AN EXPLORATION OF COMPARABILITY ISSUES IN EDUCATIONAL RESEARCH: SCALE LINKING, EQUATING, AND PROPENSITY SCORE WEIGHTING

by

Tong Wu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Educational Research, Measurement, and Evaluation

Charlotte

2023

Approved by:

_____
Dr. Stella Kim

_____
Dr. Carl Westine

_____
Dr. Claudia Flowers

_____
Dr. Anne Cash

ABSTRACT

TONG WU. An Exploration of Comparability Issues in Educational Research: Scale Linking, Equating, and Propensity Score Weighting. (Under the co-direction of Dr. STELLA KIM and Dr. CARL WESTINE)

This three-article dissertation aims to address three methodological challenges to ensure comparability in educational research, including scale linking, test equating, and propensity score (PS) weighting. The first study intends to improve test scale comparability by evaluating the effect of six missing data handling approaches, including listwise deletion (LWD), treating missing data as incorrect responses (IN), corrected item mean imputation (CM), imputing with a response function (RF), multiple imputation (MI), and full information likelihood information (FIML), on item response theory (IRT) scale linking accuracy when missing data occur within common items. The relative performance of these six missing data treatment methods under two missing mechanisms is explored with simulated data. Results show that RF, MI, and FIML produce fewer errors for conducting scale linking, whereas LWD is associated with the most errors regardless of testing conditions. The second study aims to ensure test score comparability by proposing a new equating method to account for rater errors in rater-mediated assessments. Specifically, the performance of using an IRT observed-score equating method with a hierarchical rater model (HRM) is investigated under various conditions. The newly proposed equating method leads to comparable bias, SE, and RMSE to that of a traditional IRT observed-score equating method with the use of generalized partial credit model (GPCM) as normal raters scoring the new test forms. However, when aberrant raters are involved in the scoring process, the HRM IRT observed-score equating method generally produces more accurate results in bias and RMSE, though generates comparable SEs to the traditional method. The third study

examines the performance of six covariate balance diagnostics when using PS weighting method with multilevel data. Specifically, a set of simulated conditions is used to examine the ability of within-cluster and pooled absolute standardized bias (ASB), variance ratio (VR), and percent bias reduction (PBR) methods to identify a correct PS model. In addition, the association between the balance statistics and the bias in treatment effect is explored. Within-cluster ASB and PBR are observed to be associated with the most accurate results in the choice of PS model as compared to other diagnostics. Pooled ASB is found to have the highest association with the treatment effect bias. By advancing the methodology for addressing comparability issues, the dissertation intends to enhance the validity and improve the quality of educational research.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Stella Kim, my dissertation co-chair, for providing invaluable advice throughout my PhD journey. Her expertise and constructive feedback challenged me to think critically and significantly enhanced the quality of my work. I am deeply grateful for her “counseling services”, without which this dissertation and the publication of the first article in Educational and Psychological Measurement would not have been possible.

I am immensely grateful to Dr. Carl Westline, my advisor and dissertation co-chair, for his guidance at UNC Charlotte. Working with him on his grant not only provided financial support but also the opportunity to present my research at several conferences. His mentorship helped me generate ideas for the third paper of this dissertation, and I am thankful for the invaluable lessons he taught me.

My appreciation also goes to the other committee members, Dr. Claudia Flowers and Dr. Anne Cash, for their contributions to my work. Their suggestions and comments provided me with a fresh perspective and helped me think deeply about the implications of my studies, which enhanced the quality of my work.

I would also like to thank the Educational Leadership department faculty, including Dr. Dika, Dr. Lim, Dr. Wang, Dr. Newton, Dr. Martin, and Dr. Sadaf, for creating a supportive and safe environment for international students like me. In addition, the friendships I established while studying at UNCC, particularly with Ting, Tuba, Zhi, Yi, Kristin, Yue, and Maorong, were crucial to my achievements.

I am also thankful for the external support I received, including funding from the UNC System Office and the opportunity to intern at the Center for Assessment. I express my

appreciation to Dr. Michelle Boyer for inspiring the second paper of this dissertation during our work together in the summer of 2020. I am also grateful to Dr. Brooke Houck and Dr. Jill Bryant for providing me with a part-time job as a psychometrician at the National Board of Examiners in Optometry, which allowed me to gain a deeper understanding of the roles and responsibilities of a psychometrician.

I express my gratitude to my current supervisors at Riverside Insights, Dr. JP Kim, and Dr. Hyeonjoo Oh, and my colleagues, Dr. Mike Custer, Dr. Hsin-Ro Wei, and Huan Liu, for their support and encouragement, which have contributed to my growth as a qualified psychometrician.

Finally, I thank my family, including my parents, grandmas, uncles, aunts, and in-laws, for their understanding for the past few year years. I am grateful to my husband Fei for his patience, especially in the last year when he took care of the household. I am thankful for his presence in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASB	absolute standardized bias
ATE	average treatment effect
ATT	average treatment effect on treated participants
CINEG	common-item nonequivalent groups design
CM	corrected mean
CR	constructed response
DRIFT	differential rater functioning over time
FIML	full information maximum likelihood
GMFRM	generalized many-facet rater model
GPCM	generalized partial credit model
GRM	graded response model
HRM	hierarchical rater model
ICC	intraclass correlation coefficient
IN	incorrect
IPTW	inverse probability of treatment weighting
IRT	item response theory
LWD	listwise deletion
MAR	missing at random
MC	multiple-choice
MCAR	missing completely at random
MFRM	multifaceted Rasch model
MI	multiple imputation

ML	maximum likelihood
MNAR	missing not at random
NAEP	National Assessment of Educational Progress
NCES	National Center for Educational Statistics
ORMSE	overall/marginal Root Mean Square Error
OSE	overall/marginal standard error
PBR	percent bias reduction
PCM	partial credited model
PS	propensity score
RCT	randomized controlled trial
RF	response function
RMSE	Root Mean Square Error
SB	standardized bias
SE	standard error
VR	variance ratio

INTRODUCTION

The topic of comparability has received extensive attention in measurement and statistical research (e.g., Berger, 2006; Engelhard & Crocker, 1995; Green, 1995; Kolen, 1999). Depending on specific purposes and contexts, comparability can be interpreted differently. Therefore, researchers need to provide detailed and accurate information to create effective communication with stakeholders (Elliott, 2013). Notably, ensuring comparability is one of the key steps to maintain the validity of analyses or inferences under many circumstances (Berman et al., 2020; Evans & Lyons, 2017; Newgard et al., 2004).

In the field of measurement, achieving comparability means that "users could be assured that students with the same score are equally proficient with respect to the knowledge and skills a test was intended to measure" (Berman et al., 2020, p. 2). According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), maintaining the comparability of test scores "enables test users to make comparable inferences based on the scores for all test takers (p. 59). In other words, test scores should be interchangeable regardless of test time, location, form, mode, or other factors. Comparability is significant in ensuring test fairness, such that examinees who are assigned to a particular form are not inadvertently disadvantaged over the others who take different forms of a test. With such practices, stakeholders, such as school admission officers, evaluators, test takers, and parents can make accurate interpretations of test results through valid test scores comparisons. Beyond the strict definition of test score interchangeability, Evans and Lyons (2017) stressed the value of viewing comparability from a broader perspective. Specifically, comparable inferences need to be guaranteed for different school districts and across assessment systems to ensure state accountability when various local assessments are administered.

Securing comparability related to test scores and across districts and assessment systems can be achieved through various approaches. Statistical procedures such as linking and equating can be used to adjust for differences in scales, group abilities, and test scores (Kolen & Brennan, 2014). Judgement methods can also be employed, particularly for performance-based assessments, where expert judges ensure comparability of the judgements made by primary judges (Adams, 2007; Evans & Lyons, 2017).

Ensuring comparability is also crucial in intervention studies, in which the actual effect of the intervention is more likely to be obtained if comparable treatment and control groups are created. Randomized controlled trial (RCT) is an ideal research design because all subjects have an equal probability of being assigned to a group, which will produce two comparable groups. The comparability between groups indicates that covariates, both observed or unobserved, such as age, ethnicity, gender, etc., between treatment and control groups are similar. Thus, any difference between treated and untreated subjects can be directly attributed to the treatment effect on the outcome measure (Greenland et al., 1999).

However, in practice, randomization is often not practical or unethical, especially in education. For example, it is almost impossible for researchers to arbitrarily assign students into treated or untreated classes or schools. In non-randomized studies, the comparability of the characteristics between the two groups is unknown. Moreover, the probability of being assigned to a treatment group often relates to subject propensities, making it challenging to obtain comparable groups. If there is an imbalance in covariates between the groups, the treatment effect cannot be accurately estimated due to confounding issues (Greenland & Neutra, 1980). Thus, it is essential to use techniques such as propensity score matching, weighting,

subclassification, and covariate adjustments (Bai & Clark, 2018) to account for systematic differences between groups and ensure the validity of estimated treatment effects.

Though multiple approaches can be employed to achieve comparability of different types or within contexts, there are several factors that can pose a threat to it. This dissertation aims to address three methodological challenges associated with comparability issues in educational research. The dissertation is formatted with three studies, each of which tries to address one specific challenge based on simulated data. The challenges are presented below:

1. Which missing data handling method should be used to achieve accurate scale transformation for maintaining comparability among examinees taking different test forms?
2. How can test score comparability be maintained if responses on different test forms are graded by raters with varying levels of severity and variability?
3. Which covariate balance diagnosis is the most effective for selecting a more accurate propensity score model and constructing comparable groups when weighting is used with multilevel data?

The dissertation aims to provide guidance to educational researchers in identifying more accurate methods for achieving comparability in different educational contexts, including scale linking, test equating, and propensity score (PS) weighting. The dissertation examines three different instances related to comparability where researchers face decision-making in the design or analysis of their study. Each challenge is elaborated on below, and the respective challenge is briefly addressed in each of the three studies.

Chapter/ Study 1

To reduce the probability of cheating and enhance test security, large-scale testing programs often administer multiple forms of a test. However, within the item response theory (IRT), ensuring that test scores remain comparable across different forms of a test can be a difficult task. Common factors that can affect test score comparability may include differences in examinee abilities associated with different test forms and variations in the difficulty level of each test form. The first study in this dissertation aims to minimize the negative impact of ability differences between two forms to ensure test comparability.

The item response theory scales are characterized by the indeterminate property, which means that the location and spread of the scale are arbitrary. Conventionally, computer software utilizes a standard scale, specifically a normal distribution - $N(0, 1)$, for placing abilities estimated from different groups (Kim & Kolen, 2007; Mislevy & Bock, 1990). Under a random group design, examinees from two groups (old and new group) are considered to be equivalent with respect to ability, which is often achieved by incorporating a spiraling procedure during data collection. Parameters estimated from the two equivalent groups are automatically placed on a common standard scale, enabling the estimates to be compared directly.

However, it is often the case that test forms are administered to examinees from different populations. Then, the calibrated item parameter estimates of the two forms based on separate calibration are placed on two distinct standard scales due to the indeterminacy property. Improper scale linking could result in test results from a group of high-achieving test takers and a group of low-achieving test takers being incorrectly claimed to be at similar levels, undermining the validity of test scores. In addition, the credibility of different psychometric activities, such as test equating and differential item functioning, is threatened.

To address the issue, statistical procedures are often employed to adjust the differences related to group abilities. Scale linking is one such procedure used to transform parameter estimates, including item and ability parameters, onto the same scale. Although different definitions of linking were proposed in the literature (Dorans et al, 2007; Kim & Kolen, 2014), scale linking in this study refers to a procedure specifically used under a common-item nonequivalent groups design (CINEG). With this design, two groups of test takers with different abilities are provided with parallel forms of a test including a set of common items. Since the parameters are calibrated separately, there is no guarantee that the two sets of parameter estimates will be on the same scale.

In practice, two major IRT scale linking approaches are used: characteristic curve methods and moment methods. The moment transformation methods include the mean-mean method (Loyd & Hoover, 1980) as well as the mean-sigma method (Marco, 1977). These methods use the mean of a and b parameters, and the mean and standard deviation of b parameter, respectively, to perform the transformation (Kolen & Brennan, 2014). In contrast, characteristic curve methods, including the Haebara (1980) and Stocking-Lord methods (1983), consider all parameters when estimating transformation coefficients, whereas moment methods only take the first one or two moments of item parameters into account.

Previous literature has explored the topic of scale linking from various perspectives (e.g., Kim & Kolen, 2007, 2022; Kim & Lee, 2006; Vale 1986; von Davier & von Davier 2007). However, an overlooked challenge in scale linking is how to accurately transform the scale in the presence of missing responses. This problem should be investigated because without a proper treatment of missing responses, item and examinee parameters estimates obtained from

examinees associated with different forms can be affected. As a result, inaccurate linking coefficients are likely to compromise the validity of test results.

Missing responses in other contexts have been handled in multiple ways. For example, listwise deletion (LWD) or treating the missing data as incorrect responses (IN) are the two most used methods. More advanced methods have also been used in practice, including substituting missing data with a corrected mean (CM, Bernaards & Sijtsma, 2000), response function imputation (RF, Sijtsma & van der Ark, 2003), multiple imputations (MI), and full information maximum likelihood (FIML, Enders, 2001), each of which has its own advantages. Both CM and RF are suitable for categorical data, a data type that is often found in large-scale tests (Bernaards & Sijtsma, 2000). MI and FIML have shown high levels of accuracy in previous research (Finch, 2008; Mislevy, 2017).

The impact of these methods on scale linking has not yet been investigated. It is likely that various methods of handling missing responses will influence linking coefficients, affecting the accuracy of item and person parameters. This can ultimately compromise the comparability of test scores between two test forms. Therefore, it is crucial to investigate the performance of different methods to treat missing responses and understand how they impact the maintenance of a common IRT scale between two test forms.

Chapter/ Study 2

Test equating, as a statistical procedure, is often used in large-scale assessments to achieve comparable/interchangeable scores across multiple test administrations. If parameters from different groups are placed on the same scale, equating can then be conducted “to adjust scores on test forms so that scores on the forms can be used interchangeably” (Kolen & Brennan, 2014, p.2). Equating ensures that examinees' scores on one test form can be directly compared to

scores on another test form, allowing stakeholders to interpret test results accurately and make valid comparisons.

Study 2 discusses how to achieve test score comparability across different test administrations for rater-mediated assessments that involve human raters. Equating becomes more complex in such assessments because scores assigned by raters can be influenced by various factors. According to Wolfe (2020), in general, rating outputs, such as rating quality, rating speed, and rating attitudes, can be largely influenced by three types of inputs for human-scoring tests, including the characteristics of the raters (e.g., level of proficiency, experiences, education, etc.), the content of the responses (e.g., linguistic feature profiles of essays, handwritten or keyboarded essays, gender and ethnicity of examinees, etc.), and the context of the scoring process (e.g., training raters have received, the focus of the scoring criteria, testing mode, etc.). Researchers have used rater effects to describe the patterns of rater errors.

The literature has identified four common rater effects: severity/leniency, centrality/extremity, accuracy/inaccuracy, and halo effects (Myford & Wolfe, 2003; Wolfe, 2020). In addition, rater behavior can change over time, which is commonly referred to as differential rater functioning over time (DRIFT; Wolfe et al., 1999). Both rater errors and DRIFT can threaten the validity and reliability of a test score (Anthony et al., 2020; Engelhard, 1994; Hoyt, 2000) and therefore affect the accuracy of test equating. For example, in IRT observed-score equating, if rater effects are not faithfully accounted for, they can affect both the conditional and marginal distributions of observed number-correct scores and the estimated distributions of latent traits. Therefore, rater effects can ultimately influence the accuracy of score adjustments between two forms. DRIFT, on the other hand, affects the consistency of

ratings across multiple administrations of a test, making test scores across different time points incomparable.

Several rater models have been proposed to handle measurement errors caused by the use of human raters (e.g., Lincare, 1989; Patz et al., 2002; Qiu et al., 2021). These models have been used in many studies to quantify the bias and variability associated with raters and to investigate their effect on test scores (e.g., Congdon & McQueen, 2000; Harik et al., 2009; Hung et al., 2012; Kassim, 2011; Nieto & Casabianca, 2019). However, how to handle rater errors and rater drift have not been fully investigated in the field of test equating. It remains unclear how to obtain test score comparability while adjusting for discrepancies in form difficulty and accounting for rater effects. An innovative equating method was proposed in study 2 to ensure score comparability across different test administrations by accounting for human raters.

