied across individuals with diverse levels of test-taking efforts. Contrasting results without any occurrence of item position effects were also reported; no item position effects were found for the mathematical competency test of the Austrian Educational Standards (Hohensinn et al., 2011). The possible reasons discussed for non-significant effects were different testing conditions such as testing time (Marso, 1970), age (Robitzsch, 2009), and ability (Debeer & Janssen, 2013).

Relevant Concepts

Both the significant and non-significant results share certain relevant factor effects: 1) other closely related context effects such as fatigue and practice effects; 2) person predictors such as age, ability/achievement level, and test-taking effort; 3) speed effect due to the time limitation.

Fatigue and practice effects. Fatigue and practice effects are those most frequently discussed when interpreting item position effects; as shown in previous studies, items that tend to be more difficult at the end of a test or become more difficult during a test might have been affected by examinee exhaustion. This fatigue effect could also be related to decreasing test-taking efforts or motivation (Weirich et al., 2017). The practice effect had the opposite role. Items decreasing in difficulty toward the end of a test or becoming less difficult during a test might be due to the practice effect as examinees become familiar with the test conditions or content.

Person predictor effect. Although these effects are mentioned repeatedly in the interpretation of item position effects and appear reasonable, examining person predictors in a model is becoming increasingly necessary. Specifically, item position effects are not clearly observable as each examinee only experiences each item once at a specific position (Weirich et al., 2017). Debeer and Jenssen (2013) claimed that capturing individual differences and including them in the model would help explain item position effects. They suggested exploring "the why question" of the effects by including person predictors. Examples of person factors are age, ability, test-taking effort, and test motivation. For age, Robitzsch (2009) detected different item position effect patterns for third- and fourth-grade students. In a study that included student ability in the model, the higher-ability group showed a smaller item position effect (Debeer & Janssen, 2013) and

a study that explored moderation between the effect and test-taking efforts found significant interdependence (Weirich et al., 2017). Wise and DeMars (2005) suggested considering test motivation in low-stakes assessments because students' test motivation might differ in such assessments. The discrepancy between different person factors would result in biased estimates no matter how well designed the assessments unless these person factors are kept constant.

Speed effect. Finally, the speed effect is closely related to the item position effect (Le, 2007). The speed effect is an inevitable concern in testing as most examinees have to stop before they can attain their best result (Schweizer & Ren, 2013). This is an essential factor to consider with the item position effect in that the effect implies test-takers' lack of time, which can also cause missing data; the speed effect was often discussed in prior studies paired with the fatigue effect when interpreting the item position effect (Le, 2007; Marso, 1970). Although the given time and number of items are carefully balanced through repeated pretests to ensure that tests can be operated as a power test such that almost all students have sufficient time, time limits remain due to operational circumstances in large-scale assessments. Therefore, item position effects could occur if some examinees do not have the opportunity to complete their test and thus cause unanswered questions and missing data (Hohensinn et al., 2011). Specifically, quantitative tests (i.e., mathematics) showed different results for non-significant item position effects in power tests and significant item position effects in speed tests (Flaugher, Melton, & Myers, 1968).

Happ and Förster (2018) combined the fatigue, person predictor, and speed effects and suggested that fatigue, test motivation, effort effects, and missing values should be associated. Numerous missing values from non-responses rather than from the sampling

design might be caused by low test motivation and effort or limited time given to solve all items. In a study that examined the number of missing values with test items becoming less difficult toward the end of the test, the results revealed that the number of missing values was not consistent with difficulty changes; rather, they increased (Musekamp & Pearce, 2016). They concluded that this was because test motivation decreased during the test.

Modeling Item Position Effects

Modeling Item Position Effects Literature

Prior studies selected different approaches to detect item position effects such as treating item position effects as differences between test form groups (Bulut, Quo, & Gierl, 2017; Marso, 1970; Plake, Ansorge, Parker, & Lowry, 1982), modeling item positions as an item attribute in the IRT model (Albano, 2013; Debeer & Janssen, 2013; Fischer, 1973; Weirich, Hecht, & Böhme, 2014), and modeling a larger model in which IRT models are embedded with additional factors (Bulut et al., 2017).

Comparing groups. A general approach for detecting item position effects is grouping examinees as randomly equivalent groups by received test forms and then comparing them. Marso (1970) compared group means after randomly assigning different formats of the Quick Word Test (QWT; Borgatta & Corsini, 1960). The test formats had item arrangement designs such as random, descending order of difficulty, and ascending order of difficulty. An analysis of variance (ANOVA) was used to investigate the group differences in achievement scores and individual testing times; furthermore, test anxiety score was utilized as a classification factor. Thus, different item arrangements based on difficulty did not appear to meaningfully affect achievement scores or testing times, nor

was there a significant difference in performance between students with high- and low-test anxiety.

Plake et al. (1982) investigated the effects of item arrangement with test anxiety; they identified the effects of item arrangement, knowledge of arrangement, and test anxiety on perceived performance and difficulty in a multiple-choice mathematics test. In addition, they examined the interaction effects between gender, item arrangement, knowledge of arrangement, and test anxiety. Analysis of covariance (ANCOVA) and multivariate analysis of covariance (MANCOVA) were used to evaluate the fixed effects and differences between item arrangement types—easy/hard, uniform or spiral cyclical, and random order—for male and female students. The results showed that male students outperformed female students when items were ordered from easy to hard.

Bulut (2015) applied differential item functioning (DIF) analysis to this topic; as multiple test booklets are widely used in large-scale assessments, the effect of test booklets and gender-based DIF was examined. Specifically, the interaction effect between the test booklet and gender on students' performance—i.e., whether differently ordered yet identical items impacted the gender DIF pattern in reading assessments—was checked. The Mantel–Haenszel (MH; Mantel & Haenszel, 1959) approach was used to detect uniform DIF items and the Breslow–Day (BD; Breslow & Day, 1982) method was used to detect non-uniform DIF items. The difficulty levels for DIF items detected by the MH and BD methods did not show a consistent pattern. Thus, the DIF patterns from both MH and BD were inconsistent; the number and difficulty levels of non-uniform DIF items varied across test booklets while the number of uniform DIF items was similar between booklets for both genders.

Including in the IRT model. An alternative approach to modeling item position effects is treating the effect as an item attribute and directly including it in the IRT model. By applying explanatory item response models (Wilson & De Boeck, 2004), item responses can be modeled as a function of potential predictors such as item attributes and person characteristics; the linear logistic test model (LLTM; Fischer, 1973) is one such model that allows item position to count as an item characteristic predictor and can thus account for some of the item difficulty variance (Weirich, Hecht, et al., 2014). Specifically, the LLTM decomposes the item parameters of the Rasch model as a linear function of predictors and enables the examination of testing condition applications (Kubinger, 2009). Starting with the Rasch model where examinee j with ability parameter θ_j solves item i with difficulty parameter β_i,

$$P(u_{ij} = 1 | \theta_j, \beta_i) = \frac{e^{(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}}.$$

Difficulty parameter β_i can be decomposed into two components by adding parameter δ_{ik} for the item position effect:

$$\eta_{ij} = \theta_j - \beta_i$$

$$\eta_{ijk} = \theta_j - (\beta_i + \delta_{ik}).$$

Now, η_{ijk} becomes the logit of the response of examinee j to item i at position k:

$$logit[u_{ijk} = 1] = \eta_{ijk} = \theta_j - (\beta_i + \delta_{ik}),$$

where β_i is the difficulty parameter of item i at the reference position. This can be extended to a two-parameter logistic (2PL) model:

$$logit[u_{ijk} = 1] = (\alpha_i + \delta_{ik}^{\alpha})[\theta_j - (\beta_i + \delta_{ik}^{\beta})],$$

where α_i is the discrimination parameter of item i at the reference position, δ_{ik}^{α} is the discrimination parameter of item i at position k, and δ_{ik}^{β} is the difficulty parameter of item i at position k. Extensions to different models are possible within this LLTM framework. Debeer et al. (2013) modeled item position effects across items and individual differences in effects; they conducted simulation and empirical studies using the PISA 2006 assessment data. The results from both studies found considerable individual differences associated with item position effects. Not all examinees showed consistent effects although they were significant; for example, items tended to be more difficult in the latter part of the test while some others became easier. They concluded that item position effects should be interpreted as a person-dependent trait rather than generalized fatigue or practice effects.

Considering the limitation of the LLTM, namely its lower accuracy due to the model lacking an error term in the prediction of item position δ_{ik}, LLTM with an additional error term ε was suggested (De Boeck, 2008):

$$\eta_{ijk} = \theta_j - (\beta_i + \delta_{ik})$$

$$\eta_{ijk} = \theta_j - (\beta_i + \delta_{ik} + \varepsilon_{ik}).$$

The implication of an error term in LLTM $+ \varepsilon$ in the generalized linear mixed-models (GLMM) framework indicates that δ_{ik} now consists of a fixed effect and a random effect, enabling the more precise prediction of the item position effect as the item difficulty cannot be fully explained by the item position itself (Weirich et al., 2014). Specifically, a fixed effect is a predicted effect of an item position and a random effect represents uncertainty in the prediction. Weirich et al. (2014) conducted a simulation study with the design's conditions (completely balanced, partially balanced, or unbalanced), the

magnitude of position effects (none, linear, weak nonlinear, or medium nonlinear), and the sample size (2,000 or 4,000) to determine whether the application of LLTM + ε in modeling position effects was appropriate. Thus, the LLTM + ε approach seemed adequate for modeling the effect, particularly when the design was balanced. In an unbalanced design, item position effects were less likely to be identified.

Another example of estimating the item position effects as fixed and random effects in IRT models in the generalized linear model (GLM) framework is the hierarchical generalized linear model (HGLM) approach (Albano, 2013). Similar to the aforementioned models, η_{ij} is the logit of the correct response in the HGLM:

$$\eta_{ij} = \beta_{0j} + \sum_{q=1}^{N-1} \beta q_j X_{qij} + \beta_{Nj} p_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{qj} = \gamma_{q0}$$

$$\beta_{Nj} = \gamma_{N0},$$

where the intercept γ_{00} is the average item difficulty effect at the reference position and the remaining effects are the differences from the grand mean. The random effect of u_{0j} is the ability parameter of examinee j, X_{qij} is an item indicator where $X_{qij} = 1$ when $q = i$ ($X_{qij} = 0$ otherwise), γ_{q0} is the effect of item $q = i$, p_{ij} is the position of item $q = i$ for examinee j, and γ_{N0} is the fixed effect for the position, i.e., the item position effect. This approach is also known as the multilevel modeling approach and it permits the estimation of several effects across complicated data structures. Albano (2013) found significantly different impacts for item positions on item difficulty in the GRE and detected 12 potential DIF items: six quantitative items and six verbal items.

Embedding IRT models in a larger model. A final extension of these approaches is to include additional item position-related factors and examine item position and other factor effects simultaneously; a binary factor analysis (FA) model that is often used to estimate the IRT model can be expanded to structural equation modeling (SEM). When a latent independent variable and observed dependent variables are included in a regression model, it is considered an FA model (Ferrando, Anguiano-Carrasco, & Demestre, 2013). As a one-factor FA model can be applied to the IRT model:

$$y_i = \lambda_i \eta + \varepsilon_i,$$

where y_i is the latent response function for underlying items, η is the latent trait, λ_i is the factor loading for item i, and ε_i is the residual term for item i; this can be defined as a measurement model with a latent trait (e.g., ability) in the SEM framework. In the structural model, the causal relationship between the latent trait and other relevant variables such as item position can be examined:

$$y_i = \lambda_i \eta + \beta_i p_i + \varepsilon_i,$$

where p_i is the position of item i for all examinees and β_i is the effect size of the item position (Bulut et al., 2017). In other words, IRT models can be embedded in a full SEM model with additional variables such that the causal relationship between these latent and observed variables and parameters can be examined while taking account of all related factors. Bulut et al. (2017) conducted a study to introduce and represent the appropriateness of the SEM framework for modeling item position effects in large-scale assessments. An empirical study was conducted using the statewide reading assessment followed by a simulation study. The results showed that the SEM approach is suitable for modeling various types of item position effects from the empirical study; furthermore, the

SEM model contributes to successful parameter recoveries even with small sample sizes. Therefore, they concluded that this approach could accurately detect item position effects. In addition, the advantages of the SEM framework for modeling item position effects have been discussed: 1) the SEM approach can handle both dichotomous and polytomous items to be examined for the effects; 2) evaluating both item position effects and IRT models is possible as the IRT model can be embedded in the full model as in the FA model; 3) the approach enables multiple latent or observed variables underlying the circumstances to be incorporated into the model, which will contribute to a more precise interpretation.

In summary, various methods have been developed for modeling item position effects such as comparing groups (Marso, 1970; Plake et al., 1982), the LLTM approach (Debeer & Janssen, 2013; Fischer, 1973), the GLMM approach (Weirich et al., 2014), the HGLM approach (Albano, 2013), and the SEM approach (Bulut et al., 2017). This dissertation adopts both the comparing groups approach and the SEM approach; the groups were compared with a recent technique—propensity score matching—and the IRT models were embedded in the SEM model with a person predictor.

Relevant Theoretical Frameworks

Propensity score matching. Although comparing the equivalence of test form groups with different item orders is a general approach to modeling item position effects, the predominant limitations of such a comparison are: 1) it is limited to a random assignment of examinees in treatment and control groups; 2) item position effects may remain undetected when the effects cancel out due to confounding effects (Debeer et al., 2013). Matched sampling aims to overcome the discrepancy between treatment and control groups; this method selects a control group that is similar to the treated group in

terms of covariate distributions (Rosenbaum & Rubin, 1985). Background variables are often regarded as covariates.

Specifically, average causal effects (ACEs) can be estimated via propensity score matching. ACE is defined as

$$ACE = E(Y_{i1}) - E(Y_{i0}),$$

where Y_{i1} denotes the response when treated and Y_{i0} denotes the response when untreated, which can be regarded as another type of group comparison approach. In addition, ACE for the treated or untreated are popular alternatives depending on the group of interest (Schafer & Kang, 2008):

$$ACE_{treated} = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1),$$

$$ACE_{untreated} = E(Y_{i1}|T_i = 0) - E(Y_{i0}|T_i = 0).$$

Estimating ACE requires assumptions of stable unit treatment value assumption and strong ignorability assumption. A stable unit treatment value assumption assumes that an examinee's treatment assignment does not affect other examinees' outcomes and a strong ignorability assumption assumes strict independence between treatment assignment and potential outcomes, given the observed covariates (Rosenbaum & Rubin, 1983). Rosenbaum and Rubin (1983) defined the propensity score as the conditional probability of being treated given covariates:

$$P(T_i = 1|X_i),$$

where T_i denotes the treatment received by examinee i and X_i denotes the vector of covariates for examinee i. Using propensity scores, examinees from treatment and control groups are chosen based on their identical propensity scores, thus both groups could have balanced covariate distributions. ACE estimated after propensity score matching reduces the bias in covariates.

IRT framework. The IRT approach is another important framework to discuss. IRT (Lord & Novick, 1968) enables the measurement of an unobserved hypothetical variable θ (ability or a latent trait) by assuming a relationship between individuals' abilities and item responses (Baker & Kim, 2004). Ability is interchangeably used as attitudes, personalities, or other latent traits. Several different IRT models are applicable depending on item formats, the number of traits, and the number of item parameters. Regarding the item scoring, there are dichotomous and polytomous IRT models for binary responses and responses with multiple categories, respectively. Regarding the number of traits, there are unidimensional and multidimensional IRT models for a single trait and for multiple traits; based on the number of item parameters, there are one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) IRT models. This section presents the 1PL model—also called the Rasch model—and the 2PL model as a dichotomous IRT model and the graded response model (GRM; Samejima, 1969) as a polytomous IRT model. All presented models assume unidimensionality.

The Rasch model (Rasch, 1960) is the simplest IRT model with a single parameter; the model presents the probability of a correct response for item i by examinee j from the following:

$$P\left(X_{ij} = 1 \middle| \theta_j, \beta_i\right) = \frac{\exp\left(\theta_j - \beta_i\right)}{1 + \exp\left(\theta_j - \beta_i\right)},$$

where X_{ij} is the response of examinee j to item i, θ_j is examinee j's ability level, and β_i is item i's difficulty. The difference between ability level and item difficulty determines the probability of correctly answering the item.

The 2PL model (Birnbaum, 1968) extends the Rasch model by adding a discriminant parameter; the model predicts the probability of examinee j getting item i correct:

$$P\left(X_{ij}=1\middle|\theta_j,\beta_i,\alpha_i\right)=\frac{\exp\left[\alpha_i\left(\theta_j-\beta_i\right)\right]}{1+\exp\left[\alpha_i\left(\theta_j-\beta_i\right)\right]},$$

where α_i is a discrimination for item i. The 2PL model is suitable for situations in which items are unequally affected by ability level because it includes another item parameter that differentiates items.

The GRM (Samejima, 1969) is one of polytomous IRT models; it is used when item responses are ordered into multiple categories such as Likert scale responses. The GRM model is a generalized 2PL model that has multiple thresholds rather than a single threshold. Therefore, a conditional probability for having an examinee's response fall into a certain category k is involved and written as:

$$P_k^*(\theta)=\frac{\exp\left[\alpha_i(\theta-\beta_{ik})\right]}{1+\exp\left[\alpha_i(\theta-\beta_{ik})\right]},$$

where k is the categories for item i, α_i is the discrimination for item i, and β_{ik} is a threshold parameter between categories. When there are k categories, the category response probabilities of $P_0^*(\theta)$, $P_1^*(\theta)$, … $P_{k-1}^*(\theta)$ are estimated, $P_0^*(\theta)=1.0$ and $P_{k-1}^*(\theta)=0.0$ for all θ:

$$P_k(\theta)=\ P_k^*(\theta)-P_{k+1}^*(\theta).$$

For example, when there are four categories:

$$P_0(\theta)=\ 1.0-P_1^*(\theta)$$

$$P_1(\theta)=\ P_1^*(\theta)-P_2^*(\theta)$$

$$P_2(\theta)=\ P_2^*(\theta)-P_3^*(\theta)$$

$$P_3(\theta) = P_3^*(\theta) - 0,$$

where $P_0(\theta)$–$P_3(\theta)$ represent the conditional probability of an examinee responding in each category given their ability level.

In this study, 2PL and GRM are chosen for embedding in a larger model because: 1) 2PL appeared robust to a violation of an IRT assumption of local independence when examining item position effects (Wells, Subkoviak, & Serlin, 2002); 2) to keep the number of item parameters consistent across models with different item scoring as GRM is a generalized version of the 2PL IRT model.

SEM framework. The SEM framework enables the embedding of such IRT models in a larger model with additional variables (Bulut et al., 2017). In addition, the SEM model consists of two parts: the measurement part and the structural part.