Chapter/ Study 3

It is well acknowledged that randomization is often essential in educational research, as it ensures that all subjects have an equal chance of being assigned to a group, producing two equivalent groups. However, many intervention studies in educational context are not RCT. To achieve comparable subjects in treatment and control groups on the observed propensities, Rosenbaum & Rubin (1983; 1984) developed the PS method by matching the subject's probability of being selected in the treated group to mimic a randomized experiment. One of the assumptions for using the PS method is that all factors related to treatment assignment or outcomes are measured.

The PS method requires a series of steps (Powell et al., 2020). The first step involves calculating the effect size of the between-group differences on all covariates before conducting the PS methods. This step helps to understand the extent to which the equivalence between the

treatment and control groups needs to be improved. In the second step, a propensity score is created for each participant, representing the probability of being selected in the treatment group. Typically, a logistic regression is used to estimate the propensity score by using group membership as the outcome variable (1 for the treatment group and 0 for the control group) and the covariates as predictors. Then, a PS score is applied to each individual using one of the four popular approaches: matching, stratification, weighting, and covariate adjustments. The third step involves assessing the covariate balance between the treatment and control groups. If the difference between the two groups meets a certain criterion, step 4 is carried out. Otherwise, the covariates used in step 2 need to be changed, or the logistic regression model should be adjusted. Step 3 is repeated until the equivalence between the two groups is achieved. In the last step, a treatment effect is estimated by examining group differences on the outcome variable. For each of these four steps, researchers need to decide which method is most suitable for their study from a pool of available methods. Scholars have conducted extensive research to provide rationales for each decision point in using the PS methods by comparing their performance under specific conditions (Austin, 2009; Austin, 2014; Jacovidis et al., 2017; Stone & Tang, 2013).

Among the four steps, the third step, covariate balance diagnostics, is crucial to ensure the comparability of covariates between groups, which influences the validity of the estimated treatment effect. Using an effective covariate balance diagnostic is instrumental in selecting a correct PS model. If the covariate balance diagnostic fails to provide accurate information about the degree of imbalance between two groups, the de facto imbalanced covariates might be incorrectly assessed as balanced. In this case, the undetected imbalance in the covariates can still influence the accuracy of the estimated treatment effect.

Multiple covariate balance diagnostics have been researched with single-level data (e.g., Ali et al., 2015; Austin, 2009). However, limited research has considered the complex data structure inherent in educational settings, where data are hierarchical (e.g., students nested within a class), in evaluating the performance of covariate balance diagnostics. Equivalence between comparison groups can also be achieved by applying PS weighting. Weighting can be conducted through multiplying the observations of the measured variable by a weight based on the propensity scores (Bai & Clark, 2018). When applying PS weighting with multilevel data, using the diagnostics constructed for single-level data may not be sufficient in detecting group imbalance. This is due to the fact that traditional diagnostics only consider variations among individuals, ignoring the information regarding the variations on covariates among clusters. Improper use of covariate balance diagnostics may lead to a misunderstanding of the degree of group imbalance, hindering the accuracy of treatment effect estimation. Therefore, there is a need to examine different covariate balance diagnostics to gain more insight into their applications with multilevel data, considering the use of PS weighting.

Significance

It is essential for educational researchers to ensure comparability for different purposes with the consideration of specific contexts. However, this goal can be difficult to achieve due to a variety of challenges. One such challenge is the selection of an appropriate method, which requires careful consideration of previous research findings as well as the researcher's experience. The dissertation aims to advance the methodology for addressing comparability issues in educational research through three studies.

Both the first and second studies present advancements in methodologies to achieve comparability among test scores of multiple test forms. Test stakeholders can benefit from the

studies in different ways. For example, test takers can be treated fairly by receiving valid scale scores. School administrators can interpret test scores accurately to make better decisions.

The first study of the dissertation uses simulated data to explore the role of six missing data handling approaches in IRT scale transformation. The objective is to aid psychometricians and researchers in determining the most suitable method for handling missing responses, thereby ensuring a fair comparison of IRT parameters. The second study addresses a gap in the literature by proposing an IRT observed-score equating method that accounts for rater error using a hierarchical rater model. Using simulated data to evaluate the new equating model, this study intends to offer a new equating tool for researchers and practitioners. The newly proposed equating method is expected to improve the interchangeability of test scores and help ensure test fairness when data involve errors attributable to human raters.

The third study aims to identify a suitable covariate balance diagnostic for PS weighting with multilevel data to achieve balanced comparison groups. By examining a more comprehensive list of covariate balance diagnostics in the context of PS weighting, which has not been previously explored with nested data structures, this study contributes significant value to existing knowledge. By appropriately utilizing covariate balance diagnostics, researchers can better comprehend the effect of an intervention/program, providing policymakers with evidence to make informed decisions.

This dissertation fills the gap in the literature by addressing comparability issues in education, including scale linking, test equating, and PS weighting. Based on the findings of the dissertation, educational researchers will be able to choose the optimal methods in data processing and analysis procedures to ultimately improve study validity and quality of educational research.

CHAPTER/ STUDY 1: EVALUATING THE EFFECTS OF MISSING DATA HANDLING METHODS ON SCALE LINKING ACCURACY

Wu, T., Kim, S. Y., & Westine, C. (2022). Evaluating the Effects of Missing Data Handling Methods on Scale Linking Accuracy. *Educational and Psychological Measurement*, 00131644221140941.

In large-scale assessments, multiple forms of a test are often administered to multiple groups of test takers whose ability distributions are not equivalent. As a result, parameters that are calibrated separately for each group will be on different scales due to the indeterminacy property (S. Kim & Kolen, 2007; Kolen & Brennan, 2014) of the item response theory (IRT). This property makes it difficult to directly compare person and item parameters across different groups. More importantly, the differences between IRT scales need to be adjusted to conduct various psychometric work, such as item analysis, differential item functioning analysis, form construction, and so on. To deal with this issue, scale linking is usually used to position estimates from different groups to be on a common scale (Kolen & Brennan, 2014; Lee & Lee, 2018). Specifically, linking is indispensable when tests are administered under the common-item nonequivalent groups design (CINEG), where groups differ in ability and, thus, forms are built to share a set of common items in an effort to achieve scale comparability.

Research on IRT scale linking has focused mainly on understanding the efficiency of using various linking approaches (e.g., moment methods and characteristic curve methods; Baker & Al-Karni, 1991; S. Kim & Kolen, 2006; Kim & Lee, 2006; von Davier & von Davier, 2007), examining the performance of different calibration methods (Hanson & *Béguin*, 2002; S. Kim & Kolen, 2006; Lee & Ban, 2009), or extending scale linking to tests with complex structures or with multiple constructs (S. H. Kim & Cohen, 1998; Kim & Kolen, 2006; S. Kim & Lee, 2006;

Li et al., 2004; Li & Lissitz, 2000; Oshima et al., 2000). These studies often assumed that responses collected from examinees are complete before dealing with linking issues. Unfortunately, achieving a complete dataset is almost infeasible in reality. For instance, in Programme for International Student Assessment (PISA) 2012, missing data were found to range from 0.3% to 13.3% at the item level for mathematics and from 0.8% to 10.5% for reading items (OECD, 2012).

Under the CINEG design, linking coefficients are often estimated separately based on the responses to common items. If missing data that occur with the common items are not handled properly, estimated item and ability parameters are possible to be affected, influencing the accuracy of scale linking estimates. In that case, the estimated abilities of examinees will also be biased after being transformed to the common scale due to the inaccurate linking coefficients. Thus, it is crucial to understand the appropriate method to deal with missing data in the scale linking context such that sound decisions can be made to ensure test score validity and improve test fairness.

Several missing data handling approaches have been explored within the IRT context (Cetin-Berber et al., 2019; Finch, 2008; Pohl et al., 2014; Shin, 2016), such as their impacts on the estimation accuracy of item and ability parameters. However, the performance of the methods on scale linking is still unknown. This study was conducted to fill this gap in the literature and to provide implications for practitioners and researchers on how to deal with missing responses for correctly placing group abilities on a common scale. Specifically, missing data treatment approaches were examined under two missing data mechanisms for a set of simulation conditions, including examinees' ability distributions, ratio of common items, missing rates, percentages of common items involving missing data, and test lengths. Six missing data

handling approaches investigated in this study included a) listwise deletion (LWD), b) treating the missing data as incorrect responses (IN), c) substituting the missing data with corrected mean (CM; Bernaards & Sijtsma, 2000), d) applying imputation using response function (RF; Sijtsma & van der Ark, 2003), e) using multiple imputation approach (MI; Rubin, 1987), and f) utilizing full information maximum likelihood method (FIML; Enders, 2001a, 2001b; Finkbeiner, 1979) .

Scale Linking Approaches

This study focuses on the impact of missing data treatment methods on two IRT characteristic curve methods, including Haebara (1980) and Stocking-Lord methods (1983), because they have been found to yield more accurate results compared with the moment methods (Hanson & Béguin, 2002; S. Kim & Kolen, 2006; S. Kim & Lee, 2006; LeBeau, 2017). Both the Haebara and Stocking-Lord methods search for optimal scale transformation constants (slopes and intercepts) to mitigate the difference between characteristic curves across common items. However, they differ in that the Haebara method finds the coefficients by minimizing the differences between item characteristic curves, whereas the Stocking-Lord approach achieves the goal by reducing the differences between test characteristic curves (See Kolen & Brennan, 2014 for more details).

Missing Mechanisms

According to Rubin's theorems (1976), missing data are grouped into three mechanisms, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), depending on how the probability of missingness is related to the missing data. Data that are MCAR imply that the cause of the missingness is completely random. Under the MAR mechanism, the probability of having a data point missing is not dependent on the missing point itself. Instead, it is linked to some additional measured variables (e.g., total scores

over non-missing items can be considered a measured variable). For MNAR data, the likelihood of a missing response cannot be explained to any measurable variables, but caused by unmeasured variable(s) (e.g., missingness depends on the individual's ability). Missing data handling approaches have shown to perform differently based on the type of missing mechanisms (e.g., Cetin-Berber et al., 2019; Finch, 2010; Robitzsch & Rupp, 2009). Mislevy and Wu (1996) suggested considering different missing data mechanisms when estimating parameters for obtaining more accurate results. Sachse et al. (2019) analyzed the PISA data from 35 countries and identified that missing mechanisms and the presence of missing data were considerably different across multiple time points, countries, and domains. Through a simulation study, these two factors were proved to have significant impacts on trend estimates for large-scale assessments.

Large testing organizations tend to deal with missing responses based on the missing types, such as non-administered items, omitted items, and non-reached items. For instance, the National Assessment of Educational Progress (NAEP) assessment considers missing responses that appear before the last observed response as omissions and treats as fractionally correct, whereas the missing responses at the end of a block of items are considered as non-reached items and treated as not present (National Center for Educational Statistics [NCES], 2008). In real-world practice, it is very unlikely that common items are not administered or not present to examinees unless there is a technical or administrative issue, which is rare to occur. Although missing responses on common items could possibly fall under all three missing data mechanisms, it is more reasonable to assume that responses are omitted or non-reached for common items. Previous research found that both omitted items and non-reached items can be affected by the examinee's proficiency levels (Köhler et al., 2015a; Rose et al., 2010; Sachse et

al., 2019). However, they each can be associated with distinctive factors. The presence of omitted items could be attributed to item format (Köhler et al., 2015a), and item difficulty (Rose et al., 2010), whereas non-reached items are possibly influenced by one's motivation level and test-taking strategy (Mislevy & Wu, 1988). These findings indicate that missingness for omitted items and non-reached items are likely to be MAR and MNAR rather than MCAR. The current study, as the first to examine missingness in the scale linking context, intended to identify the most appropriate methods to deal with omitted items under both MAR and MNAR missing mechanisms.

Missing Data Handling Methods

Listwise Deletion and Treating Missing Data as Incorrect

In this study, six missing data handling approaches were investigated either because they have been found to yield accurate estimation of parameters, or they are easy to implement and have been commonly used (Cheema, 2014; Finch, 2008; Hawthorne et al., 2005). In terms of practicality, two of the most straightforward and easiest methods to apply are probably LWD and IN. However, treating missing responses with LWD by excluding the entire record of an examinee if any single value is missing was found to be associated with biased estimates (Enders, 2001b; Robitzsch & Rupp, 2009; Sinharay et al., 2001), except when it is under the MCAR mechanism (van Ginkel et al., 2010, 2020). The use of LWD inherently reduces overall sample size and, in turn, lowers statistical power. Considering field testing where a small number of examinees are involved, applying LWD with missing data seems to be more problematic because little data may remain if examinees who miss answering any items will be removed.

IN assumes that the missingness is MNAR. Specifically, missingness is believed to be resulted from test takers' limited knowledge or skills to perform the tasks, and thus missing

responses are treated as incorrect. However, in practice, missingness can also be related to examinees' gender, motivation, self-efficacy, and the enjoyment of working on the subject (Di Chiacchio et al., 2016), which are not MNAR. Research has shown that IN is associated with biased parameter estimates or theta estimates (De Ayala et al., 2001; Mislevy, 2017; Robitzsch & Rupp, 2009; Rose et al., 2010), but this approach is still commonly used in practice for handling missing data due to its relatively simple application.

In addition to the traditional methods introduced earlier, various imputation methods that are more complex and computationally demanding have been proposed to treat missingness in educational assessments. Imputation is a process where, "the missing values are filled in, and the resultant completed data are analyzed by standard methods" (Little & Rubin, 2020, p.24). Both single imputation methods, including CM and RF, and multiple imputation (MI) method are described subsequently. In addition, FIML is also discussed and examined in this study.

Corrected Item Mean Substitution

The CM method incorporates a weight to represent the examinee's performance relative to the average performance of all examinees on the non-missing items (Bernaards & Sijtsma, 2000). For instance, a higher value is imputed to the missing item response when the examinee's performance on non-missing items is above the average performance. The imputed value X_{ij} for examinee i on missing item j is found as (Bernaards & Sijtsma, 2000; Finch, 2008),

$$X_{ij} = \left\{ \frac{PM_i}{\frac{1}{N_i}\sum_{k_i} IM_{k_i}} \right\} IM_j, \tag{1.1}$$

where PM_i is the mean score for examinee i that can be obtained by averaging the available scores for examinee i across all non-missing items. IM_j is the mean score for item j, which can be calculated by averaging all non-missing scores over examinees on item j. N_i indicates the count of non-missing items for examinee i. IM_{k_i} is the mean score for non-missing item k of

examinee i, which can be obtained by averaging all non-missing scores over examinees on item k. The ratio in the bracket indicates the relative performance of examinee i with PM_i being divided by the mean of item means across all non-missing items. If the ratio is > 1, then examinee i will have a higher score than the average score on item j of all examinees, and vice versa. For dichotomous item responses, the imputed value is rounded to either 0 or 1 (see Bernaards & Sijtsma, 2000; Finch, 2008 for more details).

Response Function

The RF approach is a method using nonparametric regression to impute values for missing data based on a latent trait parameter of an examinee (Sijtsma & van der Ark, 2003). This method assumes that an examinee's score is related to a latent parameter denoted by the rest score, $R_{(-j)i}$, which represents the total score of examinee i on all the items of a test except for missing item j. To impute data using RF, the first step is to calculate the rest score $\hat{R}_{(-j)i}$:

$$\hat{R}_{(-j)i} = PM_i(J-1), \tag{1.2}$$

where PM_i is as defined earlier and J is the count of items over the test. The next step is to calculate the possibility of examinee i obtaining the rest score $\hat{R}_{(-j)i}$, but the method used in the calculation varies by the value of $\hat{R}_{(-j)i}$. When $\hat{R}_{(-j)i}$ is an integer, the possibility is calculated by dividing the number of examinees who have obtained the same rest score and answered the target missing item correctly by the total examinees. If $\hat{R}_{(-j)i}$ is not an integer, the possibility is obtained using linear interpolation based on the percentage of the examinees who got item j correct with the largest integer rest score below $\hat{R}_{(-j)i}$ and the proportion with the smallest integer rest score above $\hat{R}_{(-j)i}$. In the final step, an imputed value for a missing item is drawn from a Bernoulli distribution with the estimated possibility (see Sijtsma & van der Ark, 2003 for more details).