The measurement part allows latent variables to be estimated from a set of observed responses. A measurement model—or factor analysis model—is essentially a regression model with manifest dependent variables and a latent independent variable (Ferrando et al., 2013). Measurement models can be formulated as:

$$x = \Lambda_x \xi + \delta$$

$$y = \Lambda_y \eta + \varepsilon,$$

where x and y are observed responses, Λ_x and Λ_y are the factor loading matrices for latent variables ξ and η, respectively, and δ and ε are the residuals.

The structural model links the latent variables to other latent variables or observed covariates. A structural model can be formulated as

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta,$$

where B and Γ are the linear effect matrices and ζ is a residual, which permits the examination of the relationships between these latent variables. In addition, they are capable of easily extending the model with other variables including other latent variables and manifest variables; therefore, the direct and indirect effects among variables can be explored.

Missing data. This section also briefly addresses the missing data mechanism. Missing data is considered a lack of response for various unknown reasons (Schafer & Graham, 2002). In testing, possible circumstances that can cause missing data could be an examinee's refusal to answer, lack of knowledge, lack of time, simple mistake, or attrition during the test; specifically, lack of time to finish an exam would be closely related with the item position effect as mentioned when introducing the speed effect. Such non-responses were often ignored by simply applying a case deletion method; however, missingness could be relevant data. In detail, in cases of missing data with a specific reason such as lack of time or motivation rather than a simple mistake or refusal to answer, simply ignoring missing data could result in biased parameter estimation, less reliable generalizability for results, and information loss (Dong & Peng, 2013). Thus, Rubin (1976) classified missingness based on the complexity of the missingness pattern as missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR).

Denoting complete data as Y_{com}, observed data as Y_{obs}, and missing data as Y_{mis}, missing data is defined as MAR when the distribution of missingness R does not depend on Y_{mis},

$$P(R|\mathrm{Y}_{com}) = P(R|\mathrm{Y}_{obs}).$$

MCAR is a special case of MAR in which the distribution of missingness R does not depend on either Y_{mis} or Y_{obs}:

$$(R|Y_{com}) = P(R).$$

Meanwhile, data can be defined as MNAR when the distribution of missingness R depends on Y_{mis}; MNAR is no longer an ignorable nonresponse and assumed to reflect meaningful information such as examinee dropout (Schafer & Graham, 2002).

To date, methods for handling missing data include case deletion, pairwise deletion, averaging available items, single imputation, and multiple imputation; case deletion—also known as listwise deletion—simply means that units whose data is incomplete are dropped and is by far the most common method for dealing with missing data. However, this may cause bias in estimates unless the data is MCAR. Pairwise deletion eliminates data selectively and estimates parameters using different sets of sample units; it is regarded as less biased than listwise deletion. The method of averaging the available items is motivated by the fact that items are exchangeable when they are from a reliable measure, a variable's mean value replaces the missing response for the same variable. Although this method is not generally accepted because it increases the sample size while decreasing the reliability, studies have shown that this mean substitution method behaves well in practice (Schafer & Graham, 2002). Imputation is the process of replacing missing data with estimated values when partial but useful information is available; this is considered more efficient than listwise or pairwise deletion because it keeps all samples and makes use of important information. Single imputation substitutes each element of missing data with a single value and multiple imputation replaces each element of missing data with a set of plausible values that can account for the variability. Beyond these methods, various approaches to handle missing

data have been actively developed including full information maximum likelihood (FIML) estimation and weighting methods.

Chapter III – METHODS

This dissertation consists of two studies: Study 1 dealt with prerequisite analysis using a real-world dataset and in Study 2, main data analyses were conducted using the same dataset with missing data techniques applied to the original dataset—cases were deleted according to missing data types and imputed—to answer the research questions.

This chapter includes three sections: the first introduces the research questions and proposed models, the second describes the details of Study 1, and the third describes the details of Study 2.

Research Questions and Proposed Models

Research Questions

This study's purpose is to examine the item position effect while taking the student motivation effect into account. To this end, the following research questions are addressed:

Research Question 1: Do item positions significantly affect mathematics achievement? (Prerequisite analysis)

Research Question 2: Do item position and student motivation affect mathematics ability?

Research Question 3: Do item position and student motivation effects on mathematics ability differ for different types of data missingness?

Research Question 1 was answered with empirical data (TIMSS 2015 grade 8 mathematics data) in Study 1. Research Questions 2 and 3 were investigated with

manipulated TIMSS datasets—missing data techniques applied to the original TIMSS dataset with missing data resulting in complete-cases datasets—in Study 2.

In Study 1, Research Question 1 was motivated by the question of whether there is a significant item position effect. To check the robustness of simple group comparison analysis with TIMSS data, propensity score matching analysis was conducted to derive the causal inference accounting for relevant covariates including socioeconomic status, person predictor, and psychological factor variables. By defining treatment as receiving items at the end and control as receiving the same items at the beginning, the average treatment effect was obtained by estimating the propensity scores and then using them to match individual cases on covariates from the treatment and control groups (Morgan, Frisco, Farkas, & Hibel, 2010). The total score of the selected mathematics items was used as an outcome for comparison.

In Study 2, the effects of student motivation and item position on mathematics ability for both complete and imputed datasets were investigated to answer Research Questions 2 and 3. Three SEM models were proposed to fit the data: math ability (θ_{Math}, referred to as θ_M) and motivation ($\theta_{motivation}$, referred to as θ_m) were estimated under the 2PL model and a GRM within such SEM models. Then, five types of data with different missing data scenarios—a complete dataset, an MCAR dataset with two imputation methods, and an MNAR dataset with two imputation methods—were fitted to each of these models. Model fits and effect sizes were examined.

Proposed Models

Item position and student motivation effects were modeled under a structural equation modeling (SEM) framework. Taking advantage of the SEM framework, item

response theory (IRT) models with latent variables such as math ability (θ_M) and motivation (θ_m) underlying a set of item responses were defined using the following measurement models as factor models:

$$\theta_M = \lambda_{\eta i}\theta_M + \varepsilon_i$$

$$\theta_m = \lambda_{\xi i}\theta_m + \delta_i,$$

where $\lambda_{\eta i}$ is the factor loading for mathematics item i and $\lambda_{\xi i}$ is the factor loading for motivation item i with measurement errors ε_i and δ_i. The causal relationships among these latent variables and item positions as an observed variable were tested in the structural part of the SEM model:

$$\theta_M = \gamma_1\theta_m + \gamma_2 P_i + \zeta_1,$$

where P_i is the position of item i, γ_1 is the effect size of student motivation on math ability, γ_2 is the effect size of item position on math ability, and the remaining terms are the same as above. This study proposed three SEM models: $Model\ 0$ (null model with no effect), $Model\ 1$ (item position effect model), and $Model\ 2$ (item position and moderation effect model) as follows:

$$Model\ 0\text{: } Null\ model$$

$$Model\ 1\text{: } \theta_M = \gamma_2 P_i + \zeta_1$$

$$Model\ 2\text{: } \theta_M = \gamma_1\theta_m + \gamma_2 P_i + \zeta_1.$$

Figures 1, 2, and 3 each present the proposed models.

Study 1

Data

Real-world data analysis was conducted to investigate the causal inference of item positions. TIMSS 2015 grade 8 mathematics data from the United States was used as an exemplary large-scale assessment data that employs a matrix sampling design to detect any existing item position effect. Furthermore, wide ranges of student background variables from this data were used as relevant covariates.

Many large-scale assessments use a matrix-sampling design to maximize available items while considering an appropriate number of items that students can answer carefully without undue burden. In TIMSS, the entire pool of items is grouped into subsets of 14 student booklets, and each student completes one. In detail, approximately 12–18 items are packaged into each block and there are 28 blocks in total (14 in mathematics and 14 in science). Then, these blocks are assembled in various combinations as student booklets where each booklet contains four blocks (two in mathematics and two in science). The mathematics blocks come first in half of the booklets and the science blocks come first in the other half (Martin et al., 2015). Therefore, each item appears in two booklets at two different positions, thus allowing a comparison between the two groups of student responses. Table 1 represents the booklet design and allocations of 28 item blocks in TIMSS 2015. Among these, item block 01 from booklets 1 and 14 were chosen to represent common items at different positions – at the beginning and at the end.

Item block 01 consisted of 17 mathematics items in either multiple-choice format or constructed-response format. Multiple-choice questions were given one point each and constructed-response questions were given one point for partial credit and two points for

full credit. The distribution of items across content domains (number, algebra, geometry, data, and chance) and cognitive domains (knowing, applying, and reasoning) in each item block were allocated to match the distribution across the whole item pool as closely as possible. All students had 90 minutes to work on items (45 minutes for each part) and were expected to spend 22.5 minutes on each item block on average. The testing time was agreed among the National Research Coordinators from participating countries from previous assessments (Martin et al., 2015). The response data were scored using the syntax from the TIMSS database and individuals' total scores were calculated.

Relevant covariates were selected referring to prior studies that have explored what predictors affect test performance. Student background variables such as gender, the language in which tests were taken at home, and socioeconomic status-related variables such as breakfast at school and home educational resources were chosen as relatively stable predictors of student achievement (Wu, Debeer, Buchholz, Hartig, & Janssen, 2019). In addition, students' attitudes toward school and mathematics were selected as exploratory student psychological factor covariates. In detail, attitude toward school (BSBM15A–BSBM17G) includes statements such as "I like being in school" and "I feel like I belong at this school." Attitude toward mathematics includes "student likes mathematics" (BSBM17A–BSBM17I), "student feels confident in mathematics" (BSBM19A–BSBM19I), and "student values mathematics" (BSBM20A–BSBM20I); each includes statements such as "I enjoy learning mathematics," "I usually do well in mathematics," and "It is important to do well in mathematics." Table 2 presents the covariate variable names and descriptions. In this study, student motivation was operationally defined as intrinsic motivation by which a person is internally engaged in an activity itself (Deci, 1972) since intrinsic motivation makes more sense in a low-stakes

assessment; “student likes mathematics” responses were used to measure student motivation. In addition, some items were reverse-coded as necessary.

There were 1,236 examinees from 246 schools in the data after removing missing data from the initial 1,438 responses: There were 619 students who received items at the beginning (booklet 1) and 617 students who received items at the end (booklet 14). The data were composed of 596 (48%) male students and 640 (52%) female students.

Data Analysis

Descriptives. To understand the samples’ overall characteristics, descriptive statistics were employed prior to the main data analysis. Student background variables, socioeconomic status-related variables (such as gender and educational resources at home), and an outcome variable “total score” were explored; distributions were displayed from the whole sample and from each item position group.

Propensity score matching. Causal inference can be made by balancing covariate distributions using propensity scores as an additional analysis in order to provide the robustness of the simple group comparison using the data. A propensity score is the probability of receiving a treatment given observed covariates and potential outcomes (Rosenbaum & Rubin, 1983) defined as:

$$P(T_i = 1|X_i),$$

where T_i denotes the treatment received by examinee i and X_i denotes the vector of covariates for examinee i. The ACE for the treated can be obtained by propensity score matching. This takes the probability of receiving treatment into account and balances a wide range of covariates between groups; thus, systematic differences can be adjusted.

Propensity score matching is useful for estimating the causal impact by handling selection bias without randomization (Keller & Tipton, 2016).

To investigate and quantify the effect of item position while controlling for student motivation, treatment was defined as getting items at the end of a test. In other words, those receiving items at the end were defined as the treatment group (item block 01 from booklet 14), and those receiving items at the beginning were defined as the control group (item block 01 from booklet 1). In addition, student background variables, socioeconomic status-related variables, and student psychological factor variables were incorporated as covariates.

After examining the initial covariate imbalance and the mean difference in the mathematics total score between students who had the same items at the start and end, propensity score matching was conducted to calculate their average treatment effect. The key assumptions, which are the strong ignorability assumption that potential outcomes are independent of the treatment received given the covariates and that the propensity scores are in the range 0–1, were met.

Initially, students' propensity scores were estimated by fitting a logistic regression model. Overlaps were assessed both before and after eliminating non-overlapping cases to compare the distribution for each group and propensity scores were re-estimated using a subset of just the overlapping cases. Then, the propensity scores of students who received items at the end were matched with those of students who had received the same items at the beginning. After checking for the covariate balance such that treated and untreated cases had nearly identical distributions of covariates, an average treatment effect for the treated was calculated.

Study 2

Data

To examine whether considering student motivation effect with item position improves the model with and without missing data in different types, several missing data scenarios were applied to the complete dataset of TIMSS 2015 grade 8 mathematics data from the United States. Namely, cases were deleted and imputed to a complete dataset. First, to investigate the structural relationship between variables without missing data, a complete dataset was processed using the matched dataset as a result from Study 1; A subset consisting of just the overlapping cases from propensity score matching (n = 1,210) was used in the analysis. Then, to evaluate the performance of models with different missing data types under realistic conditions, missing data techniques were applied to the complete dataset.

While the general features of the data remain consistent with Study 1 as they share the same data, the responses for 17 mathematics items and 23 student motivation items were selectively used in Study 2 to fit the model. As mentioned earlier, student motivation items refer to "student likes mathematics" questions as student motivation is operationally defined as intrinsic motivation. In addition, in this part of the study, 17 mathematics item responses were recorded as dichotomous (0 or 1) since IRT models were employed by treating only fully correct responses as correct (1) and partially correct or incorrect responses as incorrect (0) for partial credit items. For the 23 student motivation questionnaires in a Likert scale form, a GRM was employed.

On this complete dataset, different types of missing mathematics data were manipulated and fitted to the SEM models to mimic how the students' dynamics were associated with item position and person predictor effects. Therefore, the data in Study 2

consisted of a complete dataset, imputed datasets with combinations of two missing data types, and two imputation methods:

Data 1: Complete dataset

Data 2: Missing values at MCAR, imputed by treating as incorrect

Data 3: Missing values at MCAR, imputed by random forest

Data 4: Missing values at MNAR, imputed by treating as incorrect

Data 5: Missing values at MNAR, imputed by random forest.

Using the "mice" package in R, responses were randomly sampled and set to NA for MCAR, and the last few responses from booklet 14 were set to NA for MNAR to represent the situation where students could not answer all questions. In other words, a monotonic missing data pattern was employed in MNAR; a monotone missing data pattern appears present if Y_j —the response of ordered item j —is missing, then all items Y_j with $k > j$ are also missing (Dong & Peng, 2013). Figure 4 presents the monotonic missing pattern used when generating missing data in this study. Then, the generated missing data were replaced in either method "treated as incorrect" or "random forest": when treating as incorrect, missing data were imputed as incorrect (0); for random forest imputing, the missing data were replaced with predicted values using multiple decision trees.

Data Analysis

Item calibration. Before fitting SEM models, 17 mathematics items were calibrated to describe the individual and overall item characteristics; although parameters were estimated within the SEM model, they were estimated prior to model fitting in a separate step to explore the items' descriptions. The item difficulty (b) parameter and

item discrimination (a) parameter estimates were obtained under the 2PL model using the "mirt" package in R. In addition to item parameters, item and test information for the whole dataset and each item position group were delivered. The mirt package is capable of analyzing various unidimensional and multidimensional IRT models with dichotomous and polytomous item responses (e.g., 1PL, 2PL, 3PL, and GRM). It uses multiple estimation methods including the expectation-maximization algorithm (EM; Bock & Aitken, 1981) and Metropolis–Hastings Robbins–Monro (MH-RM; Cai, 2010).

SEM model fitting. The causal relationship among various types and numbers of variables can be examined using the SEM framework. As SEM combines the features of factor analysis and regression, both measurement and structural models can be defined in an SEM model; measurement models of

$$x = \Lambda_x \xi + \delta$$

$$y = \Lambda_y \eta + \varepsilon,$$

and the structural model of

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta.$$

IRT models can be embedded into the SEM model as measurement models; a unidimensional IRT model is considered a one-factor FA model (Ferrando et al., 2013). Latent variables ξ and η correspond to ability parameters and the IRT-based latent trait θ. Fitting SEM models is useful for modeling item position effects with other relevant factors using the IRT framework.

To examine the effects of student motivation and item position on math ability in this study, student motivation and math ability are both treated as latent variables and item position is treated as a categorical grouping variable. The proposed SEM model ($Model\ 2$) can be written as:

$$\theta_M = \lambda_{\eta i}\theta_M + \varepsilon_i$$

$$\theta_m = \lambda_{\xi i}\theta_m + \delta_i,$$

where $\lambda_{\eta i}$ is the factor loading for mathematics item i, $\lambda_{\xi i}$ is the factor loading for motivation item i with measurement errors ε_i and δ_i, and

$$\theta_M = \gamma_1\theta_m + \gamma_2 P_i + \zeta_1,$$

where P_i is the position of item i, γ_1 is the effect size of students' motivation on math ability, and γ_2 is the effect size of item position on math ability.

The basic assumptions of error terms and the linear relationship assumption are met. The exogenous-to-endogenous latent variables are linear, as are the observed variables and their associated latent variables. In addition, there are constraints to ensure model identification: the mean and variance of the latent variables (θ_M and θ_m) are constrained to 0 and 1, which are assumed to follow normal distributions $\theta_M \sim N(0,1)$ and $\theta_m \sim N(0,1)$, respectively. In addition, the factor loadings of $\lambda_{\eta 1}$ and $\lambda_{\xi 1}$ are fixed at 1.

Model fits and effect sizes were compared to examine whether the student motivation effects on student ability varied between item positions under different student motivation levels and under different data missingness types. Three SEM models, $Models$ 0, 1, and 2, were specified and fitted for each dataset:

$$Model\ 0\text{: } Null\ model$$

$$Model\ 1\text{: } \theta_M = \gamma_2 P_i + \zeta_1$$

$$Model\ 2\text{: } \theta_M = \gamma_1\theta_m + \gamma_2 P_i + \zeta_1.$$

$Model$ 0 is the null model in which every variable is unrelated. In other words, all the covariances and paths are fixed at zero. On such a baseline model, $Model$ 1 adds the item

position effect, and then $Model\ 2$ adds the student motivation effect as a psychological factor. $Model\ 2$ is the proposed full model in this study. The "lavaan" package in R was used to fit the model. The package implemented diagonally weighted least squares (DWLS) estimation by default for categorical data; this estimation method is recommended when variables in data are not continuous (Xia & Yang, 2019).

Initially, a complete dataset was fitted to these three models. Model fit indices from each model were compared and effect sizes from the best model were examined. Then, the imputed missing data of MCAR and MNAR with two methods were fitted to these models using the "mice" package in R. Again, model fit indices were compared to determine whether any missingness type was associated with item position and student motivation effects on mathematics ability. Followed by model fit comparisons, effect sizes from each dataset were assessed to detect whether there was any meaningful implication when a certain type of data missingness existed.

To evaluate the model-data fit, the root mean square error of approximation (RMSEA; Steiger, 1990), comparative fit index (CFI; Bentler, 1990), and Tucker–Lewis index (TLI; Tucker & Lewis, 1973) were used under DWLS. These are the most common model fit indices in SEM:

$$RMSEA = \sqrt{\frac{\widehat{\Delta_M}}{df_M(N-1)}},$$

$$CFI = 1 - \frac{\widehat{\Delta_M}}{\widehat{\Delta_B}},$$

$$TLI = 1 - \frac{\chi_M^2/df_M}{\chi_B^2/df_B},$$

where M and B are defined for both the hypothesized model and the baseline model, $\widehat{\Delta_M} = \max(0,\ \chi_M^2 - df_M)$, and $\widehat{\Delta_B} = \max(0,\ \chi_B^2 - df_B)$. As seen from the definitions, while RMSEA compares a hypothesized model to a perfect model, CFI and TLI evaluate how far such a model is from a baseline model; smaller RMSEA and larger CFI and TLI values indicate better model–data fit. The general cutoff criteria of the indices are RMSEA < .05; CFI > .95; TLI > .95 (a 90% confidence interval for RMSEA is additionally addressed).

Chapter IV – RESULTS

This chapter presents study results. As with the results from Study 1 (conducted with real-world data), the average treatment effect on mathematics achievements were identified and compared. From Study 2 (conducted with missing data imposed on real-world data as described in the previous chapter), model–data fit indices and effect sizes were compared using five datasets based on data missingness and imputation types.

This chapter includes three sections: the first presents the detailed results from Study 1, the second describes the detailed results from Study 2, and the third summarizes the main findings.

Study 1

Descriptives

The treatment group (n = 619) was comprised of those who received the items at the end and the control group (n = 617) of those who received items at the beginning. Figure 5 presents both item position groups.

From the 1,236 responses in both item position groups, 899 (73%) students answered that they always spoke the test language at home, 220 (18%) selected “almost always,” 106 (9%) selected “sometimes,” and 11 (1%) selected “never.” Figure 6 presents the response distribution from the two groups. In terms of the “breakfast at school” question, which works as a socioeconomic index, 471 (38%) students answered that they never had breakfast at school, 243 (20%) selected “sometimes,” 255 (21%) selected “almost always,” and 267 (22%) answered that they always had breakfast at school. Figure 7 displays the response distribution from the two groups. Example

psychological factor covariates from the “student likes mathematics” and “student feels confident in mathematics” survey responses were additionally compared between groups in Figures 8 and 9, respectively. Overall, the covariate distributions were quite similar between groups except for a slight difference in psychological factor covariates. Lastly, the average total mathematics score was 10.375 out of 20, with a standard deviation of 4.984. Figure 10 presents the total mathematics score distributions between the two item position groups.

Simple Mean Difference

First, the initial balance was checked for the two item position groups. From the initial mean differences and variance ratio results (Table 3), most of the student background covariates and socioeconomic status-related covariates were balanced. In terms of student psychological factor covariates, the variables from “Student likes mathematics” (BSBM17A, BSBM17C, BSBM17D, BSBM17G, BSBM17H, BSBM17I) and those from the “Student feels confident in mathematics” (BSBM19A, BSBM19E, BSBM19H) questions were imbalanced although not seriously. Other psychological factor covariates were balanced. Figure 11 presents the initial covariate balance.

The *prima facie* estimate (naïve estimate) of item position effect on mathematics achievement was -0.620 (SE = 0.283) with p-value <.05. This is a simple difference in means from two item position groups without any potential covariates; results showed that students who had items at the end scored 0.620 points (items) lower in them compared to students who had the same items at the start. Figure 12 presents a boxplot displaying the difference between the two groups.

Propensity Score Matching

An average treatment effect for the treated using propensity score matching was estimated from this analysis. After estimating the propensity scores with logistic regression, the overlap was assessed before and after eliminating non-overlapping cases (Figures 13 and 14). With a subset of only overlapping data, propensity scores were re-estimated. After the overlap had been reassessed (Figure 15), 1,210 matched sets were created after running pair matching; Table 4 and Figure 16 each present the final mean differences, variance ratio results, and the final covariate balance from the matched data.

The result from the matched sets showed an almost perfect balance on covariates except for BSBM17G, BSBM17H, and BSBM19A, which also showed relatively higher mean differences before matching. Since these mean difference values have been lowered and were considered minimal, the treatment effect was calculated. The treatment effect from this model was -0.655 (SE = 0.286) with p-value <.05. Thus, the result from the propensity score matching approach was consistent with the result from prior analyses with a slight improvement in precision; students who had items at the end scored 0.655 points (items) less than students who had the same items at the beginning after controlling for relevant covariates. The matched dataset that resulted from this part of the study was used for Study 2.

Study 2

Item Calibration

Table 5 presents the overall item parameter estimates obtained from the whole dataset, including two item position groups; first, they were calibrated under the 2PL IRT model to review the items' overall descriptive statistics. For the 17 mathematics items,

the item difficulty parameters were in the range $^{-}$1.075 (M042202, the easiest item)–$^{+}$2.244 (M042302C, the most difficult item) and the item discrimination parameters were in the range 0.654 (M042159, the least discriminating question)–2.653 (M042302B, the most discriminating question). Figures 17 and 18 show item characteristic curves and item information functions, respectively, for each item.

Table 6 presents the item parameter estimates from each item position group. For the group that received items at the beginning, the item difficulty parameters were in the range $^{-}$1.620 (M042159)–$^{+}$2.017 (M042302C) and the item discrimination parameters were in the range 0.599 (M042159)–2.660 (M042240). For the group given the items at the end, the item difficulty parameters were in the range $^{-}$0.993 (M042182)–$^{+}$2.497 (M042302C) and the item discrimination parameters were in the range 0.693 (M042159)–2.843 (M042302A). The most difficult item and the least discriminating item were consistent between the groups; the overall item characteristic trends over the items within the group were also consistent between the groups.

Regarding the ranges, the calibrated item parameters indicated that equal items tended to function as more difficult and more discriminating when they appeared at the end of the test compared to at the beginning. However, examining individual item parameters revealed that the first six items (M042182–M042302A) were more difficult for students who had received them at the beginning while the latter 11 items (M042302B–M042167) were more difficult for students who had received them at the end; the first and last items were harder than the same items in the opposite location. Additionally, test information functions for each item position group were displayed to determine whether tests with the same items in different positions functioned differently (Figures 19 and 20); they presented different patterns of functions between the groups.

SEM Model Fitting

Complete data. After exploring the item characteristics of the descriptive purpose, three proposed models were used to fit the complete data, all of which converged in less than a minute. The model fit indices were compared and the coefficients for item position and motivation effects were evaluated. Table 7 shows the model fit indices from three SEM models: $Model\ 0$ exhibited a CFI of 0.984, TLI of 0.983, and RMSEA of 0.093; $Model\ 1$ exhibited a CFI of 0.984, TLI of 0.984, and RMSEA of 0.087; $Model\ 2$ exhibited a CFI of 0.997, TLI of 0.997, and RMSEA of 0.039. $Model\ 2$, with both item position and motivation variables, outperformed $Models\ 0$ and 1 with the best model fit based on all indices. $Model\ 1$, with the item position variable, only showed a very slight improvement in RMSEA compared to the baseline model.

Table 8 presents each model's estimated regression coefficients for item position and motivation effects on mathematics ability. The results indicated that increasing item position (receiving items at the end) had a significant negative effect on mathematics ability in both $Models\ 1$ and 2. In terms of student motivation, motivation had a positive effect on mathematics ability. With respect to the effect sizes based on $Model\ 2$, which was the best model, the effect size of the item position on mathematics ability was -0.143 (γ_2 = -0.143, $p < .05$), and that of student motivation on mathematics ability was 0.297 (γ_1 = 0.297, $p < .05$). In addition, including motivation in the model slightly increased the coefficient for the item position effect in $Model\ 2$ compared to the coefficient from $Model\ 1$.

Missing data. Table 9 presents a summary of the model fit information across different missing data and imputation types. First, upon comparing the models fitted with all datasets—complete dataset, MCAR and MNAR datasets imputed with the "treated as incorrect" approach, and "random forest"—$Model\ 2$ again showed a significantly better model fit than the other two models based on RMSEA and its 90% CI. The RMSEA indices for $Model\ 2$ were in the range 0.037–0.054 while those for $Model\ 0$ were 0.090–0.093 and those for $Model\ 1$ were 0.087–0.090. The results were consistent across all data types, indicating that considering student motivation and item position effects improves the model fit.

In comparing missing data types, regardless of imputation method, the model fit indices indicated that imputed MCAR data showed the best model fit, followed by imputed MNAR data and complete data in $Models\ 0$ and 1. However, a different model fit pattern was found in $Model\ 2$; while the imputed MCAR data consistently showed the best fit, the imputed MNAR data showed notably different model fit indices depending on the imputation method. In detail, the MNAR data showed equal model fit with complete data (RMSEA = 0.039 [90% CI = 0.036, 0.042]) when imputed with the random forest method; however, when imputed with the treated as incorrect method, the MNAR data showed a noticeably worse model fit (RMSEA = 0.054 [90% CI = 0.051, 0.057]) than the complete data and other dataset types in $Model\ 2$.

In terms of imputation methods, there was almost no difference in the model fit results between the two imputation methods throughout the models with either MCAR data or MNAR data in $Models\ 0$ and 1. Exceptionally, when missing data were present in MNAR in $Model\ 2$, the best-fitting model, the random forest method outperformed the

treated as incorrect approach. Figure 21 displays the overall comparison of RMSEA values between the models, missing data types, and imputation methods.

Lastly, Table 10 displays the estimated regression coefficients for item position and motivation effects on mathematics ability from $Model\ 2$ with different missing data types. Consistent with the model fit results, the MCAR data showed quite similar effect sizes to those with complete data in terms of both item position and student motivation effects, while the MNAR data showed some differences in coefficients from those of other data types. In detail, the coefficient of item position on mathematics ability with the MCAR data imputed with the treated as incorrect method was -0.150, and that with MCAR data imputed using random forest was -0.145. The effect of student motivation on mathematics ability with MCAR data imputed with the treated as incorrect method was 0.293, and that with MCAR data imputed using random forest was 0.301. With the MNAR data treated as incorrect, the item position effect size was -0.612 and the motivation effect size was 0.253, whereas the item position effect size was -0.240 and the motivation effect size was 0.287 with MNAR data imputed with random forest. All effects were significant with p-values of <.05.

Thus, when MNAR data missingness was treated as incorrect, the item position effect was inflated while the student motivation effect was smaller than those from the complete data. When MNAR data missingness was replaced using random forest, the item position effect was still large but less so. In addition, the student motivation effect size was increased and closer to those from the complete data. Thus, when missing data is present in MNAR, imputing missing data using the random forest approach overcomes biased estimates from imputing missing data as incorrect.

Summary

This work explored three research questions: First, whether item position significantly affects mathematics achievement, additionally accounting for relevant covariates mainly including psychological factor variables; second, whether including the student motivation factor effect improves the item position effect model; third, whether such effects have any meaningful association with different types of data missingness.

The results of Study 1 revealed that there is a significant item position effect when simply comparing item position groups as well as when accounting for student background variables, socioeconomic status-related variables, and psychological factor variables as covariates. Students who received items at the end scored more than half a point lower than students who had those items at the beginning before and after controlling for covariates. Following Study 1, which examined the existence of the item position effect, Study 2 suggested that one of the most commonly researched psychological factors, namely student motivation, improves the item position effect model. The full model with student motivation effect added to the item position effect performed the best among the three proposed models and in the best model, both item position and student motivation significantly affected mathematics ability. Item position showed a significant negative effect on mathematics ability while student motivation showed a significant positive effect on mathematics ability. In addition, when missing data were present, the MCAR missingness type showed a close model fit and effect sizes as those of the complete dataset, while the MNAR missingness type showed a worse model fit and inflated effect sizes, particularly when the missing data were treated as incorrect.

Chapter V – DISCUSSION

The item position effect is a source of item parameter drift, which can bias assessment analyses. In practice, locating items across test forms is a common strategy when administering tests, particularly in large-scale assessments for multiple purposes: using well-developed items multiple times; equating and scaling test items with a common item set; ensuring test security and preventing cheating; and testing as many items as possible. Although tests are carefully designed, a different item location could result in different item characteristics and different test information. This work started by asking 1) whether a significant item position effect exists; 2) whether any psychological factors are involved in the effect; and 3) whether any test-taking scenarios are involved in the effect.

Thus, this work's purpose was to propose a model that includes student dynamics with the item position effect and student motivation as a psychological factor. Additionally, the missing data mechanism was considered to represent student dynamics related to the item position effect, namely, speed effect. Three research questions were addressed to achieve this goal. To examine the item position effect on students' mathematics performance accounting for student motivation with different missing data types, as a prerequisite aspect of the study, the existence of item position effects was identified, and the robustness of the simple group comparison was checked using propensity score matching (Research Question 1). Following the first study, SEM models including student motivation with the item position effect were evaluated (Research Question 2), and different types of deleted-and-imputed missing datasets were fitted to

the SEM models (Research Question 3). All analyses used TIMSS 2015 grade 8 mathematics data from the U.S. as exemplary large-scale assessment data.

The main findings from this study were as follows: 1) A significant item position effect on mathematics achievements was identified; additionally, the treatment effect size was increased slightly after accounting for the relevant covariates of student background, socioeconomic status, and psychological variables; 2) the full model ($Model\ 2$) with both student motivation and the item position effect was revealed as the best with the complete data; 3) the MNAR missingness type was found to have meaningful information that needed to be considered in test administration.

In Study 1, propensity score matching was used to assess the pure item position effect with covariates following the simple group comparison. The results from this part of the study indicated that there was a significant item position effect depending on items' location. Students who had items at the end scored 0.655 less than those who had the same items at the beginning. This result supported the findings of prior studies that items tended to become more difficult to get correct when they appeared in a latter section of the test (Le, 2007; Meyers et al., 2008). Considering that each item was worth one point in this study, the difference was more than half a question; as plausible values were calculated in actual test administrations, the difference between item position groups will increase when scaled.

Another finding from this result was that balancing covariates gave a slightly better causal effect for item position in terms of both magnitude and accuracy. The covariates of student background variables and socioeconomic status-related variables confirmed that they were stable predictors of student performance as mentioned in previous research (Wu et al., 2019). Furthermore, the main focus on covariates of student

psychological factor variables in this study gave the rationale to proceed to propose SEM models including the student motivation effect. This led to the conclusion that person predictors, which are not content-related variables, need to be brought into the model when examining the item position effect.

Lastly, the results indicated that two item position groups were already well balanced since the average treatment effect results from the simple mean difference and propensity score matching were quite similar. Moreover, the initial imbalance was not too serious; it was expected because TIMSS is a very well-structured large-scale assessment study for which a series of assessments has been conducted every four years since 1995. However, the minimal yet significant item position effect in TIMSS indicates that there could be greater item position effects in most other assessments.

In Study 2, in addition to the significant item position effect from Study 1, a SEM model was proposed that included student motivation and item position effects; one of the psychological factor covariates from Study 1—the student motivation variable—was specifically defined as a variable with intrinsic motivation questions because intrinsic motivation would make the most sense regarding the nature of a low-stakes assessment such as TIMSS. Three SEM models from the baseline model to the full model were evaluated using a complete dataset; the full model showed the best fit as expected. While adding the item position effect only from the null model very slightly improved the model fit, including an additional student motivation effect in the item position effect model considerably improved it. The model was proposed as an extension—both revealing the item position effect and attempting to explain it; the result indicates that item position effects should be studied with explanatory person predictor variables to better understand the effect. Previous studies also suggested addressing “the why

question" on item position effects research (Debeer & Janssen, 2013; Weirich et al., 2017).

Using the best-fitting model—Model 2—the effect sizes of item position and student motivation on mathematics ability showed that the item appearing at the end negatively affected mathematics ability and that student motivation positively affected mathematics ability; both coefficients were statistically significant. This result agreed with prior studies that concluded that positioning an item in the latter part of a test negatively affected mathematics performance when controlling for motivation. From this study, we can conclude that mathematics abilities for students with lower motivation could be affected more negatively when the items appear at the end than for students with higher motivation.

One interesting additional point of note was that items showed two different patterns regarding item position groups; within an item block, the first six items tended to be more difficult when items were given at the beginning and the latter 11 items tended to be more difficult when given at the end, which was determined from the item calibration conducted as a description of items. In other words, items were deemed more difficult when at the beginning and end positions within a booklet.

Lastly, following the complete dataset, imputed missing datasets were fitted to the SEM models. Consistent with the complete dataset, the full model performed the best among data types. In terms of model fit comparison, the MCAR datasets did not show much difference from the complete data regardless of the imputation method; the fit indices were slightly better. This is unsurprising because the missing data were completely random and had no relevance to the factors in the model. Nevertheless, the MNAR datasets showed a noticeable difference depending on the imputation method;

when the MNAR missing data were imputed as incorrect and fitted to the model, the model fit was the worst. Meanwhile, when the MNAR data were imputed with the random forest method, the model fit indices were equal to those of the complete data. A worse model fit was expected when data missingness was present in MNAR since cases were dropped in a specific missing pattern for MNAR to represent the scenario in which students could not finish the test in time. However, imputing these missing data with random forest improved the model fit, while treating these missing data as incorrect showed the worst model fit. These results indicate that simply scoring missing responses as incorrect would discard meaningful test information in that there was a non-ignorable situational missingness pattern—in this study, a lack of time.

The item position and student motivation effect sizes were also influenced by the missing data type; consistent with the model fit results, MCAR data did not show a significant difference while the MNAR data showed considerably larger item position effects and generally a smaller motivation effect. Again, the inflation in the item position effect and the slight decrease in the motivation effect were more severe when the MNAR data missingness was imputed as incorrect rather than when MNAR data were imputed with random forest. This result could be considered natural; however, it gives a practical emphasis that using the item position effect as a testing bias could be huge when students are unable to finish a test but are scored as incorrect, which is the exact scenario of MNAR missing data.

The meaningful contributions of this study are as follows: first, this work started with realistic and practical curiosity regarding what psychological factor variables are considered in item position effects. The current approach in item position effect research is to add person-related predictors to explain this phenomenon. Student psychological

factor variables were mainly included in this work by extending the typically suggested relevant predictors such as students' background, socioeconomic status, and test-related variables. Second, propensity score matching approach was additionally employed to calculate the item position effect considering such covariates. Since calculating the difference between the item position groups was the first approach in previous item position effect studies, the propensity score matching approach could be an important step in calculating the treatment effect more accurately and checking the robustness of the group comparison. Third, this study embedded the IRT framework in SEM models to include all variables in one model. This approach enabled the extension of the current framework of item position effect research that only accounts for item-related factors. Once the IRT models for both dichotomous and polytomous responses are included in the SEM model, it can be extended to be unlimited to improve the understanding of the dynamics behind these effects. Lastly, the missing data mechanism was added to this work to mimic the actual scene when taking tests. Missing data types and imputation methods are often researched with importance placed on parameter recovery; however, missing data were used in this study to represent potential scenarios that could be closely related to item position effects. Thus, realistic item bias possibilities could be found by analyzing different missing data types with different imputation methods.