Both CM and RF are single imputation strategies, which generate a complete dataset by filling in a value for each missing point. Finch (2008) found that the amount of bias in difficulty estimates associated with CM was comparable with other commonly used approaches, although a larger bias was found in item discrimination estimates with this method. The use of RF led to less bias than IN and CM but was associated with higher standard errors than CM in many conditions. These two methods were included in this study because they were designed to be applied to categorical data (Finch, 2008). In addition, both methods require relatively less intensive computation. However, using one model to restore missing data does not reflect sampling variability, which might underestimate the standard error in the subsequent statistical analysis (Little & Rubin, 2020).

Multiple Imputation Algorithm

To correct the major flaw of single imputation approaches as discussed above, Rubin (1987) introduced the MI algorithm. In MI, each missing data point is imputed M times. The imputed values are estimated based on the means and variances of the observed data. A standard statistical analysis is then carried out on each imputed dataset. The M sets of data are merged into one dataset (i.e., parameter estimates) by averaging values over M sets of results. For more details about MI, see Little and Rubin (2020) and Finch (2008).

It was found that MI led to accurate results in parameter estimation (Finch, 2008; Kalkan et al., 2018; Mislevy, 2017; Schafer & Graham, 2002; Sijtsma & van der Ark, 2003). The number of imputations for implementing MI needs to be determined by investigators and different numbers of imputations potentially lead to different results (Graham et al., 2007). Some researchers argue that a large number of imputations are needed to ensure the accuracy of parameter estimation (Enders, 2010; White et al., 2011). However, running too many

imputations is often time-consuming and not realistic from a practical aspect. Parameter estimation involving MI is a computationally demanding process, but the introduction of several computer software, such as AMOS (Arbuckle, 2014) has made MI more accessible.

Full Information Maximum Likelihood

Schafer and Graham (2002) pointed out that MI and maximum likelihood are the two state-of-the-art methods to treat missing data in that both work well with MAR data. FIML uses all available information to estimate parameters without imputing incomplete data (Graham, 2009; Schafer & Graham, 2002). This algorithm is also one of the direct maximum likelihood (ML) approaches, as the linear parameter estimates are generated from the raw data, and no other procedure is required (Enders, 2001a). Enders (2001a) offered a more comprehensive description of FIML. The parameters to be estimated in FIML are obtained by maximizing the log-likelihood function as follows,

$$\log L_i = K_i - \frac{1}{2}\log|\sum_i| - \frac{1}{2}(X_i - \mu_i)' \sum_i^{-1}(X_i - \mu_i), \quad (1.3)$$

where X_i is the vector of complete data for examinee i, μ_i are the mean estimates for examinee i, K_i is a constant determined based on the number of data points in the complete dataset, and $\sum_i$ is a covariance matrix corresponding to the observed data (Cetin-Berber et al., 2019; Enders, 2001a).

Treating missing responses with FIML was found to have more accurate results than traditional missing treatment approaches, such as IN, LWD, pairwise deletion, and mean imputation and produce similar or more accurate results to MI (Cetin-Berber et al., 2019; Ender, 2001b; Xiao & Bulut, 2020). Another advantage of FIML is that it does not require additional steps to impute missing values. FIML has become a part of built-in functions in many computer

programs, such as AMOS and flexMIRT version 3.62 (Cai, 2020). Hence, practitioners or researchers may be able to easily incorporate this method into their practice.

Proper techniques for handling missing responses have been demonstrated in the literature in the context of IRT applications, mostly on the effect on the accuracy of item and ability parameter estimation (e.g., De Ayala et al., 2001; Debeer et al., 2017; Finch, 2008; Rose et al., 2015). Other scholars focused on dealing with missing responses in the context of competence tests (Köhler et al., 2015b; 2017), computerized adaptive testing (Cetin-Berber et al., 2019), and differential item functioning (Finch, 2011; Robitzsch & Rupp, 2009); however, thus far, researchers have paid relatively little attention to applications for test scale comparability. Shin (2016) investigated the effects of eight missing data approaches on vertical scaling using real data. IN was found to lead to higher discrimination and difficulty parameters, but it was associated with lower pseudo-guessing parameters. Two of the multiple imputation approaches, treating missing responses as not present (NP) and combining IN and NP (INNP), yielded similar results. In sum, different missing data treatment approaches resulted in different parameter estimates as well as vertical scaling results. However, the findings from this study have limited interpretation due to a lack of a proper evaluation criterion for real data analysis.

Research Questions

Although many studies compare missing data methods in the IRT context, the effect of these methods on scale linking was still unclear. From a practical perspective, how to handle missing responses when transforming scales into a common scale for large-scale assessments has received little attention. The primary purpose of this simulation study was to understand the relative performance of six methods to treat missing data on scale linking, with an intention to help practitioners choose an appropriate treatment.

Specifically, missing responses within two missing mechanisms, including MAR and MNAR, were examined. Also, how to handle omitted responses to common items was the focus of the current study. The following research questions guided the study:

Research Question 1: To what extent do the missing data handling approaches influence the linking accuracy within the MAR and MNAR mechanisms?

Research Question 2: To what extent does each missing data handling approach interact with the six simulation factors under investigation?

Method

Simulation Conditions

The six missing data handling methods were compared across two test lengths, including 30-item and 60-item test forms, as used in previous research (Cetin-Berber et al., 2019). As transformation coefficients are calculated based on responses to common items over examinees under the CINEG design, this study took common items-related factors into consideration. The ratio of common items varied with two levels, including 20% and 40%. Kolen and Brennan (2014) mentioned that the rule of thumb is to have a minimum of 20% of common items for tests involving more than 40 items. In addition, the percentage of common items that has missing responses varied at two levels: 20% and 40%. Taken together, for instance, if the percentage of common items with missing data is fixed at 20%, then missing responses occurred on five common items for a 60-item test form with 24 (40%) common items. Furthermore, three levels of missing rates were examined in the study, including 8%, 15%, and 30%, the rates frequently observed in practice according to previous research (Cetin-Berber, 2019; Finch, 2008).

In addition, the ability distribution of examinees taking the new test form was examined at three levels, including $N(0, 1)$, $N(0.25, 1.1^2)$ and $N(0.5, 1.2^2)$. The three different proficiency

levels were selected after consultation with prior research with an intention to consider three distinct scenarios in which the new and old groups are similar, somewhat different, or extremely different (Kang & Petersen, 2012). The examinees in the old group were drawn from a standard normal distribution $N(0, 1)$. The sample size was fixed at 3000. Table 1.1 presents all conditions considered in this study. Results under all conditions were investigated using the six missing data methods for both the Haebara and the Stocking-Lord linking approaches.

Table 1.1

Simulation Conditions

Factor	Levels	No. of levels
Examinee condition		
Ability distribution of new group	$N(0, 1)$, $N(0.25, 1.1^2)$, $N(0.5, 1.2^2)$	3
Test form conditions		
Test length	30, 60	2
Proportion of common items	20%, 40%	2
Missing data conditions		
Proportion of common items involving missing responses	20%, 40%	2
Missing rates	8%, 15%, 30%	3
Linking Approaches	Haebara, Stocking-Lord	2

Data Generation

To illustrate the simulation procedure, two test forms with 60 items and 20% common items were used as an example. First, 60 items were randomly drawn from an item pool of 800 sets of item parameters that were calibrated from real data, and were considered as the old form-Form Y. Among the items in Form Y, 20% of common items were deliberately selected such that the distributions of the common items represented the full test in terms of statistical characteristics, for example, item difficulty level. Next, the new form, Form X, was built to include common items already selected and another 40 items unique to Form X. The unique items on Form X were selected from the same item pool. In addition, efforts were spent to make

sure that the distributions of the unique items on both forms were similar. In this way, the two forms were created to have similar item characteristics to reflect operational test construction. Finally, the abovementioned repeated for two test lengths (30 and 60 items) and two levels of common items proportions (20% and 40%). Responses to items on each test form were simulated based on the IRT 3PL model using the R program (R Core Team, 2021). The means and standard deviations of item parameters used in simulations are provided in Table 1.2.

Table 1.2

Means and Standard Deviations of Item Parameters Used in Simulations

				Form Y			Form X		
Total items	Common items			a	b	c	a	b	c
60	24	Common items	M	0.71	0.15	0.23	0.71	0.15	0.23
			SD	0.21	0.74	0.06	0.21	0.74	0.06
		Unique items	M	0.74	0.12	0.28	0.74	0.08	0.25
			SD	0.30	1.08	0.12	0.23	0.84	0.11
		Total items	M	0.73	0.13	0.26	0.74	0.11	0.24
			SD	0.26	0.95	0.10	0.22	0.80	0.09
60	12	Common items	M	0.72	0.13	0.26	0.72	0.13	0.26
			SD	0.25	0.69	0.07	0.25	0.69	0.07
		Unique items	M	0.73	0.13	0.26	0.73	0.12	0.26
			SD	0.27	1.01	0.11	0.22	0.87	0.11
		Total items	M	0.73	0.13	0.26	0.73	0.12	0.26
			SD	0.26	0.95	0.10	0.22	0.83	0.10
30	12	Common items	M	0.72	0.32	0.28	0.72	0.32	0.28
			SD	0.20	0.99	0.14	0.20	0.99	0.14
		Unique items	M	0.73	0.27	0.25	0.76	0.30	0.24
			SD	0.25	1.02	0.09	0.22	1.06	0.10
		Total items	M	0.72	0.29	0.26	0.74	0.31	0.25
			SD	0.23	1.00	0.11	0.21	1.02	0.12
30	6	Common items	M	0.74	0.30	0.25	0.74	0.30	0.25
			SD	0.24	0.92	0.14	0.24	0.92	0.14
		Unique items	M	0.72	0.29	0.27	0.75	0.30	0.28
			SD	0.23	1.03	0.10	0.21	1.04	0.10
		Total items	M	0.72	0.29	0.26	0.75	0.30	0.27
			SD	0.23	1.00	0.11	0.21	1.00	0.10

For the MAR scenario, the missing data generation process followed the procedure used in the previous literature (De Ayala et al., 2001; Enders, 2004; Finch, 2008). In this study, the number-correct score over the non-missing items was viewed as the observed variable inversely related to the probability of a data point that was missing. Specifically, to conduct the simulation,

the number-correct score over the non-missing items was summed and then divided into three categories, each of which was assigned a probability of a missing response. For a 60-item test form where five common items were missing, the maximum value of the correct score over the non-missing items was 55. Thus, the possible correct scores for all examinees were classified into three categories, 0 to 18, 19 to 36, and 37 to 55. A missing probability was assigned to each category in which the higher score was inversely related to the missing probability. One condition here was that the averaged probability for each score category should be approximately equal to the desired missing percentage. For example, the probability for each category could be close to 0.4, 0.3, and 0.2 for 0 to 18, 19 to 36, and 37 to 55, respectively, such that the desired missing percentage, on average, was 30%. A range of ± 0.5% of the missing probability was allowed for each category to ensure the correct representation of the ratio of missingness while giving some extra room for the data to be generated (Finch, 2008). The probability of a missing value was then compared to a random value drawn from a uniform distribution. Based on the results of the comparison, the value was reserved or deleted in the datasets.

Under MNAR, responses to test items were assigned a corresponding missing probability such that examinees answering the items incorrectly were more likely to be assigned a higher probability of missingness (Finch, 2008). When generating data, the average probability of missing was set approximately equal to the target missing percentage. For example, to achieve an overall missing percentage of 15%, the missing probability of a correct response was assigned a value of 0.1, whereas the probability of an incorrect response was 0.27. As with MAR, the probability of a missing response was compared with a random value generated from a uniform distribution $U(0, 1)$ to decide whether to retain the value or not.

Table 1.3

Item Parameters of the Common Items with Missing Responses

Total items	Common items	Proportion of common items with missing responses	Item parameters		
			a	*b*	*c*
60	24	40%	0.46	0.87	0.29
			0.66	1.02	0.12
			0.72	1.05	0.16
			0.69	1.09	0.21
			1.02	1.62	0.17
			0.81	0.35	0.23
			0.67	0.37	0.15
			1.09	0.42	0.19
			0.89	0.47	0.21
			0.96	0.79	0.30
60	24	20%	0.46	0.87	0.29
			0.66	1.02	0.12
			0.72	1.05	0.16
			0.69	1.09	0.21
			1.02	1.62	0.17
60	12	40%	1.02	1.62	0.17
			0.46	0.87	0.29
			0.89	0.47	0.21
			1.24	0.27	0.33
			0.61	0.26	0.24
60	12	20%	1.02	1.62	0.17
			0.46	0.87	0.29
30	12	40%	0.59	1.82	0.46
			0.78	1.57	0.04
			0.73	0.97	0.15
			0.57	0.84	0.39
			0.65	0.82	0.11
30	12	20%	0.59	1.82	0.46
			0.78	1.57	0.04
30	6	40%	0.96	0.79	0.30
			0.78	1.57	0.04
30	6	20%	0.96	0.79	0.30

The deletion of data points in the study was carried out only within common items. In fact, missing data within both common and unique items of each form may influence parameter estimates because they are calibrated based on responses to all items in both forms. However, when using separate calibration, transformation coefficients are calculated based on the responses to common items, which makes it more reasonable to assume that missing data within common items lead to a more significant impact on linking coefficients, compared to when

missing occurs to unique items. This assumption led this study to consider only a situation where missing happened within common items. The parameters of the common items involving missing items are summarized in Table 1.3.

Six Missing Data Handling Approaches

The first four methods (LWD, IN, CM, and RF) were applied with the incomplete datasets using R. The R package MICE **(**van Buuren et al., 2015) was employed to implement MI. The package offers an option to conduct MI with dichotomous items by using an iterative approach to estimate logistic regressions and impute missing values using regression estimates for the dependent variable (Vidotto et al., 2015). After treating the incomplete datasets using each method, the computer program flexMIRT was used to estimate item and person parameters. For FIML, missing data were handled by using flexMIRT directly. Once item and ability parameters were obtained, POLYST (Kim & Kolen, 2003) was used to conduct scale linking with the Haebara (1980) and the Stocking-Lord (1983) methods.

Evaluation Criteria

The extent to which the estimated slopes and intercepts are deviant from the true linking relationship was used to evaluate the relative performance of the six missing data methods. If no error is involved in the linking process, the exact values of the transformation constants A and B are equal to σ_N/σ_O and $(\mu_N - \mu_O)/\sigma_O$, respectively. In the equation, μ_O and σ_O represent the mean and standard deviation of the proficiencies for the examinees on the old-form scale (θ_O), respectively. Similarly, μ_N and σ_N indicate the corresponding values of the new-form scale (θ_N). A more complete description of scale linking transformants can be found in S. Kim & Kolen (2007).

Three indices were used including: absolute bias ($BIAS$), standard error (SE), and root mean squared error ($RMSE$), which indicate the systematic, random, and overall errors in estimates, respectively:

$$Bias = \left|\frac{1}{N}\sum_{r=1}^{N}\hat{Z}_r - Z\right|, \tag{1.4}$$

$$SE = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{Z}_r - \bar{Z})^2}\ , \text{and} \tag{1.5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{Z}_r - Z)^2}\ , \tag{1.6}$$

where N is the number of replications (N = 100). In each equation, $\hat{Z}_r$ represents the estimated slope or intercept obtained at the r^{th} replication, Z indicates the criterion values of the slope or intercept, and $\bar{Z}$ refers to the average value of Z over N replications.

Results

Overall Performance of Missing Data Handling Methods

Aggregated results are presented across all study conditions for each type of missing data mechanism in Table 1.4. The results from the Haebara approach are presented only, as similar patterns were observed between the Haebara and the Stocking-Lord linking approaches.

Missing at Random. Table 1.4 shows minimal differences in linking errors among the missing data handling methods under the MAR missing mechanisms, except for LWD. Treating missing data with LWD produced a significantly large amount of errors than the other methods in recovering both slopes and intercepts. IN yielded the second largest bias and root mean squared error (RMSE) in slopes but led to the smallest errors in intercepts under MAR for the Haebara approach. CM produced small errors in slopes but yielded the second largest errors in intercepts. The other three approaches, including MI, FIML, and RF, tended to be not only associated with small errors in slopes but also in intercepts. In general, these three methods provided the most accurate linking results when the missingness was MAR.