Finally, the practical implications of this work are as follows: first, when administering different test forms, ensuring directly comparable results across forms should be heavily emphasized regardless of the test type in the test design. Large-scale assessments were developed with opportunities for refinement from multiple administrations; however, most exams still had a single administration. Therefore, more careful test design is required to measure student performance accurately without biased

item parameter estimates. Second, appropriate testing time should be provided to prevent loss of meaningful information. As lack of time appears to be a common reason for missing data in a testing context, investigating the patterns of missingness is necessary when missing data is present. Reviewing missing patterns would improve both test administration and student performance measurement. Lastly, person predictors should be taken into account since not all test takers are equally sensitive to testing dynamics. Further approaches to establish person factors such as test-taking abilities could be considered.

Limitations and Future Research

This study has certain limitations. First, regarding non-randomization in the TIMSS data, TIMSS was very carefully designed to cover non-randomization as mentioned above. It would be beneficial to check with other less balanced assessments to explore item position effects in greater depth. Second, despite the testing time being set to give students sufficient time to answer all questions properly after several years of administration, TIMSS remains a speed test rather than a power test. As item position effects could be affected by lack of time in a speed test, this work has tried to overcome this limitation by employing the missing data mechanism and directly exploring the speed issue. Finally, student motivation was estimated under GRM from TIMSS student questionnaire responses. It would make more sense with the item position effect if student motivation changes could be tracked throughout the assessment. A simulation study with such conditions would be meaningful for overcoming this limitation.

This study's current model only includes student motivation with the item position effect on mathematics ability. As student motivation in this work is referred to as

intrinsic motivation, it would be interesting to include additional variables in the model such as extrinsic motivation or confidence. For another model extension, as mentioned above, it would be more informative to measure such psychological variables at the beginning and end for inclusion in the model. A simulation study regarding such conditions could be conducted in the future. Moreover, this study had two item position groups: students who had received items at the start and at the end. As seen from the item calibration in the beginning of Study 2, item parameters showed different patterns within one item position group; more reliable findings could be obtained if item position groups were divided into four locations rather than just two. Finally, another potential extension of the model for future research would be to include the response time. As the speed effect and missing data mechanism were studied, considering the response time within the model could be an important extension to gain a better understanding of the item position effect.

REFERENCES

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*(4), 408–426.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for Educational and Psychological Testing.* Amer Educational Research Assn.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques.* CRC Press.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores.*

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

Borgatta, E. F., & Corsini, R. J. (1960). The quick word test. *The Journal of Educational Research, 54*(1), 15–19.

Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics, 23*(3), 236–243.

Breslow, N. E., Day, N. E., & Schlesselman, J. J. (1982). Statistical methods in cancer research. Volume 1—The analysis of case-control studies. *Journal of Occupational and Environmental Medicine, 24*(4), 255–257.

Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education, 3*(1), 7–16.

Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large-scale Assessments in Education, 5*(8), 1–20.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*(1), 33–57.

Davis, J., & Ferdous, A. (2005, April). *Using item difficulty and item position to measure test fatigue.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164–185.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533–559.

Deci, E. L. (1972). Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personality and Social Psychology*, *22*(1), 113–120.

deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. *Breakthroughs in Statistics: Foundations and Basic Theory*, 599–609.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. SpringerPlus, 2(1), 222.

Ferrando, P. J., Anguiano-Carrasco, C., & Demestre, J. (2013). Combining IRT and SEM: A hybrid model for fitting responses and response certainties. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(2), 208–225.

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

Flaugher, R. L., Melton, R. S., & Myers, C. T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement, 28*(3), 813–824.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science, 50*(3), 379–390.

Hancock, D. R. (2001). Effects of test anxiety and evaluative threat on students' achievement and motivation. *The Journal of Educational Research, 94*(5), 284–290.

Happ, R., & Förster, M. (2018). The correlation between vocational school students' test motivation and the performance in a standardized test of economic knowledge: Using direct and indirect indicators of test motivation. *Empirical Research in Vocational Education and Training, 10*(1), 10.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*(6), 497–509.

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics, 41*(3), 326–348.

Kingston, N. M., & Dorans, N. J. (1982). The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory. *ETS Research Report Series, 1982*(1), i–26.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*(2), 147–154.

Klein, S. P., & Bolus, R. (1983). *The effect of item sequence on bar examination scores* (No. RAND/P-6857). RAND CORP SANTA MONICA CA.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*(2), 232–244.

Le, L. T. (2007). Effects of item positions on their difficulty and discrimination: A study in PISA Science data across test language and countries.

Li, F., Cohen, A., & Shen, L. (2012). Investigating the effect of item position in computer-based tests. *Journal of Educational Measurement, 49*(4), 362–379.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement, 7*(2), 113–118.

Martin, M. O., Mullis, I. V., & Foy, P. (2015). TIMSS 2015 assessment design. *TIMSS*, 85–99.

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38–60.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education, 43*(4), 236–254.

Mullis, I. V., Martin, M. O., & Sainsbury, M. (2016). PIRLS 2016 reading framework. *PIRLS*, 11–29.

Musekamp, F., & Pearce, J. (2016). Student motivation in low-stakes assessment contexts: an exploratory analysis in engineering mechanics. *Assessment & Evaluation in Higher Education, 41*(5), 750–769.

Programme for International Student Assessment, & SourceOECD (Online service). (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* Publications de l'OCDE.

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance. *Journal of Educational Measurement, 19*(1), 49–57.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodological challenges in the calibration of achievement tests]. In O. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss, & G. Walther (Eds.), *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule*, 42–106. Weinheim, Germany: Beltz.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement.*

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279–313.

Schweizer, K.D., & Ren, X. (2013). The position effect in tests with a time limit: The consideration of interruption and working speed. *Psychological Test and Assessment Modeling, 55*(1), 62–78.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education, 2*(1), 9.

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535–548.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*(2), 115–129.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(1), 77–87.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In *Explanatory item response models* (pp. 43–74). Springer, New York, NY.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17.

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, *7*(1), 5.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10–16.

TABLES AND FIGURES

Table 1

TIMSS 2015 booklet design

Student Achievement Booklet	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

Table 2

Covariate names and descriptions

Variable name	Description
ITSEX	Sex of student
BSBG03	Often speak language of test at home
BSBG12	How often breakfast on school days
BSBGHER	Home Educational Resources/ Scaled
BSDGHER	Home Educational Resources/ Categorized
BSBG15A	Being at school
BSBG15B	Safe at school
BSBG15C	Belong at school
BSBG15D	Like to see classmates
BSBG15E	Fair teachers
BSBG15F	Proud to go to this school
BSBG15G	Learn a lot
BSBM17A	Enjoy learning mathematics
BSBM17B	Wish have not to study math
BSBM17C	Math is boring
BSBM17D	Learn interesting things
BSBM17E	Like mathematics
BSBM17F	Like numbers
BSBM17G	Like math problems
BSBM17H	Look forward to math class
BSBM17I	Favorite subject
BSBM19A	Usually do well in math
BSBM19B	Mathematics is more difficult
BSBM19C	Mathematics not my strength
BSBM19D	Learn quickly in Mathematics
BSBM19E	Mathematics makes nervous

BSBM19F	Good at working out problems
BSBM19G	I am good at Mathematics
BSBM19H	Mathematics harder for me
BSBM19I	Mathematics makes confused
BSBM20A	Mathematics will help me
BSBM20B	Mathematics to learn other things
BSBM20C	Mathematics to get into university
BSBM20D	Mathematics to get the job I want
BSBM20E	Job involving Mathematics
BSBM20F	Get ahead in the world
BSBM20G	More job opportunities
BSBM20H	Parents think Mathematics important
BSBM20I	Important to do well in Mathematics

Table 3

Initial mean difference and variance ratio of two item position groups

Variable name	St Mean Diff	Var Ratio
ITSEX	-0.055	1.004
BSBG03	-0.052	1.072
BSBG12	-0.035	0.952
BSBG15A	-0.075	0.931
BSBG15B	-0.073	1.157
BSBG15C	-0.017	1.032
BSBG15D	-0.037	1.102
BSBG15E	-0.027	1.067
BSBG15F	-0.047	1.095
BSBG15G	-0.026	1.032
BSBM17A	-0.121	1.012
BSBM17B	-0.074	0.994
BSBM17C	-0.114	0.991
BSBM17D	-0.114	1.164
BSBM17E	-0.079	0.979
BSBM17F	-0.089	0.981
BSBM17G	-0.138	0.986
BSBM17H	-0.141	0.919
BSBM17I	-0.097	0.942
BSBM19A	-0.120	1.092
BSBM19B	-0.039	0.998
BSBM19C	-0.099	1.044
BSBM19D	-0.034	1.079
BSBM19E	-0.111	1.080
BSBM19F	-0.067	1.042
BSBM19G	-0.078	1.106

BSBM19H	-0.106	1.029
BSBM19I	-0.085	0.947
BSBM20A	-0.087	1.100
BSBM20B	0.010	0.906
BSBM20C	0.086	0.918
BSBM20D	0.049	1.030
BSBM20E	-0.055	1.027
BSBM20F	0.020	1.014
BSBM20G	0.041	0.983
BSBM20H	0.065	0.916
BSBM20I	0.056	0.932
BSBGHER	0.047	0.849
BSDGHER	-0.044	0.861

Table 4

Final mean difference and variance ratio of two item position groups

Variable name	St Mean Diff	Var Ratio
ITSEX	-0.050	1.004
BSBG03	-0.046	1.054
BSBG12	-0.017	0.956
BSBG15A	-0.054	0.904
BSBG15B	-0.056	1.096
BSBG15C	-0.021	1.042
BSBG15D	-0.034	1.080
BSBG15E	-0.021	1.065
BSBG15F	-0.036	1.052
BSBG15G	-0.029	1.043
BSBM17A	-0.092	0.991
BSBM17B	-0.058	0.993
BSBM17C	-0.085	1.002
BSBM17D	-0.086	1.149
BSBM17E	-0.065	0.980
BSBM17F	-0.075	0.970
BSBM17G	-0.109	0.980
BSBM17H	-0.110	0.925
BSBM17I	-0.084	0.953
BSBM19A	-0.101	1.073
BSBM19B	-0.048	1.004
BSBM19C	-0.091	1.048
BSBM19D	-0.042	1.095
BSBM19E	-0.094	1.077
BSBM19F	-0.050	1.044
BSBM19G	-0.066	1.097

BSBM19H	-0.093	1.026
BSBM19I	-0.073	0.946
BSBM20A	-0.061	1.047
BSBM20B	0.013	0.891
BSBM20C	0.066	0.938
BSBM20D	0.042	1.046
BSBM20E	-0.031	1.033
BSBM20F	0.022	1.022
BSBM20G	0.034	0.985
BSBM20H	0.047	0.958
BSBM20I	0.042	0.967
BSBGHER	0.033	0.869
BSDGHER	-0.043	0.872

Table 5

Item parameters for the whole data

Item	Discrimination (a)	Difficulty (b)
M042182	0.901	-0.758
M042081	1.196	-0.134
M042049	0.848	-0.697
M042052	1.875	-0.806
M042076	0.854	-0.299
M042302A	2.529	-0.016
M042302B	2.653	0.03
M042302C	1.028	2.244
M042100	1.695	-1.057
M042202	1.454	-1.075
M042240	2.35	-0.572
M042093	2.318	0.935
M042271	1.115	-0.477
M042268	0.839	1.542
M042159	0.654	-0.993
M042164	1.856	-0.078
M042167	2.042	0.847

Table 6

Item parameters for two item position groups

	Part 1		Part 2	
Item	a	b	a	b
M042182	0.971	-0.511	0.894	-0.993
M042081	1.046	-0.122	1.380	-0.145
M042049	0.847	-0.659	0.866	-0.724
M042052	1.838	-0.759	2.013	-0.837
M042076	0.791	-0.216	0.945	-0.368
M042302A	2.320	0.006	2.843	-0.041
M042302B	2.609	-0.020	2.694	0.076
M042302C	0.961	2.017	1.118	2.497
M042100	1.682	-1.167	1.691	-0.958
M042202	1.288	-1.282	1.603	-0.915
M042240	2.660	-0.697	2.124	-0.444
M042093	2.068	0.862	2.660	1.006
M042271	1.100	-0.725	1.122	-0.245
M042268	0.847	1.478	0.835	1.604
M042159	0.599	-1.620	0.693	-0.497
M042164	1.517	-0.237	2.256	0.053
M042167	2.225	0.705	1.870	1.006

Table 7

Model fit comparison with complete data

Model	CFI	TLI	RMSEA	Lower CI	Upper CI
M0	0.984	0.983	0.093	0.090	0.096
M1	0.984	0.984	0.090	0.087	0.093
M2	0.997	0.997	0.039	0.036	0.042

Table 8

Effect Sizes on mathematics ability with complete data

	Item Position			Student Motivation		
Model	b	SE	p-value	b	SE	p-value
M0	-	-	-	-	-	-
M1	-0.140	0.062	0.025	-	-	-
M2	-0.143	0.065	0.027	0.297	0.034	0.000

Table 9

Model fit comparison with missing data

			RMSEA		
Data	Imputation	Model		Lower CI	Upper CI
Complete	-	M0	0.093	0.090	0.096
		M1	0.090	0.087	0.093
		M2	0.039	0.036	0.042
MCAR	Treat as Incorrect	M0	0.090	0.087	0.093
		M1	0.087	0.084	0.089
		M2	0.037	0.034	0.040
MCAR	Random Forest	M0	0.091	0.088	0.094
		M1	0.088	0.085	0.091
		M2	0.037	0.034	0.040
MNAR	Treat as Incorrect	M0	0.091	0.088	0.094
		M1	0.088	0.085	0.091
		M2	0.054	0.051	0.057
MNAR	Random Forest	M0	0.091	0.088	0.094
		M1	0.088	0.086	0.091
		M2	0.039	0.036	0.042

Table 10

Effect Sizes on Mathematics Ability with missing data

			Item Position			Student Motivation		
Data	Imputation	Model	b	SE	p-value	b	SE	p-value
Complete	-	M2	-0.143	0.065	0.027	0.297	0.034	0.000
MCAR	Treat as Incorrect	M2	-0.150	0.065	0.022	0.293	0.034	0.000
MCAR	Random Forest	M2	-0.145	0.065	0.026	0.301	0.034	0.000
MNAR	Treat as Incorrect	M2	-0.612	0.066	0.000	0.253	0.034	0.000
MNAR	Random Forest	M2	-0.240	0.065	0.000	0.287	0.034	0.000

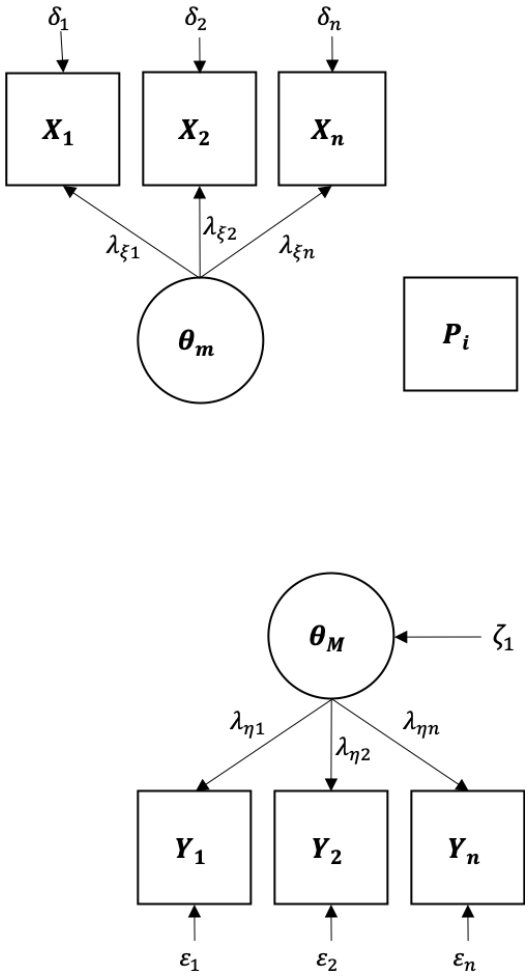

Figure 1. Null model (Model 0)

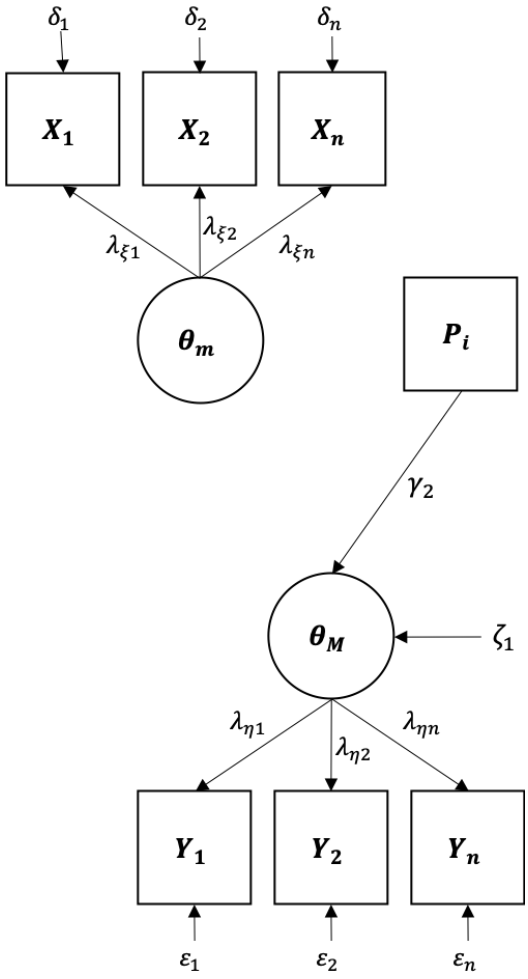

Figure 2. Item position effect model (Model 1)

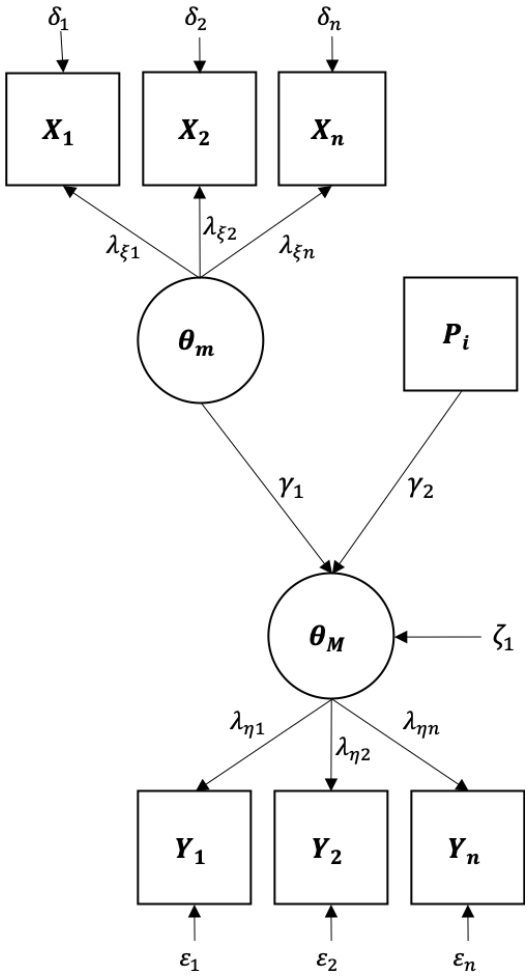

Figure 3. Item position and motivation effect model (Model 2)

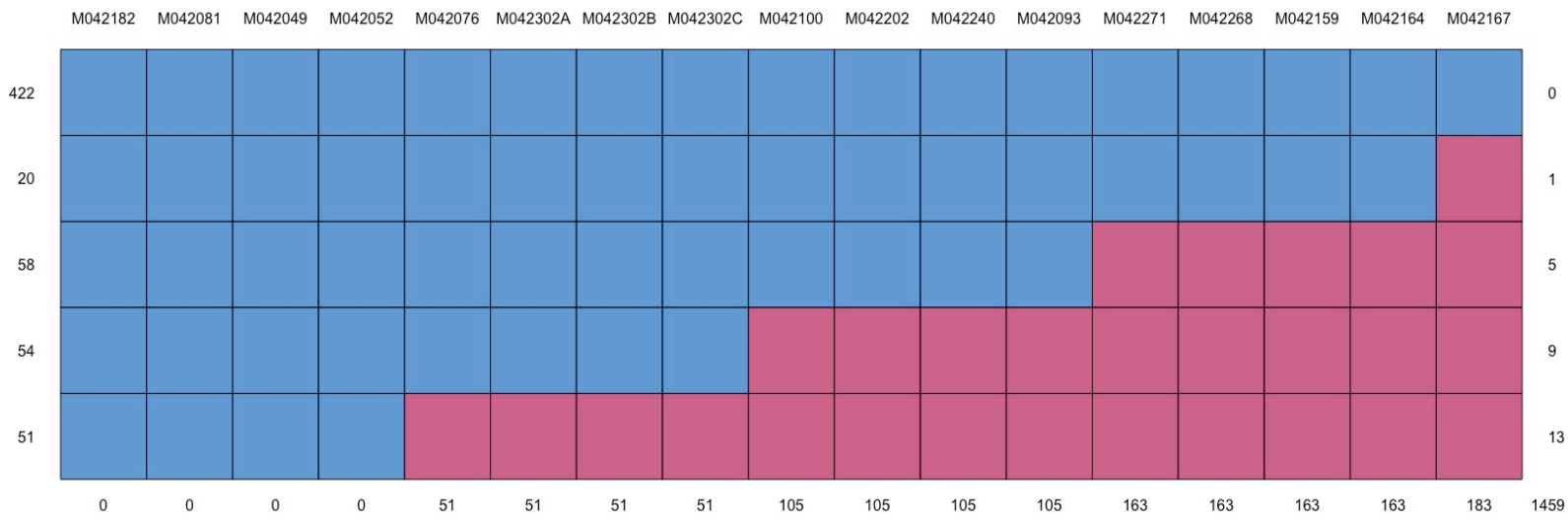

Figure 4. Monotonic missing data pattern for MNAR

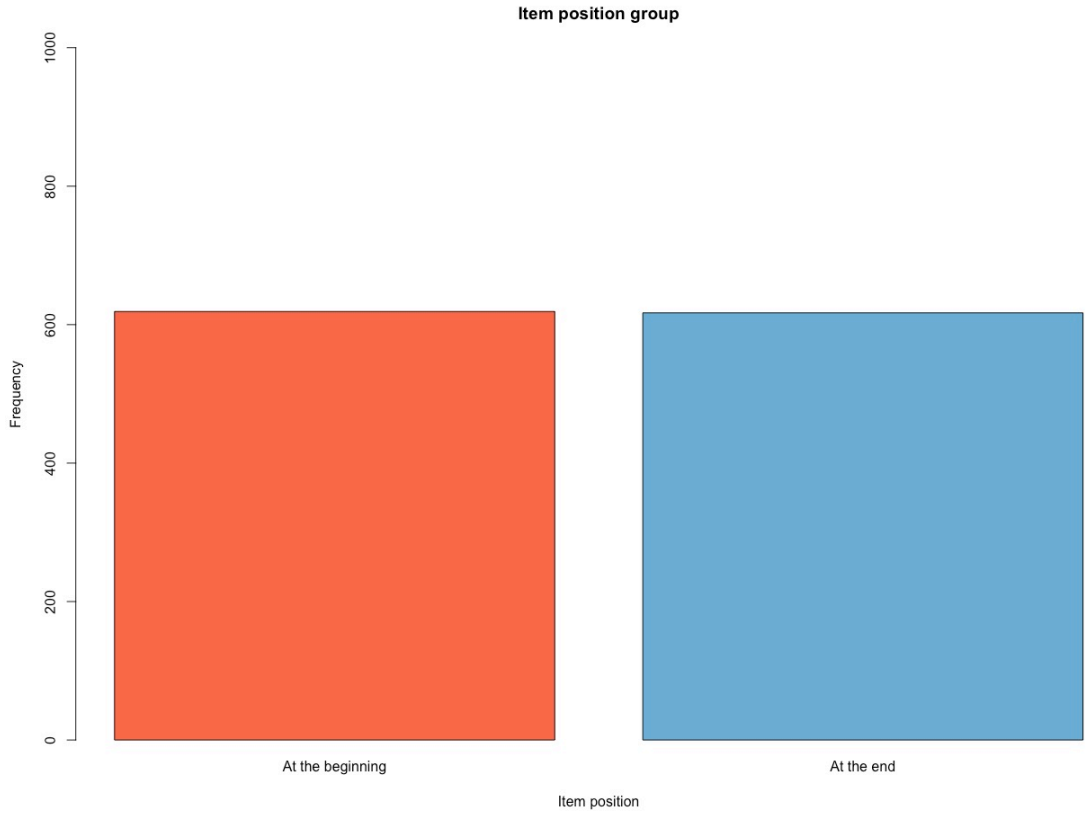

Figure 5. Item position groups

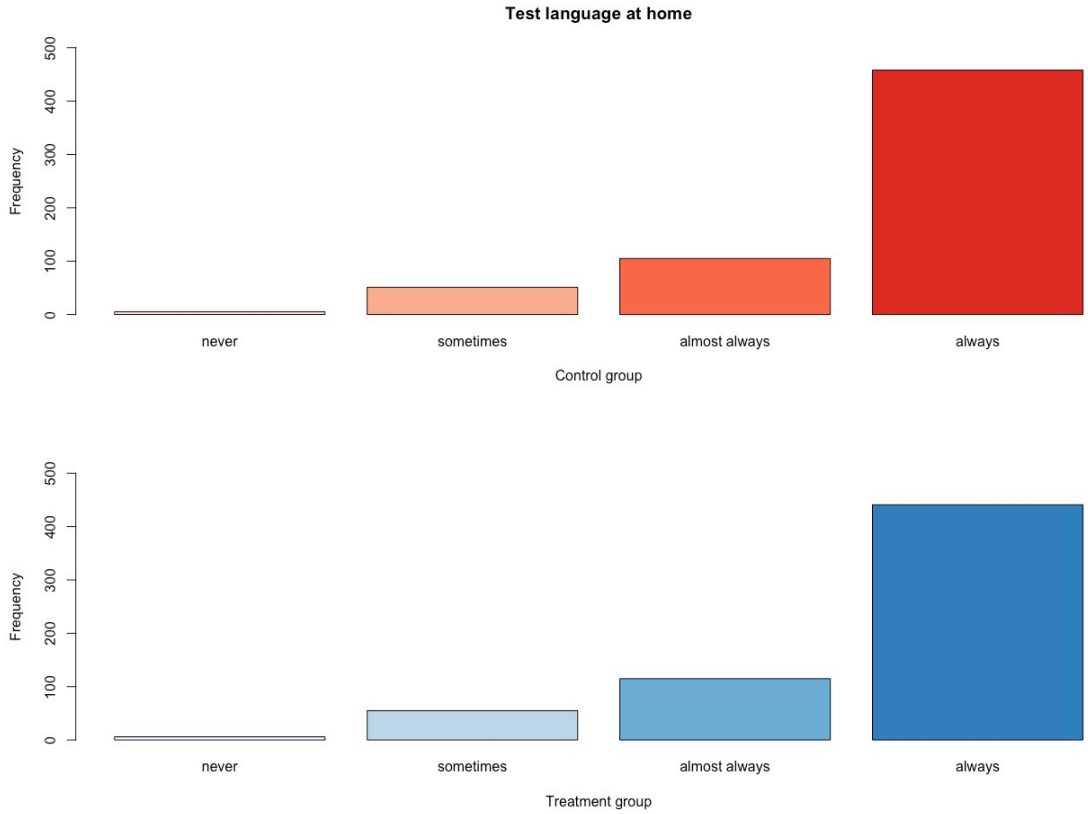

Figure 6. Test language at home response distributions

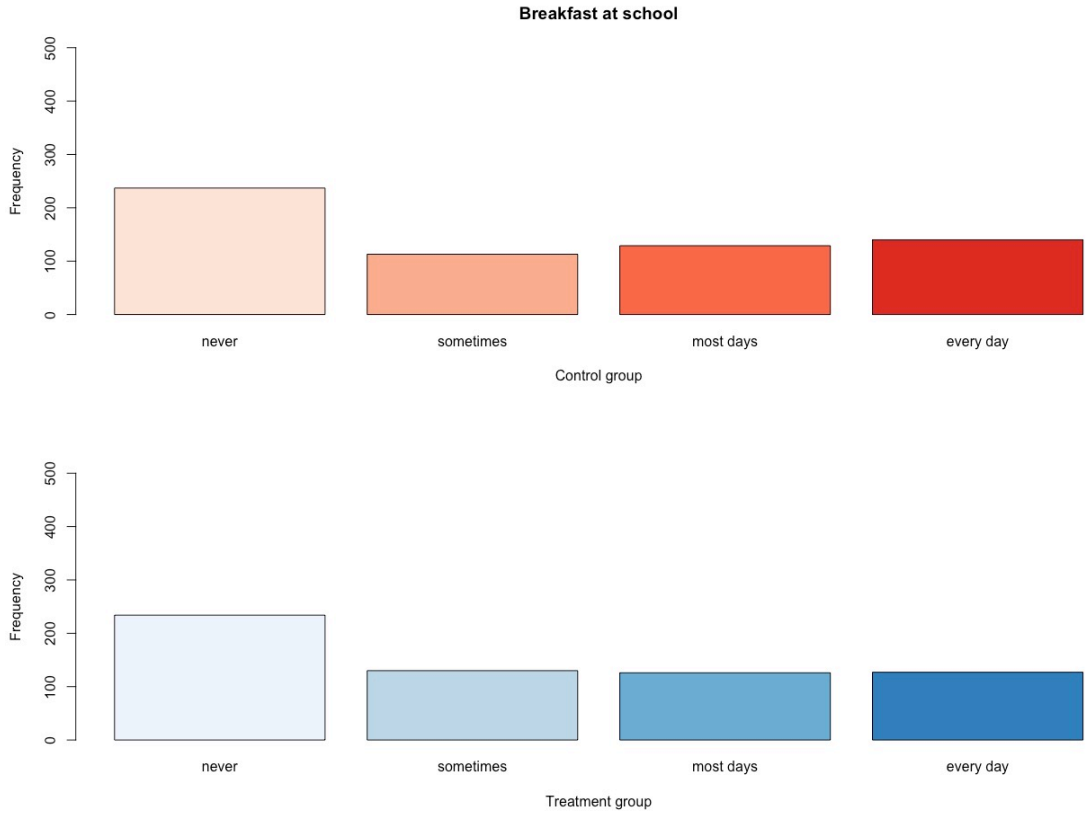

Figure 7. Breakfast at school response distributions

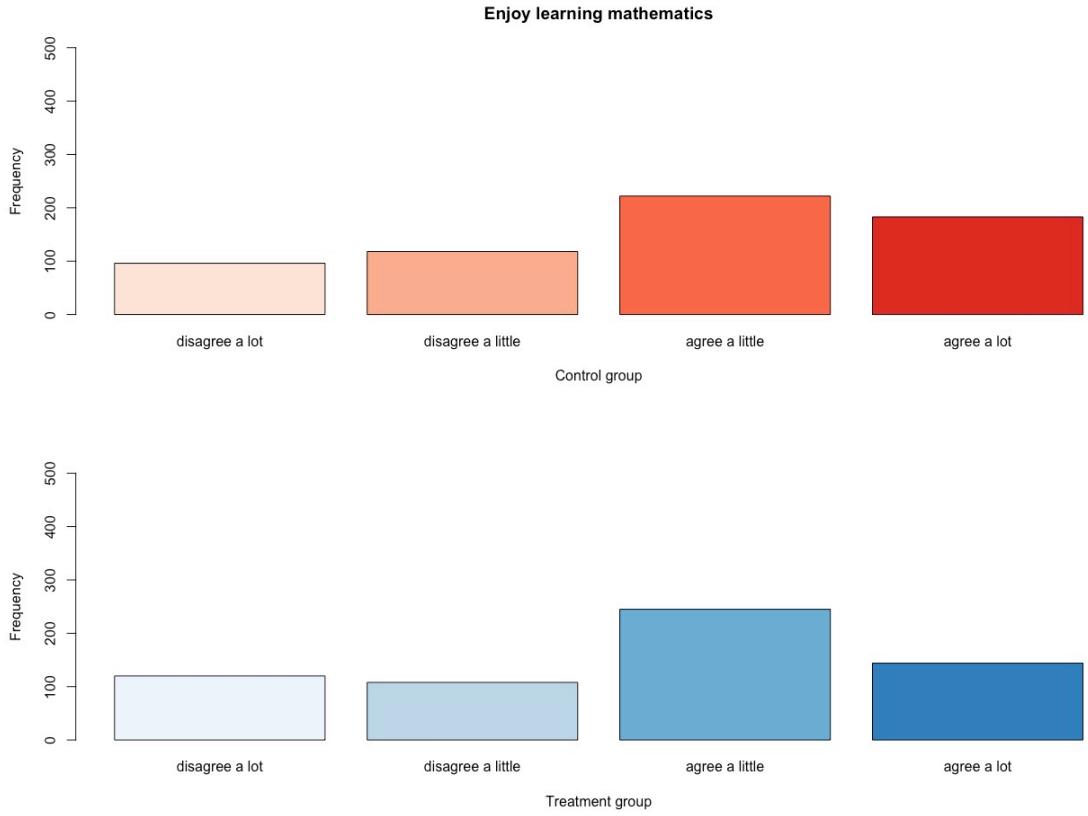

Figure 8. Example "Student likes mathematics" response distributions

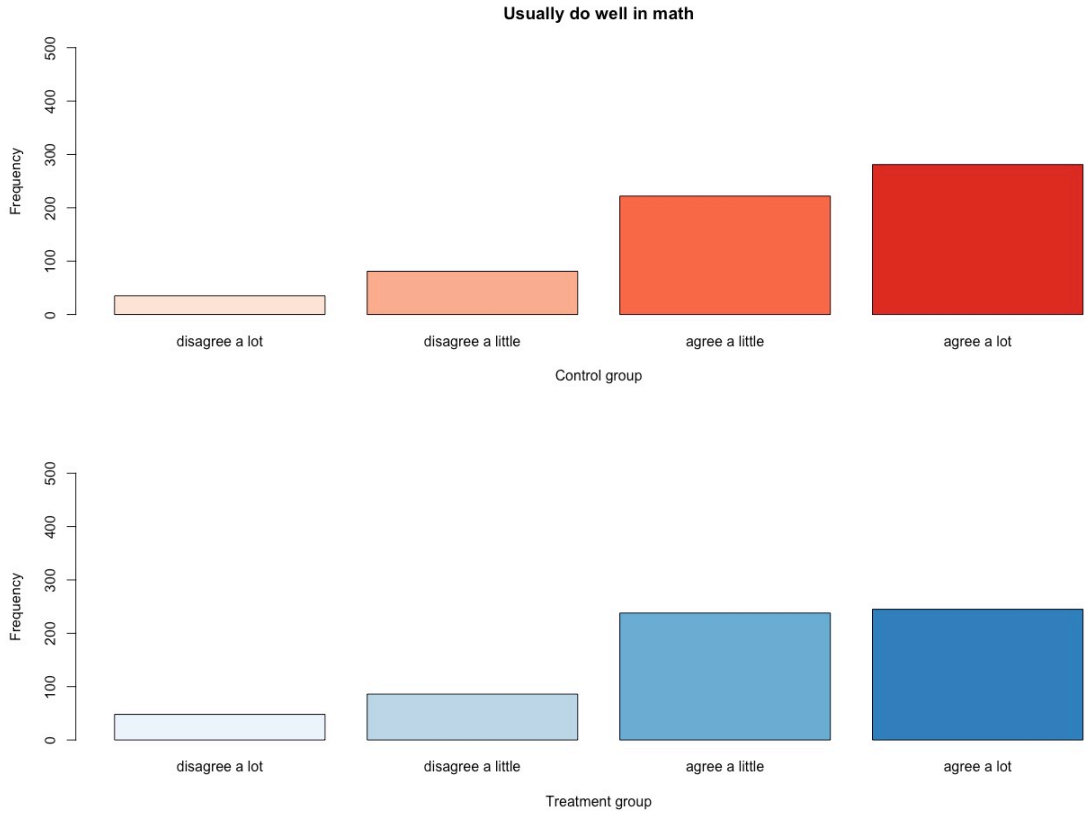

Figure 9. Example "student feels confident in mathematics" response distributions