Missing Not at Random. Although differences in the performance of the six methods were found between MAR and MNAR, the overall patterns were similar. The LWD approach yielded the largest amount of errors in both slopes and intercepts under MNAR, as can be seen in Table 1.4. Overall, differences among the approaches except LWD were minimal. In terms of the performance of the other five approaches, both CM and IN seemed to be associated with slightly larger bias and RMSE in slopes than other three approaches. For intercepts, CM yielded larger errors than the other four approaches across all three evaluation criteria, and IN generally led to small errors. As with the results for MAR, RF, MI, and FIML were likely to produce more accurate linking coefficients in both slopes and intercepts as compared with other approaches.

Table 1.4

Linking Error from Different Missing Data Handling Methods under MAR and MNAR using the Haebara Linking Approach

Linking coefficient	Missing data handling method	MAR			MNAR		
		Bias	SE	RMSE	Bias	SE	RMSE
Slope	LWD	**0.044**	**0.049**	**0.054**	**0.042**	**0.050**	**0.052**
	IN	0.033	0.034	0.040	0.034	0.035	0.041
	CM	0.030	0.035	0.037	0.033	0.036	0.040
	RF	0.029	0.035	0.037	0.030	0.036	0.038
	MI	0.029	0.036	0.036	0.029	0.036	0.036
	FIML	0.029	0.036	0.036	0.029	0.036	0.036
Intercept	LWD	**0.059**	**0.060**	**0.073**	**0.050**	**0.060**	**0.062**
	IN	0.030	0.037	0.038	0.031	0.038	0.039
	CM	0.034	0.039	0.043	0.038	0.040	0.046
	RF	0.030	0.037	0.038	0.031	0.038	0.039
	MI	0.031	0.037	0.039	0.032	0.038	0.040
	FIML	0.031	0.037	0.039	0.031	0.037	0.039

Note. The largest and the smallest values across the six methods were bolded and underlined, respectively. MAR = missing at random; MNAR = missing not at random; LWD = listwise deletion; IN = incorrect response; CM = corrected mean; RF = response function; MI = multiple imputation; FIML = full information maximum likelihood; RMSE = root mean squared error.

The Impact of Test and Examinee Conditions

Linking accuracy was calculated and compared across different test conditions for each simulation factor to better connecting the performance of missing data treatment methods with real-world practices. The results are presented in Tables 1.5 to 1.9 and discussed in the following paragraphs.

Test Length. In MAR and MNAR, most of the missing data handling methods, except for LWD, followed a similar pattern where they tended to produce slightly greater accuracy for a form with more items, as seen in Table 1.5. CM generally followed this trend, but it led to a similar amount of bias and RMSE in intercepts for a 30-item test and a 60-item test. Notably, treating missing responses with LWD showed a completely different pattern by leading to more errors, particularly in intercepts, as the number of items on a test form increased. By further examining the datasets, it was found that using LWD, responses from more examinees were deleted for a longer test, which caused greater errors in linking.

Ratio of Common Items. In general, most of the missing data treatment methods yielded slightly more accurate linking coefficients when there was a larger ratio of common items on a test form, as shown in Table 1.6. The exception was with LWD and CM. LWD yielded substantially larger errors in slopes and intercepts as the ratio of common items increased. This finding is partly explained by the fact that more responses involved missing values when there were more common items and using LWD led to deletion of those cases, which then affected the accuracy of scale linking negatively. CM yielded smaller errors for a higher ratio of common items under most conditions, but it produced larger values of bias and RMSE in intercepts regardless of the ratio of common items under MNAR. In some real-world practices, achieving a proportion of common items of 20% or 40% can be challenging. Thus, a condition where 10% of

the items are common items was also investigated in the study for the 60-item test forms. No substantial changes were identified in the results under this condition in terms of the performance of the six missing data handling methods.

Missing Rates. In general, MI, FIML, and RF were shown to be robust to missing rates: a drift was minimal in linking accuracy across the three levels of missing rates in both MAR and MNAR, which can be found in Table 1.7. In contrast, the influence of missing rates on the other methods was more apparent. LWD introduced substantial errors for a test form with higher missing rates. Note that when LWD was used, the rate of change in errors was greater for a form with a missing percentage above 15% than a form with a lower missing percentage. Also, when CM and IN were employed, larger bias and RMSE were found in slopes and intercepts for a missing percentage of 30% as compared with those for 8% and 15% under most conditions.

The Proportion of Common Items With Missing Data. The performance of RF, MI and FIML was very similar regardless how many common items involved missing data, based on Table 1.8. Similarly, the errors produced by IN and CM remained constant under most test conditions with a few exceptions. For instance, there was only a slight increase in bias and RMSE in slopes for IN as missing data were observed within more common items. Similarly, relatively more errors were also found in intercept using CM, especially under the MNAR mechanism. LWD yielded considerably more errors for a form with a larger proportion of common items containing missing values.

Ability Distribution. According to Table 1.9, the largest error was consistently found when the new group differed the most from the old group, which was $N(0.5, 1.2^2)$. Slightly more or similar linking errors were found when the new group followed $N(0.25, 1.1^2)$ as compared with $N(0, 1)$ when using IN, CM, RF, MI, or FIML. The rate of change in errors was greater

when the new group had an ability distribution of $N(0.25, 1.1^2)$ or higher as compared with the rate for a lower proficiency, which was more evident for IN and CM. LWD also yielded a large amount of errors when the new group was of a higher proficiency.

The Influence of Linking Approaches

In addition to the five simulation factors discussed, the efficiency of the six methods to treat missing data was also compared for both the Haebara and the Stocking-Lord methods. In general, a fairly consistent results were found between the two linking methods. The Stocking-Lord seemed to produce slightly more errors in slopes than the Haebara approach, with a few exceptions. For intercepts, results from the two linking methods were almost identical.

Table 1.5

Linking Error from Different Missing Data Handling Methods by Test Lengths Using the Haebara Approach

Linking coefficient	Missing data handling method	MAR						MNAR					
		Bias		SE		RMSE		Bias		SE		RMSE	
		30 Items	60 Items	30 Items	60 Items	30 Items	60 Items	30 Items	60 Items	30 Items	60 Items	30 Items	60 Items
Slope	LWD	0.041	0.048	0.048	0.050	0.051	0.058	0.039	0.044	0.048	0.053	0.049	0.055
	IN	0.035	0.030	0.038	0.031	0.043	0.037	0.035	0.033	0.039	0.031	0.043	0.040
	CM	0.033	0.027	0.037	0.032	0.041	0.034	0.034	0.032	0.039	0.033	0.042	0.039
	RF	0.033	0.026	0.038	0.032	0.041	0.033	0.033	0.027	0.039	0.034	0.041	0.034
	MI	0.032	0.026	0.039	0.032	0.040	0.033	0.032	0.026	0.039	0.033	0.040	0.033
	FIML	0.032	0.026	0.039	0.032	0.039	0.032	0.032	0.026	0.039	0.032	0.039	0.033
Intercept	LWD	0.046	0.073	0.051	0.068	0.057	0.089	0.041	0.058	0.051	0.070	0.052	0.072
	IN	0.032	0.028	0.039	0.034	0.040	0.035	0.034	0.028	0.041	0.034	0.042	0.035
	CM	0.034	0.035	0.041	0.036	0.042	0.043	0.038	0.038	0.043	0.037	0.047	0.046
	RF	0.032	0.029	0.040	0.035	0.040	0.036	0.033	0.029	0.040	0.035	0.041	0.036
	MI	0.033	0.030	0.040	0.035	0.041	0.037	0.034	0.030	0.040	0.035	0.042	0.038
	FIML	0.033	0.029	0.040	0.034	0.041	0.036	0.033	0.029	0.040	0.034	0.041	0.036

Note. MAR = missing at random; MNAR = missing not at random; LWD = listwise deletion; IN = incorrect response; CM = corrected mean; RF = response function; MI = multiple imputation; FIML = full information maximum likelihood; RMSE = root mean squared error.

Table 1.6

Linking Error from Different Missing Data Handling Methods by Proportions of Common Items Using the Haebara Approach

Linking coefficient	Missing data handling method	MAR						MNAR					
		Bias		SE		RMSE		Bias		SE		RMSE	
		20%	40%	20%	40%	20%	40%	20%	40%	20%	40%	20%	40%
Slope	LWD	0.041	0.047	0.050	0.049	0.051	0.058	0.040	0.043	0.050	0.051	0.050	0.054
	IN	0.036	0.030	0.039	0.030	0.044	0.036	0.036	0.031	0.039	0.030	0.044	0.038
	CM	0.034	0.026	0.039	0.030	0.042	0.032	0.035	0.030	0.039	0.032	0.044	0.037
	RF	0.033	0.026	0.040	0.031	0.041	0.032	0.033	0.027	0.041	0.032	0.042	0.034
	MI	0.033	0.025	0.041	0.031	0.041	0.031	0.033	0.025	0.041	0.031	0.041	0.031
	FIML	0.032	0.025	0.040	0.031	0.040	0.031	0.032	0.025	0.040	0.031	0.041	0.031
Intercept	LWD	0.048	0.070	0.052	0.068	0.060	0.086	0.041	0.058	0.051	0.070	0.052	0.072
	IN	0.032	0.028	0.039	0.034	0.040	0.035	0.033	0.028	0.041	0.035	0.041	0.036
	CM	0.035	0.034	0.041	0.036	0.043	0.042	0.036	0.040	0.042	0.038	0.044	0.049
	RF	0.033	0.028	0.04	0.035	0.041	0.035	0.033	0.029	0.041	0.035	0.041	0.036
	MI	0.033	0.029	0.04	0.034	0.042	0.036	0.034	0.029	0.04	0.035	0.043	0.037
	FIML	0.033	0.029	0.04	0.034	0.041	0.036	0.033	0.029	0.04	0.034	0.042	0.036

Note. MAR = missing at random; MNAR = missing not at random; LWD = listwise deletion; IN = incorrect response; CM = corrected mean; RF = response function; MI = multiple imputation; FIML = full information maximum likelihood; RMSE = root mean squared error.

Table 1.7

Linking Error from Different Missing Data Handling Methods by Missing Rates Using the Haebara Approach

Missing mechanism	Linking coefficient	Missing data handling method	Bias			SE			RMSE		
			8%	15%	30%	8%	15%	30%	8%	15%	30%
MAR	Slope	LWD	0.036	0.039	0.057	0.041	0.045	0.062	0.045	0.049	0.070
		IN	0.029	0.031	0.038	0.036	0.032	0.035	0.037	0.038	0.046
		CM	0.029	0.028	0.033	0.035	0.033	0.036	0.036	0.034	0.041
		RF	0.029	0.028	0.031	0.036	0.033	0.036	0.036	0.034	0.039
		MI	0.029	0.027	0.031	0.036	0.034	0.037	0.037	0.034	0.038
		FIML	0.029	0.027	0.030	0.036	0.034	0.037	0.036	0.034	0.037
	Intercept	LWD	0.046	0.053	0.079	0.042	0.051	0.085	0.055	0.065	0.100
		IN	0.029	0.030	0.031	0.035	0.037	0.037	0.036	0.038	0.039
		CM	0.030	0.033	0.039	0.036	0.039	0.041	0.038	0.042	0.048
		RF	0.030	0.031	0.031	0.036	0.037	0.038	0.037	0.039	0.038
		MI	0.030	0.032	0.032	0.036	0.038	0.038	0.037	0.040	0.040
		FIML	0.030	0.031	0.031	0.036	0.038	0.038	0.037	0.039	0.039
MNAR	Slope	LWD	0.033	0.037	0.055	0.042	0.045	0.065	0.042	0.046	0.069
		IN	0.031	0.032	0.038	0.036	0.033	0.036	0.039	0.039	0.045
		CM	0.029	0.029	0.039	0.036	0.033	0.038	0.037	0.036	0.048
		RF	0.029	0.028	0.033	0.036	0.034	0.039	0.037	0.035	0.041
		MI	0.029	0.027	0.031	0.036	0.034	0.037	0.037	0.034	0.038
		FIML	0.029	0.027	0.030	0.036	0.034	0.037	0.037	0.034	0.037
	Intercept	LWD	0.036	0.043	0.070	0.043	0.051	0.087	0.045	0.053	0.087
		IN	0.029	0.031	0.033	0.036	0.038	0.040	0.036	0.038	0.041
		CM	0.032	0.036	0.046	0.037	0.039	0.044	0.039	0.044	0.055
		RF	0.030	0.031	0.032	0.036	0.038	0.040	0.037	0.039	0.040
		MI	0.030	0.032	0.033	0.036	0.038	0.039	0.038	0.040	0.041
		FIML	0.030	0.031	0.032	0.036	0.038	0.038	0.037	0.039	0.040

Note. MAR = missing at random; MNAR = missing not at random; LWD = listwise deletion; IN = incorrect response; CM = corrected mean; RF = response function; MI = multiple imputation; FIML = full information maximum likelihood; RMSE = root mean squared error.

Table 1.8

Linking Error from Different Missing Data Handling Methods by Percentages of Common Items Involving Missing Data Using the Haebara Approach

Linking coefficient	Missing data handling method	MAR						MNAR					
		Bias		SE		RMSE		Bias		SE		RMSE	
		20%	40%	20%	40%	20%	40%	20%	40%	20%	40%	20%	40%
Slope	LWD	0.037	0.051	0.044	0.055	0.046	0.063	0.035	0.048	0.043	0.057	0.044	0.060
	IN	0.030	0.035	0.034	0.034	0.037	0.043	0.030	0.038	0.035	0.035	0.037	0.045
	CM	0.030	0.030	0.035	0.034	0.037	0.037	0.032	0.033	0.035	0.036	0.040	0.041
	RF	0.029	0.030	0.035	0.036	0.036	0.037	0.029	0.032	0.035	0.037	0.036	0.039
	MI	0.029	0.029	0.036	0.036	0.036	0.037	0.029	0.029	0.036	0.036	0.036	0.037
	FIML	0.029	0.029	0.035	0.036	0.036	0.036	0.029	0.029	0.035	0.036	0.036	0.036
Intercept	LWD	0.046	0.072	0.048	0.071	0.057	0.089	0.039	0.06	0.048	0.073	0.049	0.075
	IN	0.030	0.030	0.037	0.037	0.037	0.038	0.031	0.031	0.038	0.038	0.038	0.039
	CM	0.033	0.035	0.038	0.039	0.041	0.044	0.035	0.041	0.039	0.041	0.043	0.049
	RF	0.030	0.030	0.037	0.037	0.038	0.038	0.031	0.031	0.037	0.038	0.038	0.039
	MI	0.031	0.032	0.037	0.037	0.039	0.039	0.031	0.032	0.037	0.038	0.039	0.040
	FIML	0.031	0.031	0.037	0.037	0.039	0.039	0.031	0.031	0.037	0.037	0.039	0.039

Note. MAR = missing at random; MNAR = missing not at random; LWD = listwise deletion; IN = incorrect response; CM = corrected mean; RF = response function; MI = multiple imputation; FIML = full information maximum likelihood; RMSE = root mean squared error.