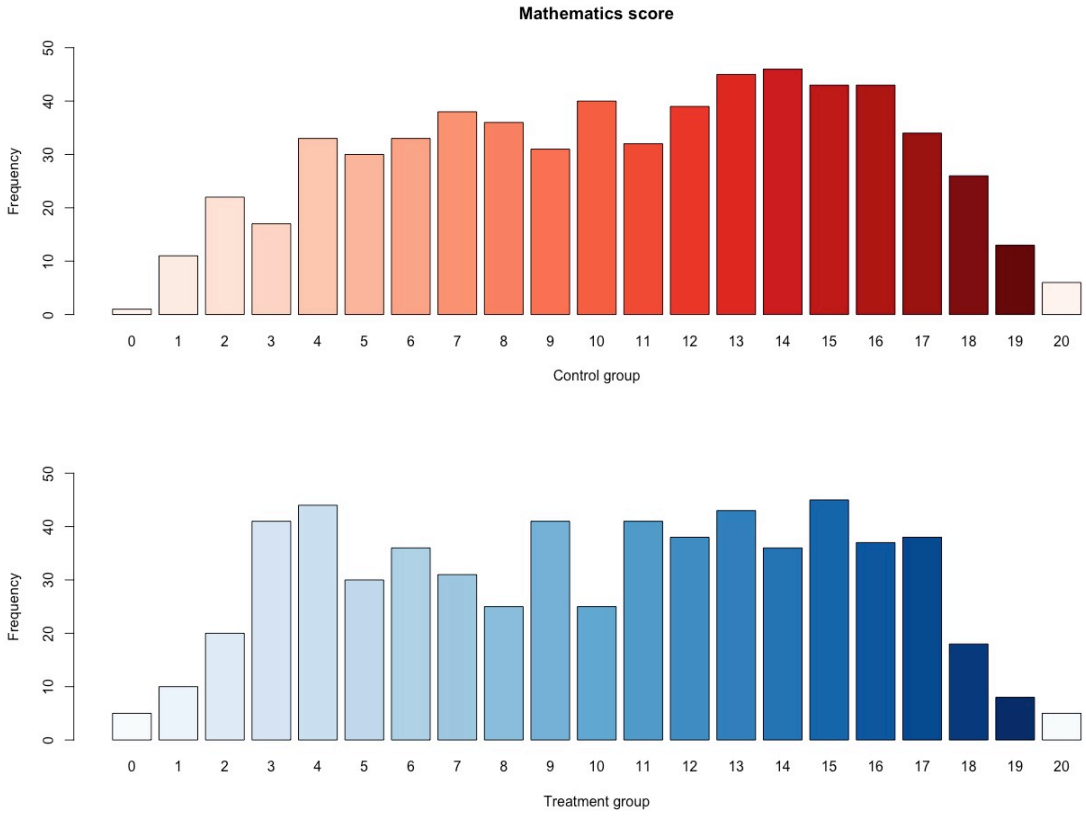

Figure 10. Total mathematics score distributions

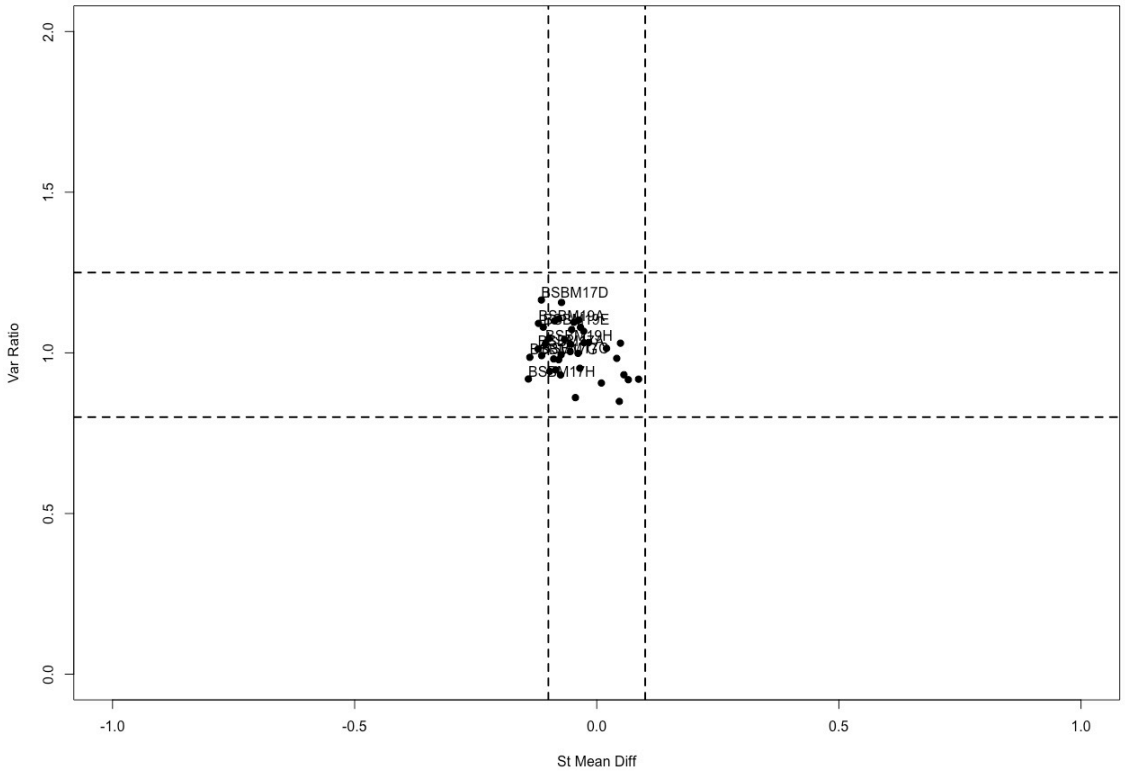

Figure 11. Initial covariate balance

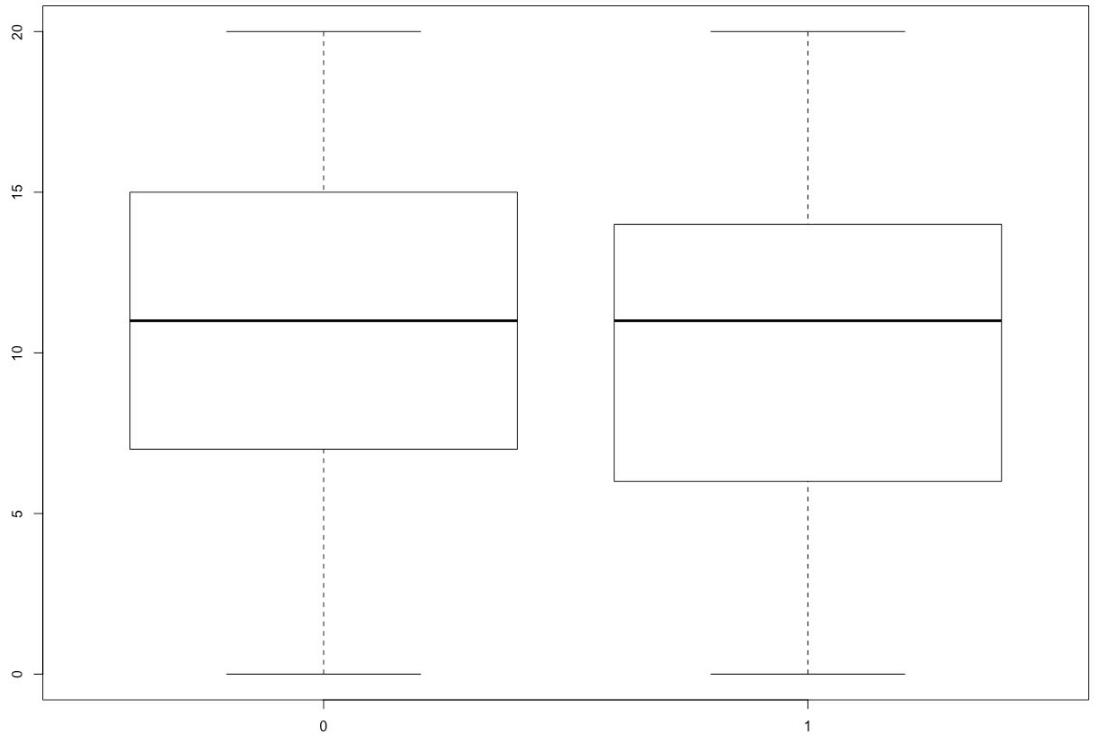

Figure 12. Initial difference between each item position group

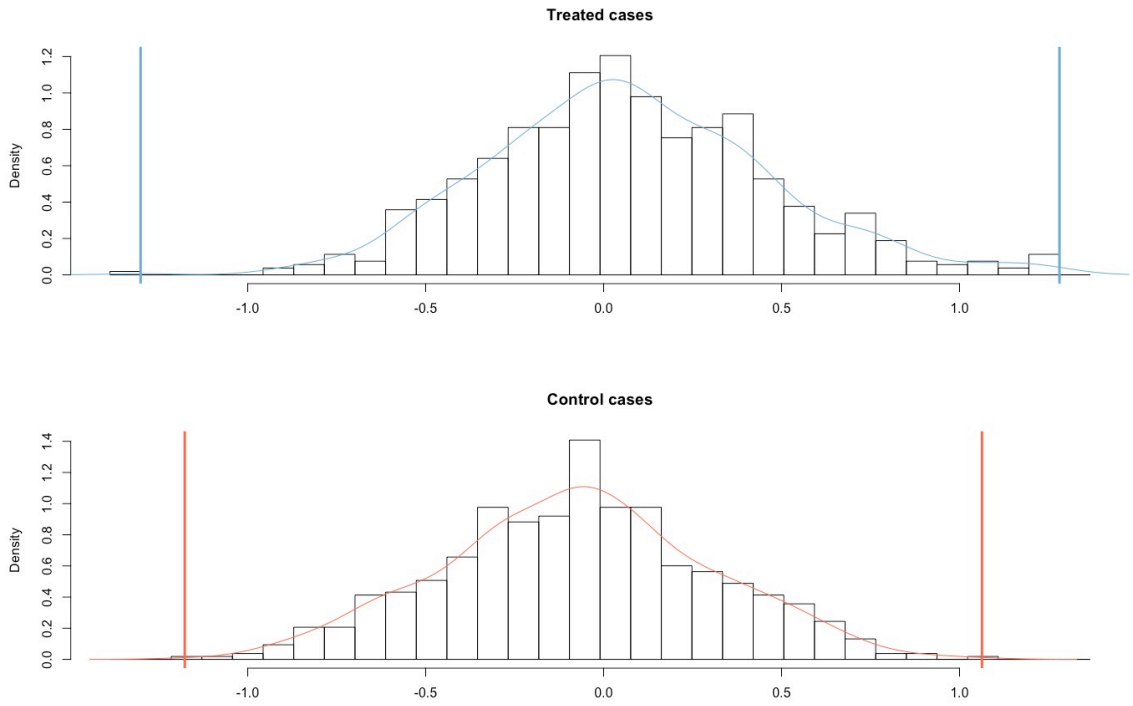

Figure 13. Initial overlap assessment

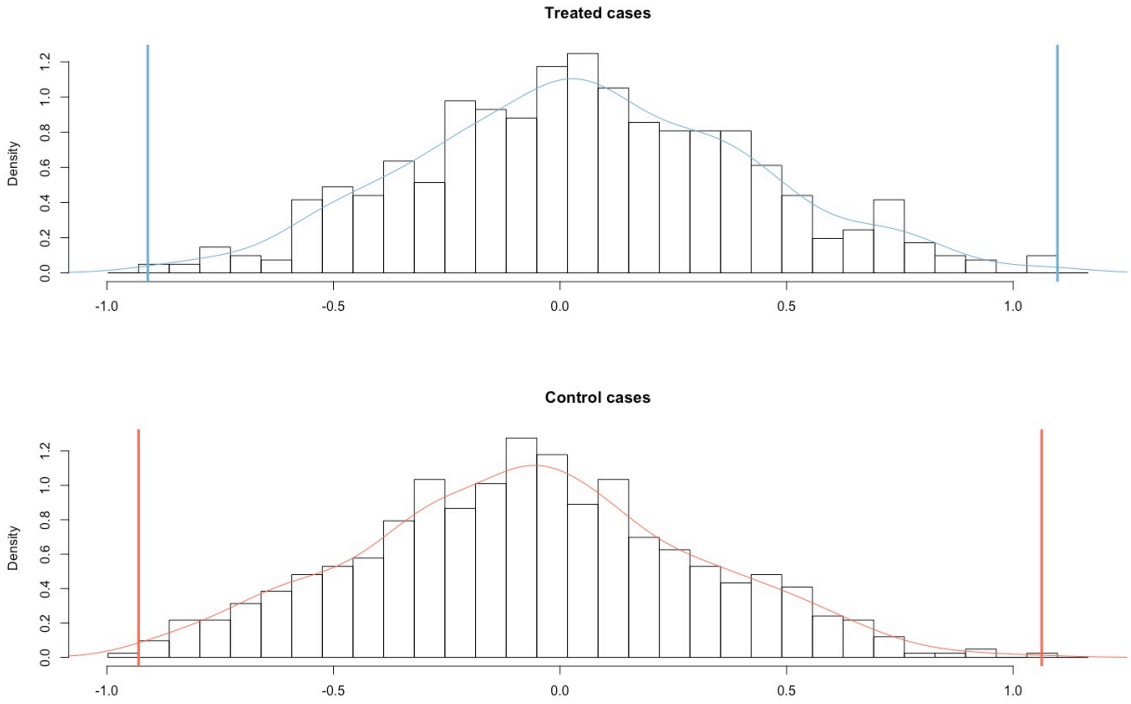

Figure 14. Overlap assessment after dropping non-overlapping cases

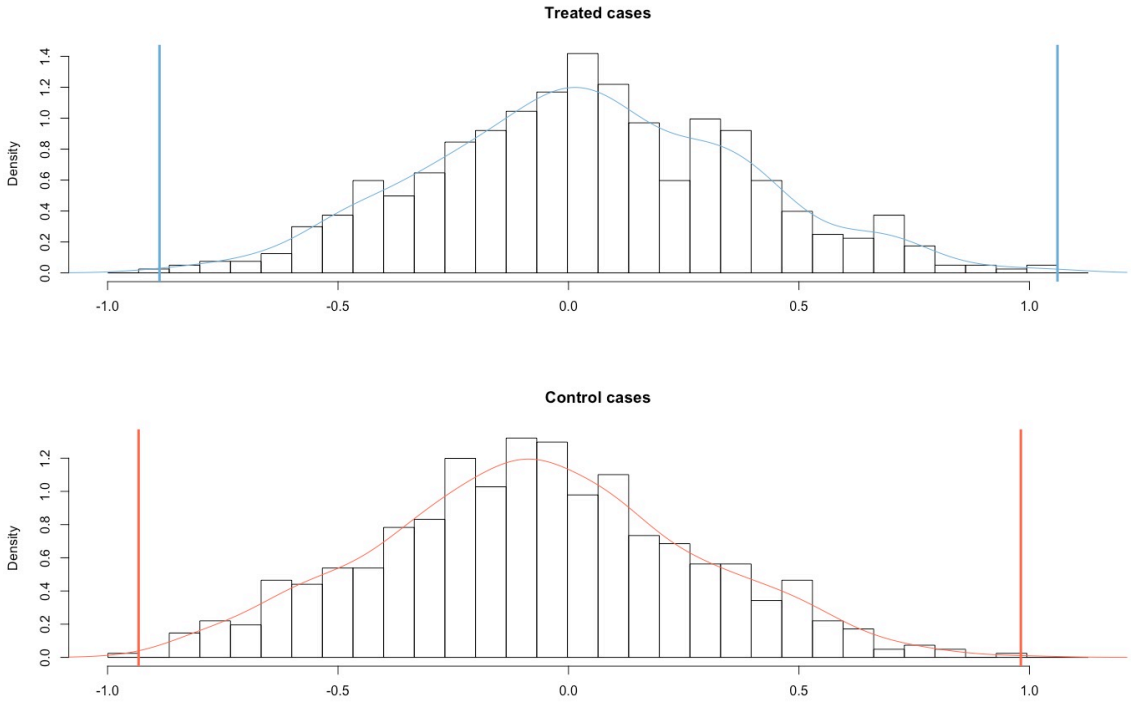

Figure 15. Final overlap assessment

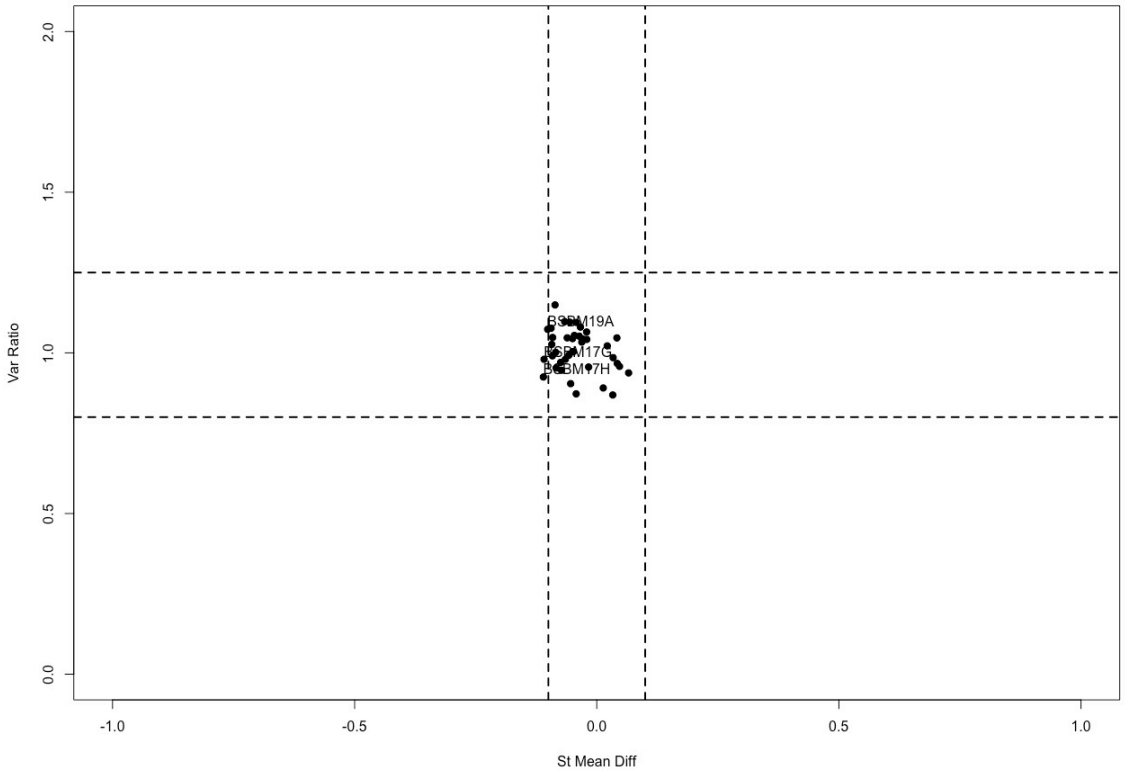

Figure 16. Final covariate balance

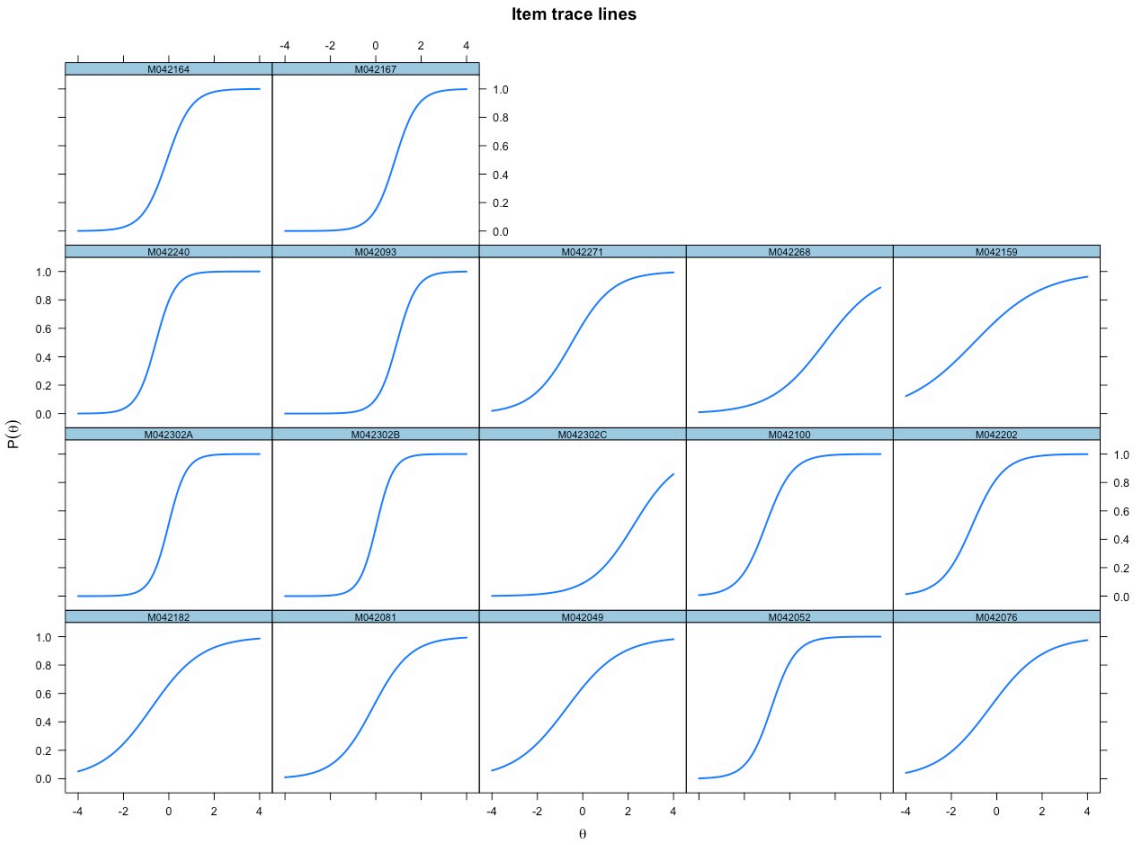

Figure 17. Item characteristic curves for mathematics items

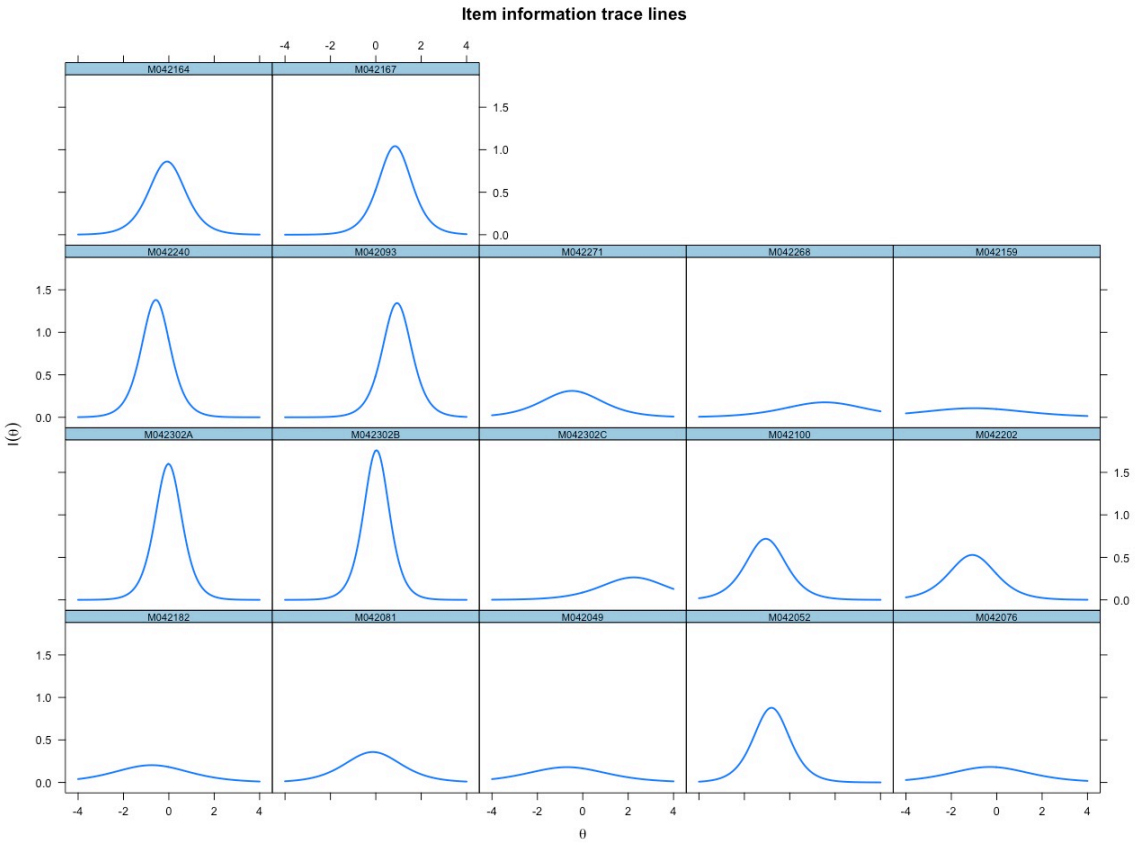

Figure 18. Item information function for mathematics items

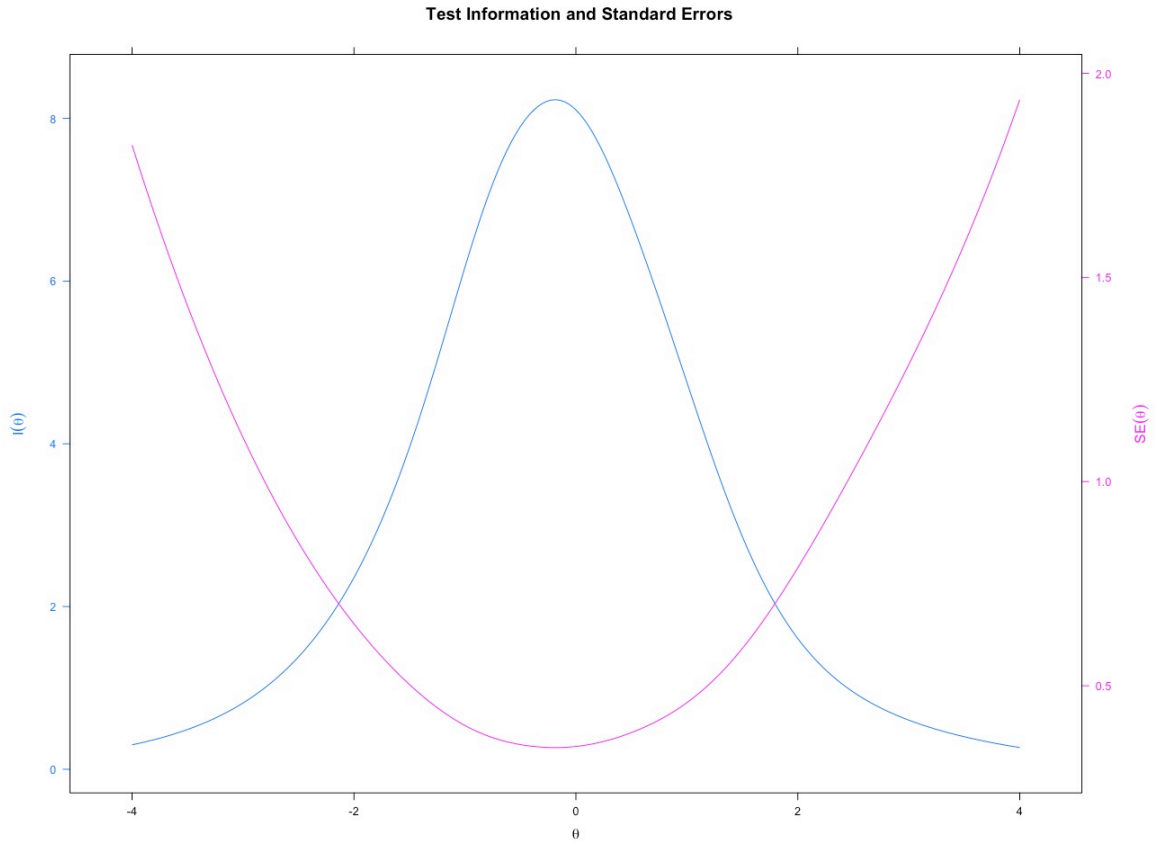

Figure 19. Test information and standard error for the beginning group

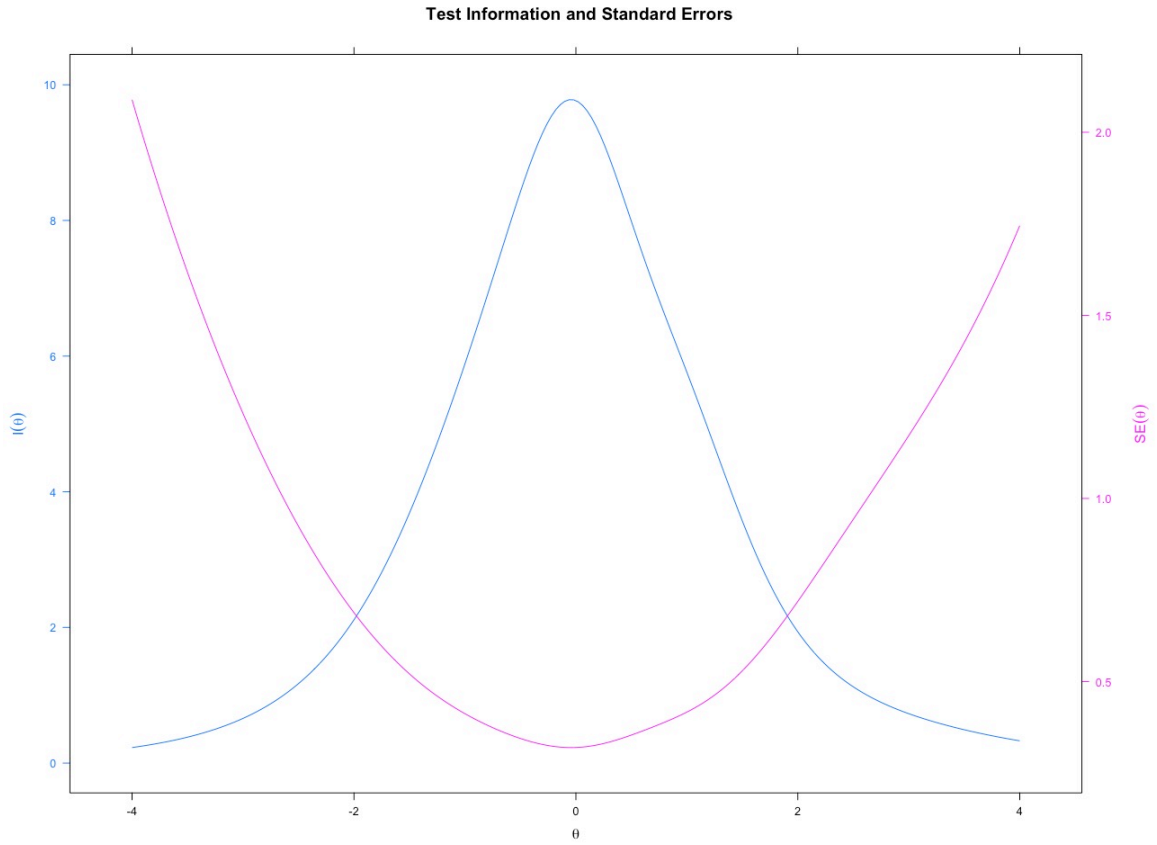

Figure 20. Test information and standard error for the end group

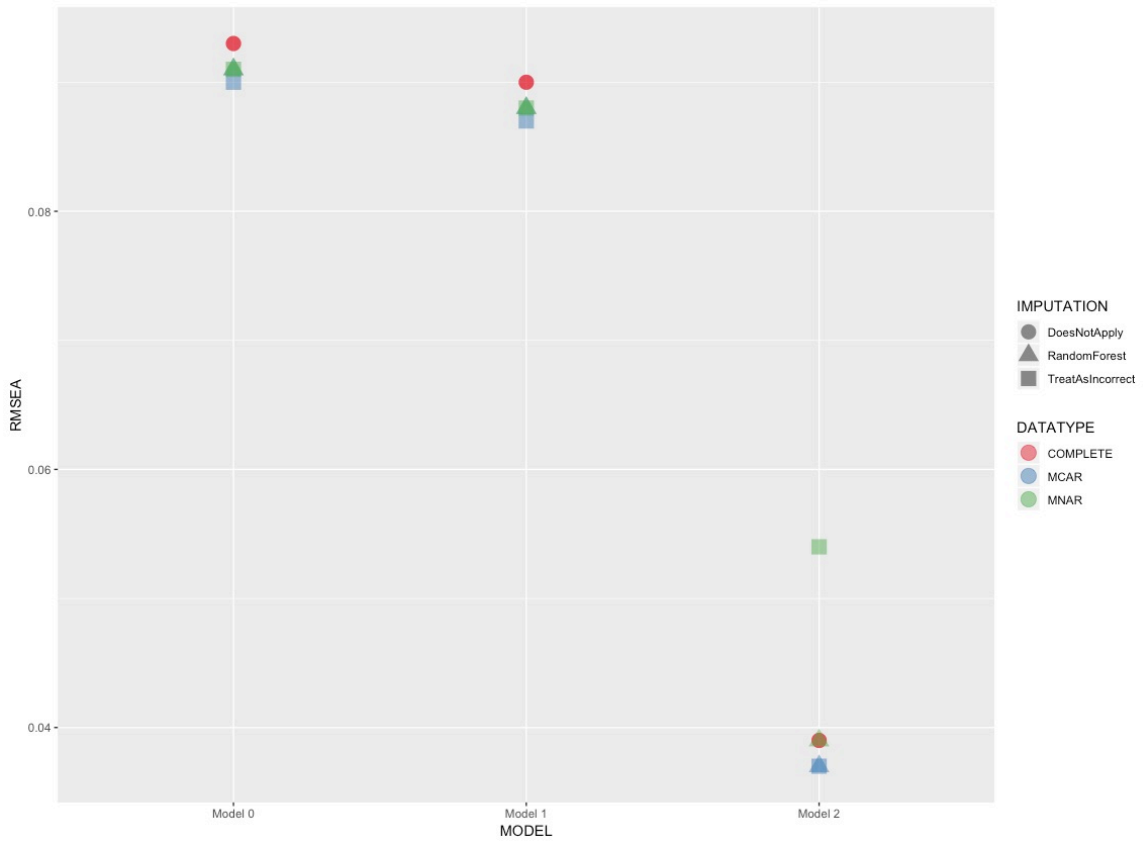

Figure 21. Overview of model fit information

Appendix A – R code

```
#### libraries & working directory
library(magrittr)
library(dplyr)
library(psych)
library(RColorBrewer)
library(optmatch)
library(mirt)
library(lavaan)
library(tidyverse)
library(mice)
library(randomForest)
library(VIM)
library(scales)
library(lavaan)
library(grid)
library(ggplot2)
library(reshape2)

setwd("/Users/nayeon/Desktop/defense/data")

#### data
score <- read.csv(file = "UsaSCR.csv", header = TRUE, sep = ",") %>%
select(IDCNTRY:M042167)

var <- read.csv(file = "BSGUSAM6.csv", header = TRUE, sep = ",") %>%
select(IDCNTRY:IDSTUD, ITSEX, BSBG03, BSBG12, BSBG15A:BSBG15G,
BSBM17A:BSBM17I,BSBM19A:BSBM20I,BSBGHER, BSDGHER, BSDG06S)

noNA <- c("IDCNTRY", "IDBOOK", "IDSCHOOL", "IDCLASS", "IDSTUD",
"BSBGHER")
var[ ,!(colnames(var) %in% noNA)][var[ ,!(colnames(var) %in% noNA)] == 9] <- NA
```

```
# consolidate & filter
data <- full_join(score, var, by = c("IDCNTRY", "IDBOOK", "IDSCHOOL",
"IDCLASS", "IDSTUD")) %>%
filter(IDBOOK %in% c(1, 14))

# drop NAs
sapply(data, function(x) sum(is.na(x)))
data <- na.omit(data)

# total score
data %<>% mutate(Tscore =
M042182+M042081+M042049+M042052+M042076+M042302A+M042302B+M04230
2C+M042100+M042202+M042240+M042093+M042271+M042268+M042159+M0421
64+M042167)

# partial credit
partial <- c("M042302A", "M042302B", "M042302C")
data[,partial][data[,partial] == 1] <- 0
data[,partial][data[,partial] == 2] <- 1

# gender
data$ITSEX[data$ITSEX == 2] <- 0

# reverse code
reverse <- c("BSBG03", "BSBG15A", "BSBG15B", "BSBG15C", "BSBG15D",
"BSBG15E", "BSBG15F", "BSBG15G","BSBM17A", "BSBM17D", "BSBM17E",
"BSBM17F", "BSBM17G", "BSBM17H", "BSBM17I", "BSBM19A", "BSBM19D",
"BSBM19F", "BSBM19G", "BSBM20A", "BSBM20B", "BSBM20C", "BSBM20D",
"BSBM20E", "BSBM20F", "BSBM20G", "BSBM20H", "BSBM20I")
data[,reverse] <- 5-data[,reverse]
data$BSDGHER <- 4-data$BSDGHER
```

```
#### study 1 ####
#### descriptive
dataProp <- data
dataProp$IPGROUP <- sapply(dataProp$IDBOOK, function (x) ifelse(x %in% 1, 0, 1))

trt <- filter(dataProp, IPGROUP == 1)
ctl <- filter(dataProp, IPGROUP == 0)
cov <- names(dataProp)[-c(1:22, 62:64)]

table(data$IDBOOK)
table(data$ITSEX)

barplot(xlab="Item position", ylab="Frequency", ylim=c(0, 1000),
      table(data$IDBOOK), main="Item position group",
      names.arg= c("At the beginning","At the end"), col= c("#FB6A4A","#6BAED6"))

par(mfrow=c(2,1))
barplot(xlab="Control group", ylab="Frequency", ylim=c(0, 500),
      table(ctl$BSBG03), main="Test language at home",
      names.arg= c("never","sometimes", "almost always", "always"),
      col= brewer.pal(5,"Reds"))
barplot(xlab="Treatment group", ylab="Frequency", ylim=c(0, 500),
      table(trt$BSBG03),
      names.arg= c("never","sometimes", "almost always", "always"),
      col= brewer.pal(5,"Blues"))

par(mfrow=c(2,1))
barplot(xlab="Control group", ylab="Frequency", ylim=c(0, 500),
      table(ctl$BSBG12), main="Breakfast at school",
      names.arg= c("never","sometimes", "most days", "every day"),
      col= brewer.pal(5,"Reds"))
barplot(xlab="Treatment group", ylab="Frequency", ylim=c(0, 500),
```

```
    table(trt$BSBG12),
    names.arg= c("never","sometimes", "most days", "every day"),
    col= brewer.pal(5,"Blues"))

par(mfrow=c(2,1))
barplot(xlab="Control group", ylab="Frequency", ylim=c(0, 500),
    table(ctl$BSBM17A), main="Enjoy learning mathematics",
    names.arg= c("disagree a lot","disagree a little", "agree a little", "agree a lot"),
    col= brewer.pal(5,"Reds"))
barplot(xlab="Treatment group", ylab="Frequency", ylim=c(0, 500),
    table(trt$BSBM17A),
    names.arg= c("disagree a lot","disagree a little", "agree a little", "agree a lot"),
    col= brewer.pal(5,"Blues"))

par(mfrow=c(2,1))
barplot(xlab="Control group", ylab="Frequency", ylim=c(0, 500),
    table(ctl$BSBM19A), main="Usually do well in math",
    names.arg= c("disagree a lot","disagree a little", "agree a little", "agree a lot"),
    col= brewer.pal(5,"Reds"))
barplot(xlab="Treatment group", ylab="Frequency", ylim=c(0, 500),
    table(trt$BSBM19A),
    names.arg= c("disagree a lot","disagree a little", "agree a little", "agree a lot"),
    col= brewer.pal(5,"Blues"))

par(mfrow=c(2,1))
barplot(xlab="Control group", ylab="Frequency", ylim=c(0, 50),