Table 1.9

Linking Error from Different Missing Data Handling Methods by Ability Distributions Using the Haebara Approach

Missing mechanism	Linking coefficient	Missing data handling method	Bias			SE			RMSE		
			N(0, 1^2)	N(0.25, 1.1^2)	N(0.5, 1.2^2)	N(0, 1^2)	N(0.25, 1.1^2)	N(0.5, 1.2^2)	N(0, 1^2)	N(0.25, 1.1^2)	N(0.5, 1.2^2)
MAR	Slope	LWD	0.037	0.042	0.053	0.046	0.048	0.053	0.047	0.052	0.064
		IN	0.026	0.030	0.042	0.032	0.034	0.037	0.033	0.037	0.051
		CM	0.026	0.029	0.035	0.032	0.034	0.038	0.032	0.036	0.043
		RF	0.026	0.029	0.033	0.033	0.035	0.038	0.033	0.036	0.041
		MI	0.027	0.028	0.032	0.033	0.035	0.039	0.034	0.036	0.040
		FIML	0.027	0.028	0.031	0.033	0.035	0.039	0.033	0.035	0.039
	Intercept	LWD	0.045	0.056	0.078	0.056	0.059	0.064	0.057	0.070	0.093
		IN	0.029	0.028	0.033	0.036	0.035	0.038	0.036	0.036	0.040
		CM	0.030	0.032	0.041	0.038	0.037	0.041	0.038	0.040	0.050
		RF	0.029	0.029	0.034	0.036	0.036	0.039	0.036	0.036	0.042
		MI	0.029	0.029	0.035	0.036	0.036	0.040	0.037	0.037	0.044
		FIML	0.029	0.029	0.034	0.036	0.036	0.039	0.036	0.037	0.042
MNAR	Slope	LWD	0.037	0.041	0.047	0.047	0.050	0.054	0.047	0.051	0.058
		IN	0.027	0.031	0.044	0.033	0.034	0.037	0.033	0.038	0.052
		CM	0.027	0.032	0.039	0.033	0.035	0.039	0.034	0.039	0.048
		RF	0.027	0.029	0.034	0.034	0.036	0.039	0.034	0.037	0.042
		MI	0.027	0.029	0.032	0.033	0.035	0.039	0.033	0.036	0.040
		FIML	0.027	0.028	0.031	0.033	0.035	0.039	0.033	0.035	0.039
	Intercept	LWD	0.047	0.049	0.053	0.058	0.060	0.064	0.058	0.061	0.066
		IN	0.030	0.029	0.033	0.038	0.036	0.039	0.038	0.037	0.041
		CM	0.032	0.035	0.047	0.04	0.039	0.042	0.040	0.043	0.056
		RF	0.030	0.029	0.034	0.037	0.036	0.040	0.037	0.037	0.042
		MI	0.030	0.030	0.036	0.037	0.036	0.040	0.037	0.038	0.045
		FIML	0.029	0.029	0.034	0.036	0.036	0.039	0.037	0.037	0.043

Note. MAR = missing at random; MNAR = missing not at random; LWD = listwise deletion; IN = incorrect response; CM = corrected mean; RF = response function; MI = multiple imputation; FIML = full information maximum likelihood; RMSE = root mean squared error

Discussion

Large-scale testing programs often administer multiple forms of a test to eliminate the chances of cheating and improve test security. One of the challenges under IRT is how to maintain scales consistently across different forms of a test. Scale linking methods are commonly used for large-scale assessments to achieve group comparability when common items are included in multiple test administrations. Under a CINEG, scale linking is a prerequisite to conducting many psychometric works when using a separate calibration method. For example, without having accurate linking coefficients, it is impossible for researchers to obtain precise equating coefficients, which will then undermine the validity of test scores. So far, many aspects of scale linking have been examined, but the understanding of the proper method to address missing responses in the process of scale linking is still rudimentary. Specifically, there is no clear guidance in selecting an appropriate approach to handling missing data with the consideration of real-world test conditions and missingness assumptions.

This study presented the results of a simulation study to understand the relative performance of using six different missing data handling methods on scale linking under two missing data mechanisms. Furthermore, a set of simulation conditions was varied, with the intent to create a comprehensive picture of the behaviors of missing data treatment methods in the context of scale linking.

In general, RF, MI and FIML consistently demonstrated their superior performance over other methods based on the overall results, although some discrepancies were found between two mechanisms. In contrast, LWD always produced the largest errors regardless of the levels of factors used for simulating the datasets. CM and IN were identified to be associated with slightly larger or similar errors as compared with RF, MI, and FIML. However, it is crucial for

researchers and practitioners to pay close attention to the specifics of test conditions when applying CM and IN in practice. For example, CM tended to generate substantially more errors when 15% of the responses to the common items were missing than that of a smaller missing rate. IN resulted in larger errors as the proficiency of the examinees taking the new form followed a distribution of $N(0.5, 1.2^2)$ as compared with a less proficient group.

One of the major findings of this study was that MI and FIML seemed to consistently yield the smallest errors under the two missing mechanisms, which is in line with previous work (Ender, 2001b; Finch, 2008; Olinsky et al., 2003; Peyre et al., 2011). Prior research found that MI and ML tended to have a similar amount of errors (Collins et al., 2001), and the current study confirmed the trend in the context of scale linking. Using MI may be computational demanding and time-consuming in practice, especially when imputation is needed for large datasets from multiple administrations. Although MICE package has made multiple imputation more accessible, researchers still need to carry out the linking procedure multiple times to obtain the final averaged linking estimates over multiple imputations. Imputation is not required for FIML, but it also needs certain software to be available.

As introduced previously, the most inaccurate results were led by LWD in both MAR and MNAR, which seemed to be consistent with what was found in previous research (Enders, 2004; Robitzsch & Rupp, 2009). Once a large proportion of responses are deleted by using LWD, the remaining sample may not well represent the whole population such that biased results can be generated. In the context of scale linking, scales estimated from two forms are supposed to be placed on a same scale based on the responses to common items. If a subgroup of examinees who underperform did not respond to those items, item parameter estimates are likely to be

influenced by not having a wide range of examinees, which will in turn affect scale linking accuracy.

A number of studies pointed out that using IN to handle missing data in the IRT context is not ideal (Cetin-Berber et al., 2019; Finch, 2008; Köhler et al., 2017; Pohl et al., 2014; Zhang & Walker, 2008). However, the linking results associated with IN presented in this study under MAR and MNAR were shown to be mixed. Specifically, this method led to slightly larger bias and RMSE than RF, MI, and FIML in slopes but had comparable or smaller results in intercepts. Notably, this method is relatively sensitive to a few simulation factors examined in this study, such as missing rate, the proportion of common items involving missing responses, and the difference in the ability distributions of the two groups for linking. Therefore, given that IN is one of the commonly used methods to treat missing responses in large-scale assessments, for example in PISA 2018 (OECD, 2020), more caution is needed to implement IN by taking different test conditions into consideration, especially in the context of scale linking.

The findings of the two single imputation methods seemed to be more complex. CM tended to produce slightly larger errors compared with other methods, except for LWD, under a few conditions. This might be related to the fact that the CM method involves the computation of an item mean using scores over non-missing responses divided by the number of examinees. With the increase in missing responses, information becomes limited in computing the item mean. By contrast, RF constantly resulted in small errors that were comparable to those of MI and FIML, which is aligned with the findings from previous studies (Finch, 2008, 2011). The calculation of RF does not require an item mean. Rather, it depends on rest scores for imputation, which accounts for examinees' information across the entire form. Based on the design of the

study, missingness only occurred within common items. Borrowing information from unique items with no missing data seemed to contribute to maintaining the accuracy of the imputation.

In terms of the simulation factors, RF, FIML, and MI led to the most accurate linking coefficients regardless of the study conditions. LWD was the most sensitive method to the choice of simulation factors under a variety of the conditions. The performance of IN and CM was slightly or moderately influenced by the simulation conditions. In general, the performance of the six missing data handling methods was consistent between the Haebara and the Stocking-Lord linking approaches. However, CM had slightly better performance when the Stocking-Lord method was used.

Conclusion and Future Research

Based on the results observed in this study, RF, MI, and FIML revealed to introduce a relatively small amount of errors for conducting scale linking. It is an important finding that RF, as a less complex method, also demonstrated robust performance under most of the test conditions investigated in the study. Other than using the two well-known missing data handling methods, MI and FIML, researchers may also consider imputing missing responses with RF for scale transformation. Another notable finding of this study is that researchers may avoid using LWD as it consistently revealed a large amount of error across various study conditions.

The list of missing data treatment approaches examined in the study is by no means exhaustive. In addition, the current study only focused on how to handle missing responses for omitted items specifically, without considering the treatment of non-reached item which might be more closely related to examinee's motivational factors. Research has demonstrated that model-based approaches can provide more accurate parameter estimates (Debeer et al., 2017; Rose et al., 2010) due to their ability in treating omitted and no-reached items differently. In

future studies, the accuracy of more complex model-based approaches on scale transformation can be investigated for missing responses caused by different reasons. In addition, the sample size was fixed at 3,000 in this study. Researchers might consider varying this factor and exploring its interaction with multiple missing data handling approaches.

The current study serves as a starting point in developing a better understanding of how to treat missing responses in maintaining IRT scales. Efforts to make linking estimates accurate when missing responses present are an essential step in ensuring the comparability of item and person parameters and the validity of test scores. The goal of the study is to assist psychometricians and researchers in selecting the most appropriate approach for dealing with missing data in scale transformation. Future work may consider extending the current study design to the context of test equating for improving test form comparability and test fairness with the presence of missing responses.

References

Arbuckle, J. L. (2014). *Amos 23.0 User's Guide.* IBM SPSS.

Baker, F. B., & Al-Karni, A. (1991*).* A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*, 147–162.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*(3), 321-364.

Cai, L. (2020). *flexMIRT®*: *Flexible multilevel multidimensional item analysis and test scoring* (version 3.62) [Computer software]. Vector Psychometric Group.

Cetin-Berber, D. D., Sari, H. I., & Huggins-Manley, A. C. (2019). Imputation Methods to Deal with Missing Responses in Computerized Adaptive Multistage Testing. *Educational and Psychological Measurement*, *79*(3), 495-511.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487-508.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351.

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

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, *54*(3), 333-363.

Di Chiacchio, C., De Stasio, S., & Fiorilli, C. (2016). Examining how motivation toward science contributes to omitting behaviours in the Italian PISA 2006 sample. *Learning and Individual Differences*, *50*, 56-63.

Enders, C. K. (2001a). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*(1), 128-141.

Enders, C. K. (2001b). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*(5), 713-740.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement*, *64*(3), 419-436.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245.

Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. *Applied Measurement in Education*, *24*(4), 281-301.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, *44*, 409–420.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206-213.

Haebara, T. (1980). Equating logistic ability scales by a weighted least square method. *Japanese Psychological Research*, *22*(3), 144-149.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24.

Hawthorne, G., Hawthorne, G., & Elliott, P. (2005). Imputing cross-sectional missing data: Comparison of common techniques. *Australian & New Zealand Journal of Psychiatry*, *39*(7), 583-590.

Kalkan, Ö. K., Yusuf, K. A. R. A., & Kelecioğlu, H. (2018). Evaluating performance of missing data imputation methods in IRT analyses. *International Journal of Assessment Tools in Education*, *5*(3), 403-416.

Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review*, *13*(2), 311-321.

Kim, S., & Kolen, M. J. (2003). *POLYST: A computer program for polytomous IRT scale transformation.* Iowa city: Iowa Testing Programs, The University of Iowa. http://www.education.uiowa.edu/casma

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357-381.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*(4), 371-397.

Kim, S., & Lee, W. C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, *43*(1), 53-76.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131-143.

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

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

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397-419.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

LeBeau, B. (2017). Ability and prior distribution mismatch: An exploration of common-item linking methods. *Applied psychological measurement*, *41*(7), 545-560.

Lee, W. C., & Ban, J. C. (2009). A comparison of IRT linking procedures. *Applied Measurement in Education*, *23*(1), 23-48.

Lee, W. C., & Lee, G. (2018). IRT linking and equating. In *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 639-673).

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*(2), 115-138.

Li, Y. H., Tam, H. P., & Tompkins, L. J. (2004). A comparison of using the fixed common-precalibrated parameter method and the matched characteristic curve method for linking multiple-test items. *International Journal of Testing*, *4*(3), 267-293.

Little, R. J., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.

Mislevy, R. J. (2017). Missing responses in item response modeling. In W. J. van der Linden (Eds.), Handbook of modern item response theory, (1st ed., Vol. 2, pp. 171-194), Cahmpman & Hall/CRC Press.

Mislevy, R. J., & Wu, P. K. (1988). Inferring examinee ability when some item responses are missing. *ETS Research Report Series*, *1988*(2), Article i-75.

Mislevy, R. J., & Wu, P. K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, *1996*(2), Article i-36.

National Center for Educational Statistics. (2008). *NAEP Technical Documentation*. https://nces.ed.gov/nationsreportcard/tdw/analysis/2000_2001/scaling_missing.aspx

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, *151*(1), 53-79.

Organization for Economic Co-operation and Development. (2012). *PISA 2012.* https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

Organization for Economic Co-operation and Development. (2020). *PISA 2018 technical report.* https://bit.ly/3zWbidA

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, *37*(4), 357-373.

Peyre, H., Leplège, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, *20*(2), 287-300.

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

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, *69*(1), 18-34.

Rose, N., von Davier, M., & Nagengast, B. (2015). Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psychological Test and Assessment Modeling*, *57*(4), 472-498.

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), Article i-53.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

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

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

Shin, A. Y. (2016). *Investigating the effects of missing data treatments on item response theory vertical scaling* [Doctoral dissertation, University of Iowa].

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*(4), 505-528.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*(4), 317-329.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201-210.

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015). *Package "mice"* [Computer software]. https://cran.r-project.org/web/packages/mice/index.html

van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of personality assessment*, *102*(3), 297-308.

van Ginkel, J. R., Sijtsma, K., van der Ark, L. A., Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item score imputation. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 17-30.

Vidotto, D., Vermunt, J. K., & Kaptein, M. C. (2015). Multiple imputation of missing categorical data using latent class models: state of the art. *Psychological Test and Assessment Modeling*, *57*(4), 542-576.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*(3), 115-124.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399.

Xiao, J., & Bulut, O. (2020). Evaluating the performances of missing data handling methods in ability estimation from sparse data. *Educational and Psychological Measurement*, *80*(5), 932-954.

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person—model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466-479.

CHAPTER/ STUDY 2: IRT OBSERVED-SCORE EQUATING FOR RATER-MEDIATED ASSESSMENTS USING A HIERARCHICAL RATER MODEL

Introduction

Multiple forms of a test are developed in large testing programs to ensure test security and fairness. Although extensive efforts are made to build parallel forms of similar test characteristics, differences across forms may still exist. Test equating, as one of the key psychometric practices, is often employed to address the discrepancy in a difficulty level between test forms. According Kolen and Brennan (2014), test equating is a process to adjust the discrepancies in difficulty across test forms, ensuring the comparability and interchangeability of scores among multiple parallel forms of a test. Proper equating of multiple test forms not only ensures the validity of a test, but also improves the accuracy and efficiency of test result interpretations for stakeholders, such as educators, policymakers, and examinees.

For multiple-choice (MC) tests, equating can be done relatively easily as long as the forms contain enough items to be used. More importantly, the scoring process for MC items is highly reliable with the aid of computers. However, rater-mediated assessments that require human raters to assign scores to examinees based on their performance on a particular construct(s) (Wang & Engelhard, 2019; Wind, 2019) is likely to introduce rater errors due to the differences in individual raters' training and experience, personal beliefs, and cultural backgrounds, etc. (Barret, 2001; Lee, 2009; Long & Pang, 2015; Wexley & Youtz, 1985; Wolfe, 2020). In fact, the presence of rater errors in rater-mediated assessments is almost unavoidable (Myford & Wolfe, 2003), making it more difficult to obtain test score interchangeability across administrations. Given that test equating plays an important role in ensuring test fairness and test score comparability, practitioners and researchers must be equipped to handle rater errors in

rater-mediated assessments in the equating process. This study aims to provide a new equating method in conjunction with a rater model that takes rater errors into account.

This study intends to incorporate a rater model into an IRT observed-score equating approach for improving the equating accuracy. Specifically, the rater model used in the study is the hierarchical rater model (HRM, Patz, et al., 2002), which faithfully models the hierarchical structure of rating scores between raters and examinees' responses.

Rater Effects and Models

Rater Effects

Previous research has investigated multiple variations of rater effects, such as rater severity/leniency, extremity/centrality, accuracy/inaccuracy, and halo effects (Engelhard, 1994; Engelhard, 2002; Myford & Wolfe, 2003; Wolfe, 2020; Wolfe & McVay, 2010), each of which can potentially influence score accuracy. Rater severity/leniency is the tendency of raters to assign consistently either higher or lower scores to examinees than what they should have based on the performance. Rater extremity/centrality indicates the deviation or lack thereof of raters' scores from the middle score of a rating scale. Rater accuracy/inaccuracy refers to the distance between raters-assigned scores and an ideal score that an examinee should have received. The halo effect occurs when raters give similar ratings to an examinee across different tasks regardless of their actual performance.