    table(ctl$Tscore), main="Mathematics score",
    col= palR(20))
barplot(xlab="Treatment group", ylab="Frequency", ylim=c(0, 50),
    table(trt$Tscore),
    col= palB(20))
par(mfrow=c(1,1))
```

```
#### propensity score matching
# initial covariate balance
source(file = "10 Functions_2.R")
smds_pfe <- smd_vr_DF(covars = cov, dat = dataProp, trt = dataProp$IPGROUP, plot =
TRUE)

# simple difference in means
pfe <- lm(Tscore ~ IPGROUP, data = dataProp)
summary(pfe)
boxplot(Tscore ~ IPGROUP, data = dataProp)

# propensity score matching
glm1 <- glm(IPGROUP ~
ITSEX+BSBG03+BSBG12+BSBG15A+BSBG15B+BSBG15C+BSBG15D+BSBG15E
+BSBG15F+BSBG15G+BSBM17A+BSBM17B+BSBM17C+BSBM17D+BSBM17E+B
SBM17F+BSBM17G+BSBM17H+BSBM17I+BSBM19A+BSBM19B+BSBM19C+BS
BM19D+BSBM19E+BSBM19F+BSBM19G+BSBM19H+BSBM19I+BSBM20A+BSB
M20B+BSBM20C+BSBM20D+BSBM20E+BSBM20F+BSBM20G+BSBM20H+BSBM
20I+BSBGHER+BSDGHER+BSDG06S, data = dataProp, family = "binomial")
ps1 <- predict(glm1, type = "response")
lin_ps1 <- log(ps1/(1 - ps1))

ovlp(trt = dataProp$IPGROUP, lps = lin_ps1)
ovlp_ind1 <- ovlp_ind(trt = dataProp$IPGROUP, lps = lin_ps1, caliper = 0.1)
sum(ovlp_ind1)
table(ovlp_ind1, dataProp$IPGROUP)
ovlp(trt = dataProp$IPGROUP[ovlp_ind1], lps = lin_ps1[ovlp_ind1])
dataO <- dataProp[ovlp_ind1,]

# on overlapping data
glm2 <- glm(IPGROUP ~
ITSEX+BSBG03+BSBG12+BSBG15A+BSBG15B+BSBG15C+BSBG15D+BSBG15E
```

```
+BSBG15F+BSBG15G+BSBM17A+BSBM17B+BSBM17C+BSBM17D+BSBM17E+B
SBM17F+BSBM17G+BSBM17H+BSBM17I+BSBM19A+BSBM19B+BSBM19C+BS
BM19D+BSBM19E+BSBM19F+BSBM19G+BSBM19H+BSBM19I+BSBM20A+BSB
M20B+BSBM20C+BSBM20D+BSBM20E+BSBM20F+BSBM20G+BSBM20H+BSBM
20I+BSBGHER+BSDGHER+BSDG06S, data = dataO, family = "binomial")
ps2 <- predict(glm2, type = "response")
lin_ps2 <- log(ps2/(1 - ps2))

ovlp(trt = dataO$IPGROUP, lps = lin_ps2)
pm1 <- pairmatch(lin_ps2, z = dataO$IPGROUP, data = dataO)
table(pm1, dataO$IPGROUP, useNA = "always")
matched_dat <- dataO[matched(pm1), ]
dim(matched_dat)

# final covariate balance
smds_pm <- smd_vr_DF(covars = cov, dat = matched_dat,
                     trt = matched_dat$IPGROUP,
                     wts = rep(1, nrow(matched_dat)), plot = TRUE)

ATT_mtch <- lm(Tscore ~ IPGROUP, data = matched_dat)
summary(ATT_mtch)

#### study 2 ####
#### descriptive
dataSem <- matched_dat
dataSem$IP <- sapply(dataSem$IDBOOK, function (x) ifelse(x %in% 1, 0, 1))

data2pl <- select(dataSem, IP, M042182:M042167)
dataGrm <- select(dataSem, IP, BSBM17A:BSBM17I)

# item calibration
resp2pl <- select(data2pl, -IP) # whole
```

```
mod2pl <- 'math = 1-17'
calib2pl <- mirt(data = resp2pl, model = mod2pl, itemtype = "2PL", SE = TRUE, verbose
= FALSE)
coef2pl <- coef(calib2pl, IRTpars = TRUE, simplify = TRUE)
items2pl <- as.data.frame(coef2pl$items)

plot(calib2pl, type='trace', theta_lim = c(-4,4), lwd=2)
plot(calib2pl, type='infotrace', theta_lim = c(-4,4), lwd=2)

# 1
resp2pl.1 <- filter(data2pl, IP == 0) %>%
  select(-IP)
mod2pl.1 <- 'math = 1-17'
calib2pl.1 <- mirt(data = resp2pl.1, model = mod2pl.1, itemtype = "2PL", SE = TRUE,
verbose = FALSE)
coef2pl.1 <- coef(calib2pl.1, IRTpars = TRUE, simplify = TRUE)
items2pl.1 <- as.data.frame(coef2pl.1$items)

# 14
resp2pl.14 <- filter(data2pl, IP == 1) %>%
  select(-IP)
mod2pl.14 <- 'math = 1-17'
calib2pl.14 <- mirt(data = resp2pl.14, model = mod2pl.14, itemtype = "2PL", SE =
TRUE, verbose = FALSE)
coef2pl.14 <- coef(calib2pl.14, IRTpars = TRUE, simplify = TRUE)
items2pl.14 <- as.data.frame(coef2pl.14$items)

plot(calib2pl.1, type='infoSE', theta_lim = c(-4,4), lwd=2)
plot(calib2pl.14, type='infoSE', theta_lim = c(-4,4), lwd=2)

#### SEM model fitting w/ complete data
dataSem %<>% select(IP, M042182:M042167, BSBM17A:BSBM17I)
```

```
# null model
model0 = '
# loadings
theta =~ l1*M042182 + l2*M042081 + l3*M042049 + l4*M042052 + l5*M042076 +
l6*M042302A + l7*M042302B + l8*M042302C + l9*M042100 + l10*M042202 +
l11*M042240 + l12*M042093 + l13*M042271 + l14*M042268 + l15*M042159 +
l16*M042164 + l17*M042167

motiv =~ l18*BSBM17A + l19*BSBM17B + l20*BSBM17C + l21*BSBM17D +
l22*BSBM17E + l23*BSBM17F + l24*BSBM17G + l25*BSBM17H + l26*BSBM17I

# thresholds
M042182 | th1*t1
M042081 | th2*t1
M042049 | th3*t1
M042052 | th4*t1
M042076 | th5*t1
M042302A | th6*t1
M042302B | th7*t1
M042302C | th8*t1
M042100 | th9*t1
M042202 | th10*t1
M042240 | th11*t1
M042093 | th12*t1
M042271 | th13*t1
M042268 | th14*t1
M042159 | th15*t1
M042164 | th16*t1
M042167 | th17*t1

BSBM17A | th180*t1 + th181*t2 + th182*t3
BSBM17B | th190*t1 + th191*t2 + th192*t3
```

```
BSBM17C | th200*t1 + th201*t2 + th202*t3
BSBM17D | th210*t1 + th211*t2 + th212*t3
BSBM17E | th220*t1 + th221*t2 + th222*t3
BSBM17F | th230*t1 + th231*t2 + th232*t3
BSBM17G | th240*t1 + th241*t2 + th242*t3
BSBM17H | th250*t1 + th251*t2 + th252*t3
BSBM17I | th260*t1 + th261*t2 + th262*t3

# factor variances
motiv ~~ motiv
theta ~~ theta
'
fit0 <- lavaan(model0, data = dataSem, std.lv=T, parameterization='theta')
summary(fit0)

# IP model
model1 = '
# loadings
theta =~ l1*M042182 + l2*M042081 + l3*M042049 + l4*M042052 + l5*M042076 +
l6*M042302A + l7*M042302B + l8*M042302C + l9*M042100 + l10*M042202 +
l11*M042240 + l12*M042093 + l13*M042271 + l14*M042268 + l15*M042159 +
l16*M042164 + l17*M042167

motiv =~ l18*BSBM17A + l19*BSBM17B + l20*BSBM17C + l21*BSBM17D +
l22*BSBM17E + l23*BSBM17F + l24*BSBM17G + l25*BSBM17H + l26*BSBM17I

theta ~ IP