Research has identified multiple factors that can contribute to differences in rater behaviors, including rater attributes (e.g., rater's conscientiousness, level of self-monitoring) (Tziner et al., 2005). In addition, different training backgrounds and experiences also have shown to contribute to impactful rater effects on scoring (Athey & Mclytyre, 1987; Borman, 1979; Leckie & Baird, 2011; Wolfe et al., 2010). In the context of IRT equating, scores that are

obscured by rater effects are likely to affect estimation of examinees' ability and item parameters, which may further threaten the accuracy of equating results.

It is not unusual to notice that rater behaviors are subject to change across time points (Casabianca et al., 2017; Harik et al., 2009; Lunz & O'Neill 1997; Myford &Wolfe, 2009), which may also influence equating accuracy (Boyer & Patz, 2019). This lack of stability of raters, referred to as differential rater functioning over time (DRIFT, McLaughlin et al., 2009; Wolfe et al., 1999), has been identified in several studies. For example, Congdon and McQueen (2000) investigated the consistency of rater severity using the elementary school writing performance data with a many-facet Rasch model (MFRM, Linacre, 1989; Wang & Wilson, 2005). Changes in rater severity was observed not only daily but also within each day. In addition, Leckie and Biard (2011) applied multilevel modeling to investigate rater performance using data from England's 2008 national curriculum English writing test. Although they found the average severity among raters was stable across time, differences in rating severity were noted between individual raters. Myford and Wolfe (2009) explored the accuracy and consistency in applying a rubric among raters over time, observing that a few raters made statistically significant changes in their levels of accuracy across time. Moreover, some raters tended to use different scale categories over time. Researchers also noticed increased centrality as the scoring process moves forward (Leckie & Biard, 2011; Myford & Wolfe, 2009). Therefore, researchers should consider the impacts of DRIFT on the accuracy of test equating. Changes in ratings across multiple administrations of a test are likely to undermine score comparability if not being properly accounted for.

Rater Models

Multiple rater models have been proposed and validated in previous research to detect and quantify rater patterns (Harik et al., 2009; Jin & Eckes, 2022; Leckie & Baird, 2011; Linacre, 1989; Patz et al., 2002; Raymond & Viswesvaran, 1993; Wang et al., 2014), such as MFRM (Linacre, 1989), generalized rater model (GRM, Wang, et al., 2014), and HRM (Patz et al., 2002). MFRM decomposes ratings into joint effects of persons, items, as well as raters on a logit scale. This model assumes that ratings are locally independent. However, this assumption is likely to be violated in practice. GRM (Wang et al., 2014) was developed as an extension of MFRM with an intention to address the local dependence problem by adding random-effect parameters to account for person-rater and person-item interactions (Robitzsch & Steinfeld, 2018).

Unlike the facet-models that directly connect scores assigned by raters with an examinee ability, HRM and its extended versions (Casabianca et al., 2016; Casabianca, 2017; DeCarlo et al., 2011; Nieto & Casabianca, 2019) use a hierarchical structure to model rating data. In the first level of HRM, raters' scores, referred as observed ratings, are perceived as the indicators of an ideal rating, which can be modeled through a signal detection process. The ideal ratings depend solely on scoring rubrics without any rater bias or variability. In the second level, the ideal ratings are considered as the categorical indicators of an examinees' true ability, which can be represented by an IRT polytomous model. In HRM, it accounts for the person-item dependence issue by introducing ideal ratings, capturing the dependence of multiple ratings on the same item (Patz et al., 2002; Robitzsch & Steinfeld, 2018). Wang et al. (2014) argued that HRM is more suitable when specific rating rubrics are available, whereas facet-models are preferred when raters are independent experts who assign scores based on their own knowledge and judgment. Given that the equating method proposed in this study is intended to be used in large-scale

assessments where raters are usually provided with standardized scoring rubrics, HRM was viewed more appropriate over facet-models. In addition, HRM is more flexible in parameter estimation because it allows the use of an existing IRT polytomous model, such as a partial credited model (Masters, 1982), a generalized partial credit model (GPCM, Muraki, 1992), and a graded response model (Samejima, 1969).

The performance of HRM and GPCM was recently examined by Song and Lee (2022), where single and double ratings were both available for constructed response (CR) tests. HRM consistently demonstrated more accurate results as compared to GPCM when rater errors were present in data. Double scoring was observed to lead to reduced errors in proficiency estimates than single scoring. Furthermore, the accuracy of proficiency estimations improved with an increased number of items and score categories.

Rater Effects on Score Comparability

Trend scoring (Tate, 1999 & 2000) was proposed to mitigate the effect of rater errors on test score comparability. Specifically, within a common item non-equivalent group (CINEG) design, the common items completed by the old-form group of examinees are scored by both the old and the new sets of raters, such that the severity of the new-form rater group can be adjusted based on the old-form rater group. Therefore, DRIFT across multiple forms of a test is likely to be reduced before conducting the equating procedure. Kim et al. (2010) examined different designs for equating CR tests using classical equating methods. It was observed that smaller bias can be obtained by applying the trend scoring method for common CR items under the CINEG design or CR items under the random groups design as compared to other CINEG designs using only CR items or external MC items as common items. However, it is worth mentioning that the adjustment with trend scoring heavily relies on the score quality of the old-form rater group. In

fact, there is no guarantee that scores assigned by the old-form rater group are reliable or consistent, and therefore, more research needs to be done to find a good solution to rater drift.

Boyer and Patz (2019) explored the impact of rater drift on two forms of a mixed-format test by varying the levels of rater bias and variability using data simulated based on HRM. Specifically, the Stocking-Lord (1983) method was used to adjust group differences across test forms under the CINEG design. Rater drift was found to have negative impacts on comparability of test scores. Higher leniency tends to result in more errors of scale linking than higher severity. Also, introducing more bias to the rating data seems to have a larger impact on scores than that of increasing variability.

Notably, there is an inconsistent use of the term *equating* in the current measurement literature. This current study adheres to the terminology used by Kolen and Brennan (2014). In practice, it is possible that conducting scale linking, namely ensuring the two sets of parameters on a common scale, is sufficient if testing programs intend to generate reported scores in some forms of IRT theta scores (θ). While theta scoring is prevalent, number-correct scoring based on raw-scores is also often considered. If number-correct scoring is to be used in practice, one additional step is conducted after scale transformation, which is here referred to as *equating*. The focus of this study is to apply HRM to an IRT observed-score equating context.

Research Purposes

Previous research has not attempted to use HRM to account for equating errors associated with rater effects. The main purpose of the current study is to fill the gap in current literature by introducing an IRT observed-score equating method that accounts for rater errors using HRM. The specific research objectives of the study are as follows:

Research Purpose 1: To propose an IRT observed-score equating method under the HRM framework to account for rater errors; and

Research Purpose 2: To compare the accuracy of equating results obtained from the newly proposed method to a traditional IRT observed-score equating method that uses GPCM for demonstrating the effectiveness of the proposed method.

The newly developed equating method was examined under a set of simulation conditions to identify condition(s) in which the new method is preferred. The next section describes the procedure to conduct IRT observed-score equating with HRM (here referred to as HRM observed-score equating) followed by the simulation process and evaluation criteria.

HRM Observed-Score Equating

Conducting the HRM observed-score equating method involves seven steps. Step A is to estimate item and person parameters using HRM. In this study, the GPCM (Muraki, 1992) is used as the IRT polytomous model because the discrimination parameter can be captured in the model in addition to the difficulty parameter.

$$\left(\xi_{ij} = \xi \middle| \theta_i, \alpha_j, \beta_j, \gamma_{j\xi}\right) = \frac{\exp\left\{\sum_{k=1}^{\xi} \alpha_j(\theta_i - \beta_j - \gamma_{jk})\right\}}{\sum_{h=0}^{K-1} \exp\left\{\sum_{k=1}^{h} \alpha_j(\theta_i - \beta_j - \gamma_{jk})\right\}}, \tag{2.1}$$

where ξ_{ij} represents examinee *i*'s ideal rating on item j, θ_i is the examinee's ability, α_j indicates the item discrimination, β_j represents the item difficulty, γ_{jk} is the item step parameter for category k of the item, and *K* is the number of rating categories.

The signal detection process modeled by HRM is a matrix of rating probabilities. A single rater *r* has a group of probabilities of assigning a score level, observed ratings X_{ijr}, to examinee i on an item j based on ideal ratings ξ_{ij}. Table 2.1 describes five score levels to be

assigned to examinees on an item, conditioning on that the probabilities in the five cells on each row must sum to one (Patz et al., 2002).

Table 2.1

The Matrix of Rating Probabilities Indicating the Signal Detection Process Modeled in the HRM

Ideal Rating (ξ)	Observed Rating (k)				
	0	1	2	3	4
0	P_{00r}	P_{01r}	P_{02r}	P_{03r}	P_{04r}
1	P_{10r}	P_{11r}	P_{12r}	P_{13r}	P_{14r}
2	P_{20r}	P_{21r}	P_{22r}	P_{23r}	P_{24r}
3	P_{30r}	P_{31r}	P_{32r}	P_{33r}	P_{34r}
4	P_{40r}	P_{41r}	P_{42r}	P_{43r}	P_{44r}

Note. $P_{\xi kr}$ ≡ P [Rater *r* rates *k* | ideal rating ξ] in each row of this matrix. From "The hierarchical rater model for rated test items and its application to large-scale educational assessment data," by R. J. Patz, B. W. Junker, M. S. Johnson, & L. T. Mariano, 2002, *Journal of Educational and Behavioral Statistics*, *27*(4), p. 349.

As Patz et al. (2002) introduced, a discrete unimodal distribution is used for each row in Table 1 to link observed ratings and ideal ratings. Each row of the matrix probabilities is expected be directly proportional to a normal density with a mean of $\xi_{ij} + \phi_r$ and a standard deviation of ψ_r. Rater bias ϕ_r is the mode of this distribution and variability ψ_r represents the spread of the distribution, represented in Equation (2.1),

$$p_{\xi kr} = P\left(X_{ijr} = k \middle| \xi_{ij} = \xi\right) \propto \exp\left\{-\frac{1}{2\psi_r^2}[k - (\xi + \phi_r)]^2\right\}. \quad i = 1, \ldots N;\ j = 1, \ldots, J;\ r = 1, \ldots, R. \tag{2.2}$$

In Equation (2.2), *k* is the observed rating category, *N* represents the number of examinees, *J* is the number of items, and *R* is the number of raters. Within HRM, a rater behavior can be described with respect to bias ϕ_r and variability ψ_r. When ϕ_r= 0, it is likely for rater *r* to select a score level that matches the ideal ratings. In this case, no rater bias is supposed to exist. When ϕ_r < 0, rater *r* tends to rate examinee responses more severely, vice versa. With respect to the variability/unreliability of a rater, rater *r* tends to be more reliable as ψ_r becomes closer to 0.

Step B is to build an ideal-rating distribution for an examinee's true ability and an observed-rating distribution for an ideal rating using HRM. For step C, the conditional observed-rating distribution for an examinee's true ability on each item is represented by the joint distribution of these two distributions.

Step D is to obtain a conditional observed-score distribution over all items on a test form for an ability point based on the conditional observed-rating distribution using the Hanson's algorithm (1994), an extended version of Lord-Wingersky's algorithm (1984). For step E, the conditional observed-score distributions over a test can be multiplied by ability distributions of all examinees to generate a marginal observed-score distribution which involves an integration procedure. For simplicity, the integral was approximated by a summation of the probability for each examinee obtaining the score point x, which then was divided by the number of examinees, as shown in Equation (2.3) (Kolen & Brennan, 2014).

$$f(x) = \frac{1}{N}\sum_i f(x|\theta_i), \tag{2.3}$$

where *N* is the total number of examinees.

Under the HRM framework, the marginal observed-score distribution over a test form exists for each rater; therefore, multiple marginal observed-score distributions are generated for a test form. Step F is to average the marginal distributions by the number of raters for each form. Step G is to conduct equipercentile equating to find the final equating relationship based on the averaged marginal observed-score distributions for the parallel forms.

The notable difference from the traditional IRT observed-score equating method is that the two layers of a distribution jointly constitute the conditional distribution. This is because in HRM, the observed rating is viewed as the function of a latent ability θ as well as a rater effect (bias and variability). To model the rater effect, one more step is added, which makes the

equating process relatively more complicated (one additional integration process is needed). However, if there is sufficient evidence of a non-negligible amount of rater effects, the increased complexity in model estimation and computation will be more than offset by the expected increased accuracy.

Method

Simulation Process

Given that this study mainly focuses on illustrating the new equating model, a random group design is used to avoid additional complexity associated with CINEG as scale linking is additionally necessary for this design. The proposed method is compatible with other data collection designs (i.e., CINEG) in which equating can be implemented in a similar fashion given that parameters are properly transformed onto a common scale across forms.

A simulation study was carried out to examine the effectiveness of the HRM observed-score equating method. Two parallel forms used for conducting equating were generated, including a new form (Form X) or an old form (Form Y). Three test forms were generated, including test forms with 1) eight five-category items, 2) twelve five-category items, and 3) eight seven-category items. The discrimination parameters were generated by randomly drawing from a lognormal distribution - $(0, 0.5^2)$, following the thresholds in Kang et al.'s (2009) study. For the five-category items, four location parameters were randomly sampled from $N(-1, 1)$, $N(-0.5, 1)$, $N(0.5, 1)$, and $N(1, 1)$. For the seven-category items, two more location parameters were added in the simulation, including $N(-1.5, 1)$ and $N(1.5, 1)$.

An incomplete rating design was used in this study, meaning that not all raters scored every response. Specifically, two out of six raters were randomly selected to score responses to each item from each examinee. Form Y was scored by normal raters, whose severity and

variability are within a reasonable range and known to have minimal impact on parameter estimates (Casabianca & Wolfe, 2017). For Form X, four different types of raters assigned scores, including 1) normal raters, 2) unreliable raters who tended to provide inconsistent ratings, 3) severe raters whose ratings were more severe and exceeded the normal range, and 4) severe and unreliable raters (here referred to as severe/unreliable raters) who gave inconsistent and tighter ratings. The distributions of the parameters associated with the four different types of raters were selected based on the review of pervious literature (Boyer & Patz, 2019; Song & Lee, 2022): normal raters ($\phi_r \sim U(-0.5, 0.5)$, $\psi_r \sim U(0.3, 0.5)$), unreliable raters ($\phi_r \sim U(-0.5, 0.5)$, $\psi_r \sim U(0.5, 1.0)$), severe raters ($\phi_r \sim U(-1.0, -0.5)$, $\psi_r \sim U(0.3, 0.5)$), and severe/unreliable raters ($\phi_r \sim U(-1.0, -0.5)$, $\psi_r \sim U(0.5, 1.0)$). The level of rater bias and variability varied in this study to examine the ability of the newly proposed equating method in handling different types/amounts of rater errors.

In total, 2,000 examinees were simulated based on a normal distribution $N(0, 1)$, and randomly allocated to take either Form X or Form Y to ensure the equivalence of the two groups. Hence, each test form was administered to 1,000 examinees. The HRM rating data were simulated using R version 4.1.0 (R Core Team, 2021).

The traditional IRT observed-score equating with GPCM was considered as the baseline method for better understanding the performance of the newly introduced method. Similar to previous studies (Casabianca & Wolfe, 2017; Song & Lee, 2022), parameters estimates were obtained through using the Markov Chain Monte Carlo (MCMC) method for both models. JAGS (Plummer, 2003), a computer program for analyzing Bayesian models, was employed for parameter estimation by incorporating the R package R2Jags (Su & Yajima, 2021). Song and Lee (2022) summarized a set of prior distributions based on previous research, which were

employed in the current study. Specifically, the priors used to fit both models include discrimination parameter ~ lognormal (1.2, 1.44), difficulty parameter ~ $N(0, 6.25)$, and theta ~ $N(0,1)$. The priors for estimating rater parameters include $\phi_r \sim \mathrm{N}(0,10)$, $\log(\psi_r) \sim U(0,10)$. In sum, 11,000 MCMC iterations, 5,500 burn-in, 16 thinning and three chains were set in JAGS. As for the estimated distributions, the mean for each parameter distribution was used as the estimates.

Once HRM or GPCM parameters were obtained, IRT observed-score marginal distributions for both models were built through using R, which were then entered into an R package *equate* (Albano, 2016) to find the equating relationships. The generated equating estimates were compared against the evaluation criteria.

Evaluation Criteria

The equating relationships estimated based on HRM and GPCM were evaluated against a criterion equating relationship over 20 replications for each simulation condition. The criterion equating relationship was generated using a large-sample single group equipercentile equating, which not only minimizes the error related to the differences in examinees' abilities between the two forms but also reduces sampling error (Kim & Lee, 2016; Kim et al., 2020). The criterion equating relationship was found through the following steps. First, 100,000 examinees were simulated from a $N(0, 1)$ distribution. Then, all examinees were assumed to complete both Form X and Form Y, so responses to both forms were generated using HRM. Last, the criterion equating relationship was identified using the traditional equipercentile equating method.

Conditional results were evaluated at each score x based on three statistics. Specifically, bias, standard error (SE), and root mean squared error (RMSE), which are described as:

$$bias(x) = \frac{1}{M}\sum_{r=1}^{M}(\hat{e}_{xr} - e_x) \ , \tag{2.4}$$

$$SE(x) = \sqrt{\frac{1}{M}\sum_{r=1}^{M}(\hat{e}_{xr} - \frac{\sum_{r=1}^{M}\hat{e}_{xr}}{M})^2} \text{ , and} \quad (2.5)$$

$$RMSE(x) = \sqrt{Bias(x)^2 + SE(x)^2} \text{ ,} \quad (2.6)$$

where *M* is the number of replications, x is the raw-score point, $\hat{e}_{xr}$ represents an estimated equated score at raw score *x* in replication r, and e_x represents the criterion equated score at raw score *x*, correspondingly.

In addition, overall results were presented in terms of marginal statistics calculated over all raw-score points to understand the overall performance of each equating method. The overall/marginal bias (Obias) can be expressed as:

$$Obias = \sqrt{\sum_x h(x)[\frac{1}{M}\sum_{r=1}^{M}(\hat{e}_{xr} - e_x)]^2} \text{ ,} \quad (2.7)$$

where $h(x)$ refers to the relative frequency distribution for the scores on Form X for the population data. The overall SE (OSE) and overall RMSE (ORMSE) were obtained in a similar fashion. The smaller the calculated statistics, the more precise the equating method.

Results

Conditional Results

Conditional results, in terms of SE, bias, and RMSE, of the two IRT observed-score equating methods are provided in Figures 2.1 through 2.12. Specifically, Figures 2.1 to 2.4 provide the values of conditional SE for normal raters, unreliable raters, severe raters, and severe/unreliable raters, respectively. Similarly, Figures 2.5 - 2.8 and 2.9 - 2.12 present the equating errors in terms of conditional bias and RMSE, respectively. In each figure, conditional results for three test conditions are presented. The horizontal axis indicates the raw score points whereas the vertical axis represents the equating errors.

Conditional SE. As seen in Figures 2.1 to 2.4, both equating methods lead to a small amount of SEs over the entire score range, with the largest below 0.08. The shapes of SEs

associated with both equating methods, in general, follow bell-shaped distributions. For each rater condition, the magnitudes of SEs become slightly larger for test forms with an increased number of items or score categories. The difference in SEs between the two equating methods are not substantial. HRM produces negligible less bias in the lower and upper ends of the scale but leads to larger SEs in the middle range.

Conditional Bias. When rater errors are within a normal range, both methods lead to relatively accurate results in terms of bias, as seen in Figure 2.5. Small differences are observed between the two methods when test forms consisted of eight five-category items and twelve five-category items. Under these conditions, HRM seems to be associated with a slightly larger amount of bias than those from the baseline method in the middle range of the score scales. For the eight seven-category condition, the two equating methods result in an almost identical amount of bias over most of the score scale.

When the new forms were scored by aberrant raters, HRM yields significantly reduced bias, which is different from the pattern observed under the normal rater condition. Specifically, with the presence of unreliable raters, HRM is associated with more accurate results than the baseline method across at least two/thirds of the scales regardless of test length and score categories, as seen in Figure 2.6. Fluctuations are still identified in the bias for HRM at the lower end and the mid-range of the score scale, but the magnitude of bias is significantly smaller than that observed for GPCM. For the upper one/third of the scales, the new method leads to smaller bias whereas the baseline method produces larger negative bias. Based on Figures 2.7 and 2.8, the new method seems to produce only a small amount of bias across the score scales when either severe raters or severe/unreliable raters present. However, the baseline method tends to produce a greater amount of bias under these two rater conditions. This indicates that the

proposed equating method is more robust to the presence of rater bias. Another interesting finding is that the baseline method consistently generates more bias for a longer test when aberrant raters scored the new forms. However, this pattern is not identified in the bias for HRM, suggesting its robustness to the changes in test length.

Conditional RMSE. Under normal rater condition, the two IRT observed-score equating methods lead to relatively accurate values in RMSEs, consistent with what was observed in conditional bias. The magnitudes of RMSEs for both methods are within 0.12. However, when the new forms are scored by unreliable raters, using HRM can reduce a great amount of systematic error at the upper third of score scales as compared to GPCM, as displayed in Figure 2.10. Although there are fluctuations in the RMSEs at the lower two-thirds of score scales for HRM, the magnitude is much smaller than those of GPCM. According to Figures 2.11 and 2.12, HRM shows superior performance in terms of RMSEs relative to GPCM for the majority of the score scales when severe raters or severe/unreliable raters scored the new forms. It is worth noting that under these two rater conditions, the discrepancy in RMSE between the two methods is larger than that under the unreliable rater condition. Consistent with conditional bias, the baseline method produces more RMSEs for the forms consisting of twelve five-category items with the presence of aberrant raters, which is not found for HRM.