# thresholds
M042182 | th1*t1
M042081 | th2*t1
M042049 | th3*t1
```

```
M042052 | th4*t1
M042076 | th5*t1
M042302A | th6*t1
M042302B | th7*t1
M042302C | th8*t1
M042100 | th9*t1
M042202 | th10*t1
M042240 | th11*t1
M042093 | th12*t1
M042271 | th13*t1
M042268 | th14*t1
M042159 | th15*t1
M042164 | th16*t1
M042167 | th17*t1

BSBM17A | th180*t1 + th181*t2 + th182*t3
BSBM17B | th190*t1 + th191*t2 + th192*t3
BSBM17C | th200*t1 + th201*t2 + th202*t3
BSBM17D | th210*t1 + th211*t2 + th212*t3
BSBM17E | th220*t1 + th221*t2 + th222*t3
BSBM17F | th230*t1 + th231*t2 + th232*t3
BSBM17G | th240*t1 + th241*t2 + th242*t3
BSBM17H | th250*t1 + th251*t2 + th252*t3
BSBM17I | th260*t1 + th261*t2 + th262*t3

# factor variances
motiv ~~ motiv
theta ~~ theta
'
fit1 <- lavaan(model1, data = dataSem, std.lv=T, parameterization='theta')
summary(fit1)
```

```
# full model
model2 = '
# loadings
theta =~ l1*M042182 + l2*M042081 + l3*M042049 + l4*M042052 + l5*M042076 +
l6*M042302A + l7*M042302B + l8*M042302C + l9*M042100 + l10*M042202 +
l11*M042240 + l12*M042093 + l13*M042271 + l14*M042268 + l15*M042159 +
l16*M042164 + l17*M042167

motiv =~ l18*BSBM17A + l19*BSBM17B + l20*BSBM17C + l21*BSBM17D +
l22*BSBM17E + l23*BSBM17F + l24*BSBM17G + l25*BSBM17H + l26*BSBM17I

theta ~ motiv
theta ~ IP

# thresholds
M042182 | th1*t1
M042081 | th2*t1
M042049 | th3*t1
M042052 | th4*t1
M042076 | th5*t1
M042302A | th6*t1
M042302B | th7*t1
M042302C | th8*t1
M042100 | th9*t1
M042202 | th10*t1
M042240 | th11*t1
M042093 | th12*t1
M042271 | th13*t1
M042268 | th14*t1
M042159 | th15*t1
M042164 | th16*t1
M042167 | th17*t1
```

```
BSBM17A | th180*t1 + th181*t2 + th182*t3
BSBM17B | th190*t1 + th191*t2 + th192*t3
BSBM17C | th200*t1 + th201*t2 + th202*t3
BSBM17D | th210*t1 + th211*t2 + th212*t3
BSBM17E | th220*t1 + th221*t2 + th222*t3
BSBM17F | th230*t1 + th231*t2 + th232*t3
BSBM17G | th240*t1 + th241*t2 + th242*t3
BSBM17H | th250*t1 + th251*t2 + th252*t3
BSBM17I | th260*t1 + th261*t2 + th262*t3

# factor variances
motiv ~~ motiv
theta ~~ theta
'
fit2 <- lavaan(model2, data = dataSem, std.lv=T, parameterization='theta')
summary(fit2)

#### SEM model fitting w/ missing data
dataMissing <- matched_dat
dataMissing$IP <- sapply(dataMissing$IDBOOK, function (x) ifelse(x %in% 1, 0, 1))
head(dataMissing)

missingIP <- select(dataMissing, IP)
missing2pl <- select(dataMissing, M042182:M042167)
missingGrm <- select(dataMissing, BSBM17A:BSBM17I)

# MCAR
mcar <- ampute(missing2pl, prop = 0.30, patterns = NULL, freq = NULL,
            mech = "MCAR", weights = NULL, std = TRUE, cont = TRUE,
            type = NULL, odds = NULL, bycases = TRUE, run = TRUE)
dataMcar <- mcar$amp
```

```
# MNAR
dataMissing.1 <- filter(dataMissing, IP == 0) %>%
  select(IP, M042182:M042167, BSBM17A:BSBM17I)

dataMissing.14 <- filter(dataMissing, IP == 1)
missingIP.14 <- select(dataMissing.14, IP)
missing2pl.14 <- select(dataMissing.14, M042182:M042167)
missingGrm.14 <- select(dataMissing.14, BSBM17A:BSBM17I)

mnar <- ampute(missing2pl.14, prop = 0.30, patterns = NULL, freq = NULL,
               mech = "MNAR", weights = NULL, std = TRUE, cont = TRUE,
               type = NULL, odds = NULL, bycases = TRUE, run = TRUE)

patterns <- mnar$patterns
diag(patterns) <- 1
patterns[1:2, 17] <- 0
patterns[3:7, 13:17] <- 0
patterns[8:12, 9:17] <- 0
patterns[13:17, 5:17] <- 0

mnar <- ampute(missing2pl.14, prop = 0.30, patterns = patterns, freq = NULL,
               mech = "MNAR", weights = NULL, std = TRUE, cont = TRUE,
               type = NULL, odds = NULL, bycases = TRUE, run = TRUE)
dataMnar <- mnar$amp
md.pattern(dataMnar)

# imputation
# treated as incorrect
mcarIncor <- dataMcar # mcar
mcarIncor[is.na(mcarIncor)] <- 0
mcarIncor <- cbind(missingIP, mcarIncor, missingGrm)
```

```
mnarIncor <- dataMnar # mnar
mnarIncor[is.na(mnarIncor)] <- 0
mnarIncor <- cbind(missingIP.14, mnarIncor, missingGrm.14)
mnarIncor <- rbind(dataMissing.1, mnarIncor)

# random forest
mcarRF <- dataMcar # mcar
mcarRF <- mice(mcarRF, meth = "rf", ntree = 3)
mcarRF <- complete(mcarRF, 1)
mcarRF <- cbind(missingIP, mcarRF, missingGrm)

mnarRF <- dataMnar # mnar
mnarRF <- mice(mnarRF, meth = "rf", ntree = 3)
mnarRF <- complete(mnarRF, 1)
mnarRF <- cbind(missingIP.14, mnarRF, missingGrm.14)
mnarRF <- rbind(dataMissing.1, mnarRF)

# model fitting
# mcar
mcarIncor.fit0 <- lavaan(model0, data = mcarIncor, std.lv=T)
mcarIncor.fit1 <- lavaan(model1, data = mcarIncor, std.lv=T)
mcarIncor.fit2 <- lavaan(model2, data = mcarIncor, std.lv=T)

mcarRF.fit0 <- lavaan(model0, data = mcarRF, std.lv=T)
mcarRF.fit1 <- lavaan(model1, data = mcarRF, std.lv=T)
mcarRF.fit2 <- lavaan(model2, data = mcarRF, std.lv=T)

# mnar
mnarIncor.fit0 <- lavaan(model0, data = mnarIncor, std.lv=T)
mnarIncor.fit1 <- lavaan(model1, data = mnarIncor, std.lv=T)
mnarIncor.fit2 <- lavaan(model2, data = mnarIncor, std.lv=T)
```

```
mnarRF.fit0 <- lavaan(model0, data = mnarRF, std.lv=T)
mnarRF.fit1 <- lavaan(model1, data = mnarRF, std.lv=T)
mnarRF.fit2 <- lavaan(model2, data = mnarRF, std.lv=T)

# graph
graph <- read.csv("RMSEAresult.csv")
p21 <- ggplot(filter(graph, DATATYPE %in% c("COMPLETE", "MCAR", "MNAR")),
            aes(x = MODEL,
                y = RMSEA,
                color = DATATYPE))+
  geom_point(aes(color = DATATYPE, shape = IMPUTATION),size = 5, alpha = 0.5)
p21 + scale_color_brewer(palette="Set1")
```