Figure 2.1

Conditional SE as Normal Raters Assigning Scores on New Forms

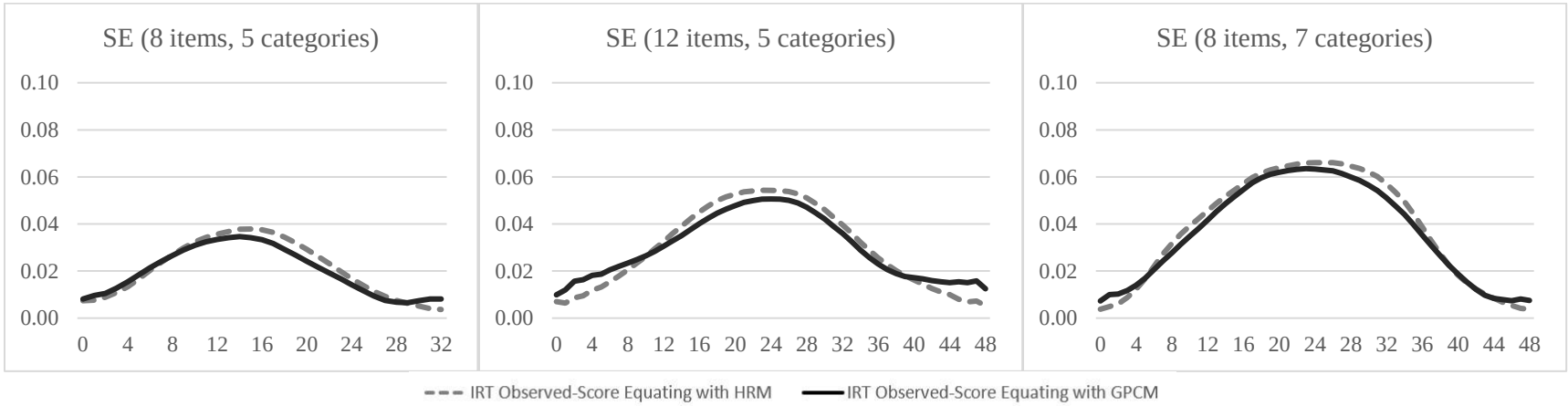


Figure 2.2

Conditional SE as Unreliable Raters Assigning Scores on New Forms

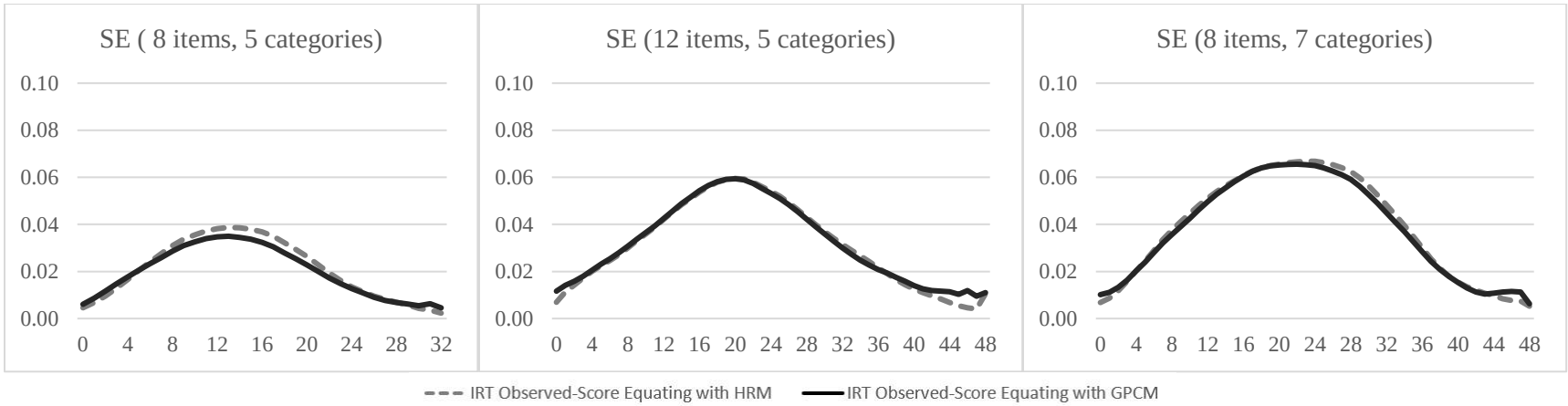


Figure 2.3

Conditional SE as Severe Raters Assigning Scores on New Forms

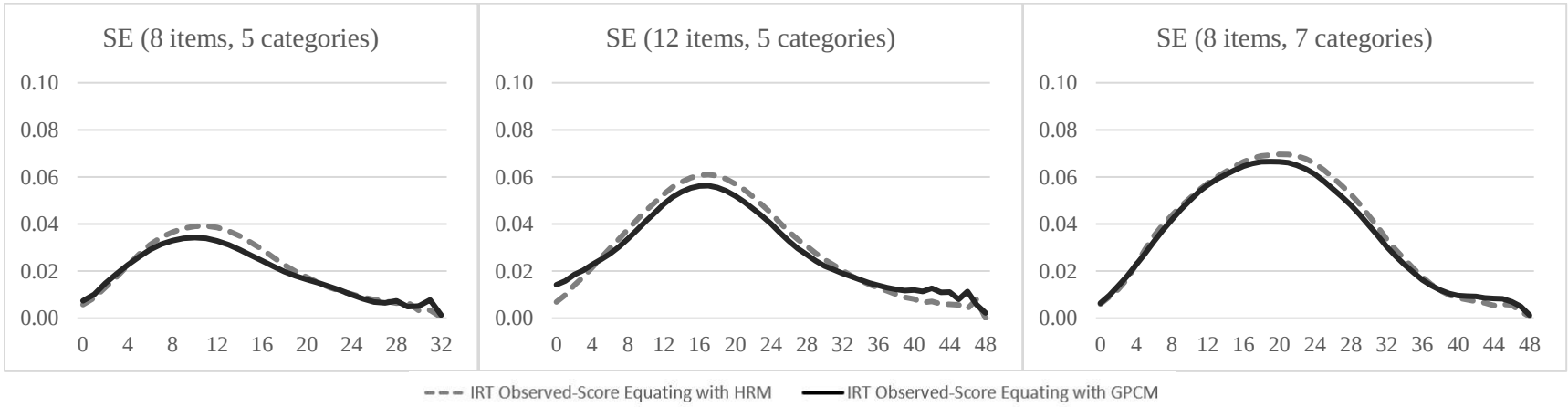


Figure 2.4

Conditional SE as Severe/Unreliable Raters Assigning Scores on New Forms

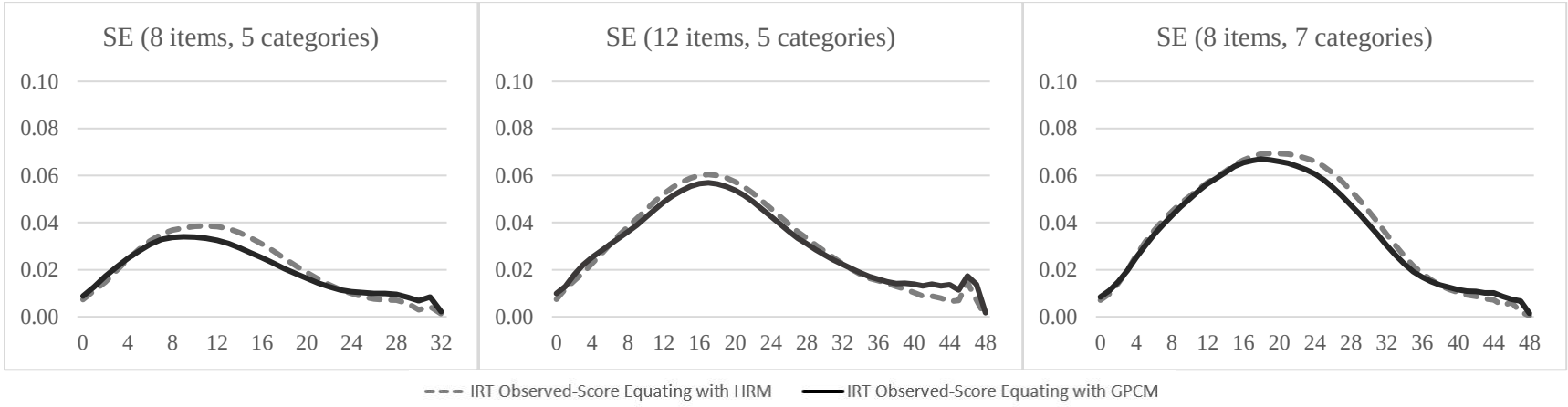


Figure 2.5

Conditional Bias as Normal Raters Assigning Scores on New Forms

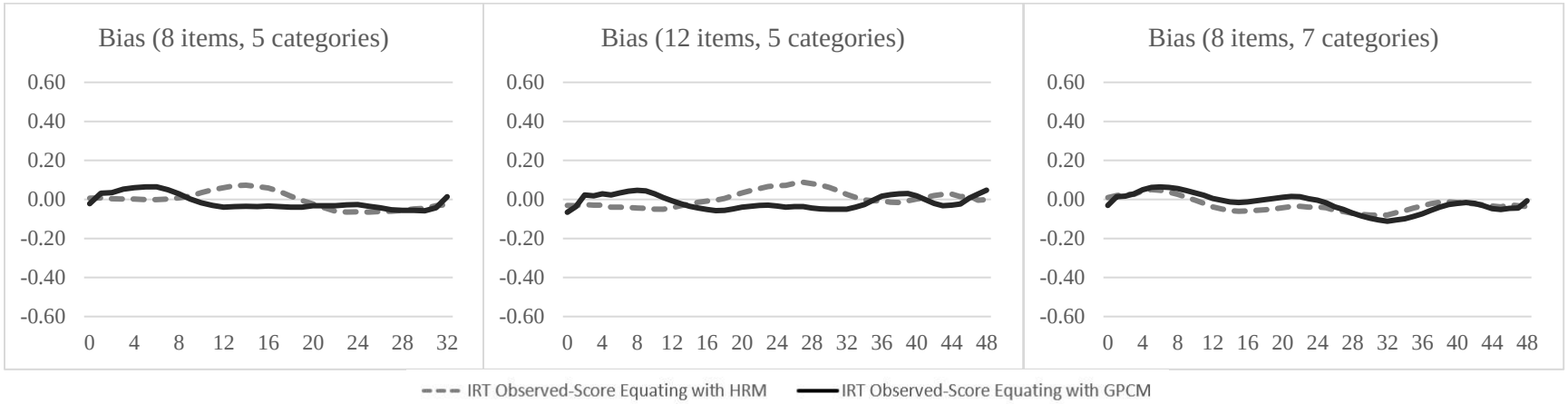


Figure 2.6

Conditional Bias as Unreliable Raters Assigning Scores on New Forms

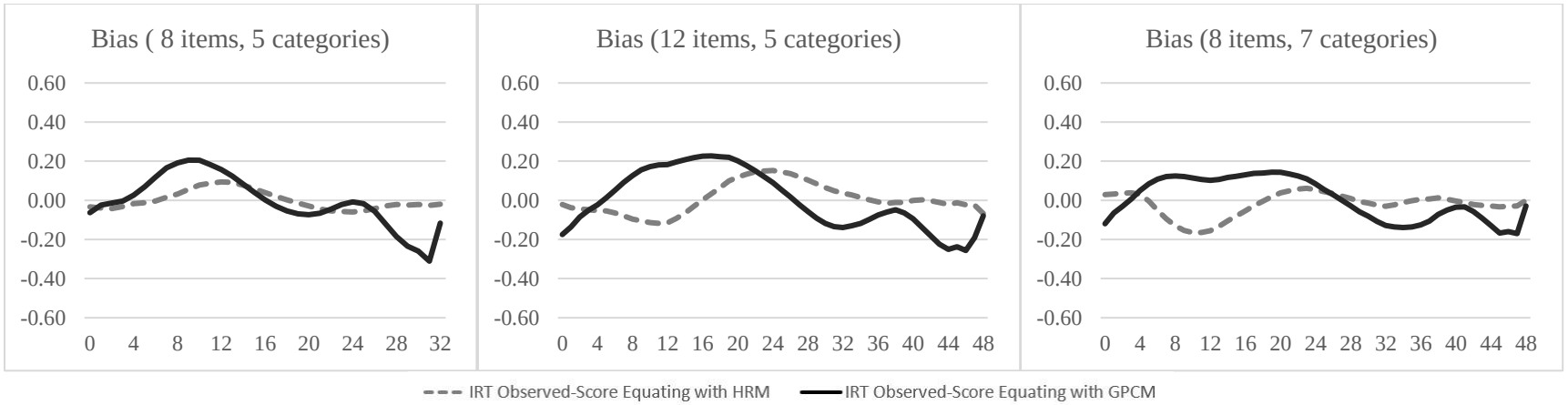


Figure 2.7

Conditional Bias as Severe Raters Assigning Scores on New Forms

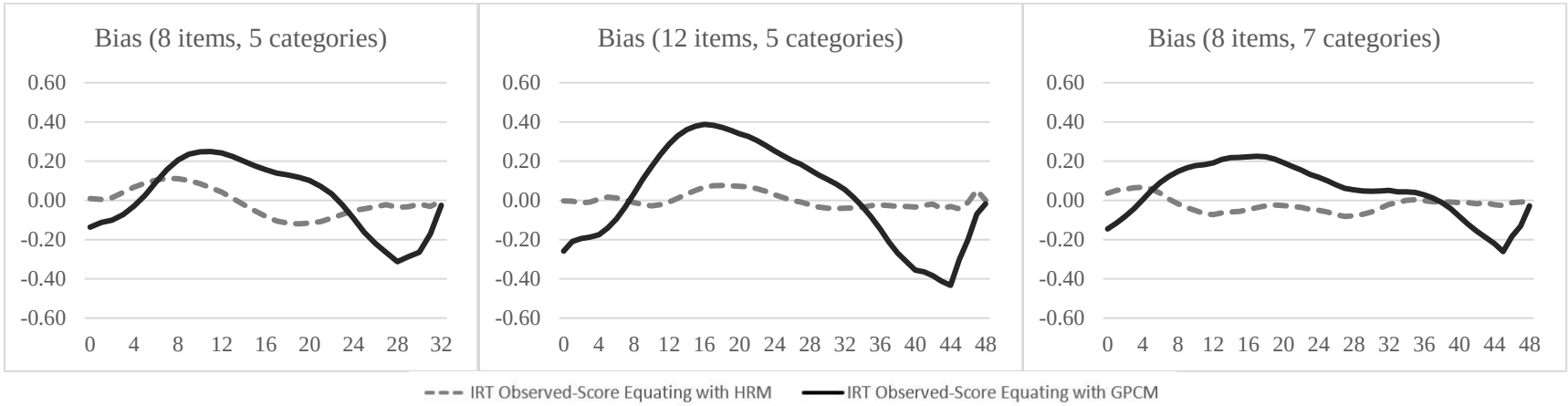


Figure 2.8

Conditional Bias as Severe/Unreliable Raters Assigning Scores on New Forms

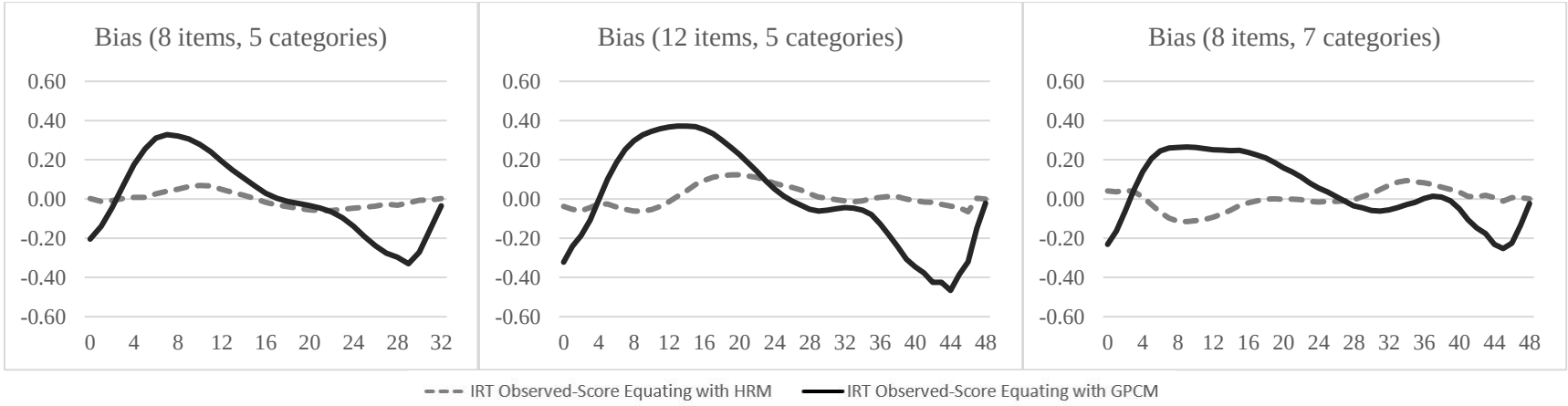


Figure 2.9

Conditional RMSE as Normal Raters Assigning Scores on New Forms

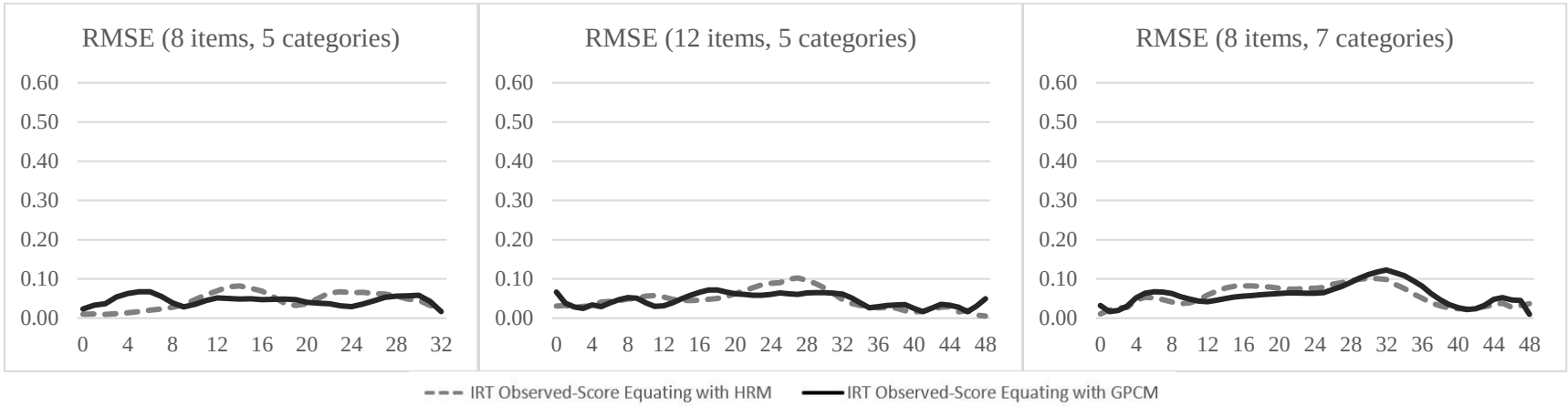


Figure 2.10

Conditional RMSE as Unreliable Raters Assigning Scores on New Forms

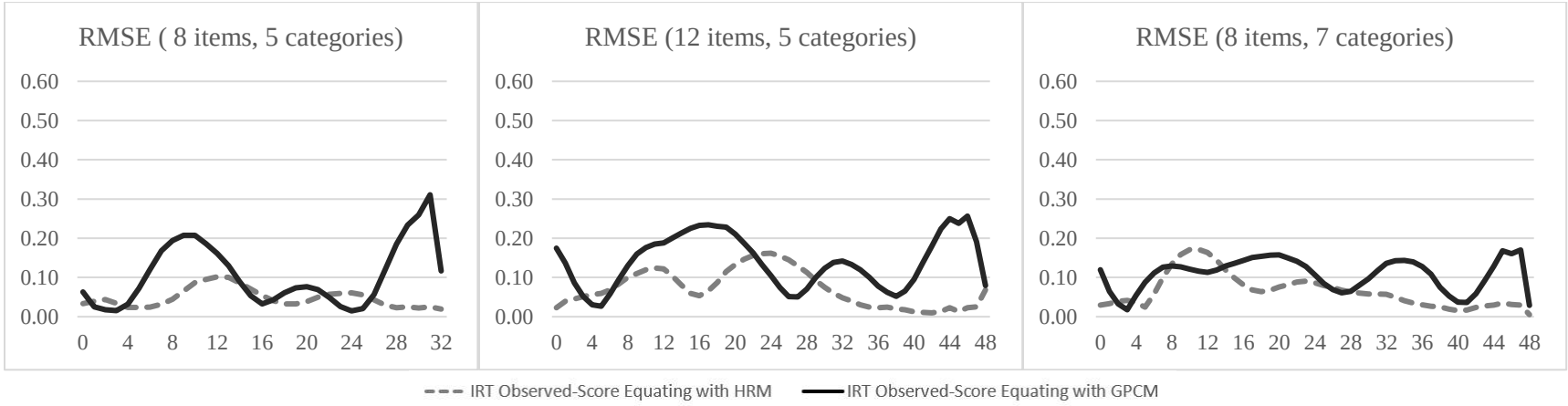


Figure 2.11

Conditional RMSE as Severe Raters Assigning Scores on New Forms

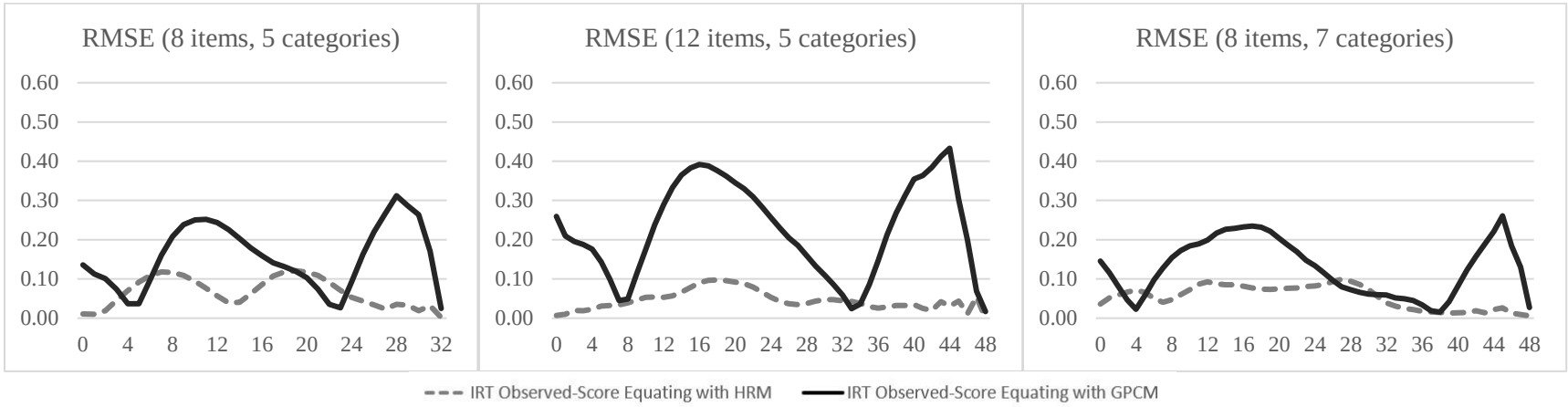
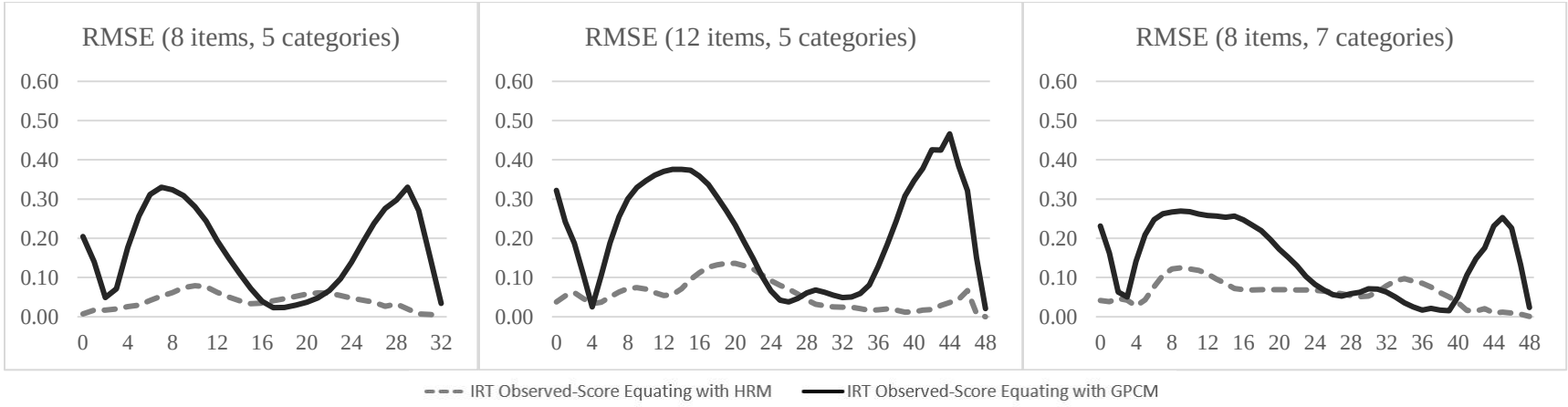


Figure 2.12

Conditional RMSE as Severe/Unreliable Raters Assigning Scores on New Forms



Overall Results

Table 2.2 provides a summary of the overall results. Specifically, the equating errors, including Obias, OSE, and ORMSE, are aggregated over the three test conditions by four types of rater errors. In terms of averaged Obias, the two equating methods produced results that are almost identical when new forms are graded by normal raters. However, different patterns are discernable when new forms are scored by aberrant raters in which HRM consistently produces smaller averaged Obias. In terms of averaged OSE, the difference between the two equating methods is minimal, which is within 0.002. It is worth noting that GPCM consistently yields smaller values regardless of the levels of rater severity and/or variability. The patterns of ORMSEs are similar to those of averaged Obias under the conditions involving aberrant raters because Obias contributes to ORMSE to a greater extent than OSE does. With normal raters, HRM leads to marginally larger values in averaged ORMSEs as compared to GPCM. Overall, given the magnitudes of the averaged Obias and ORMSE, HRM demonstrates robustness with respect to the presence of different types of rater effects (severe or unreliable) as compared to the traditional equating method based on GPCM. However, introducing rater errors has a little to no impact on SEs for both equating methods.

Table 2.2

Summary Statistics of Equating Errors across Test Conditions

Raters Scoring New Forms	Averaged Obias		Averaged OSE		Averaged ORMSE	
	HRM	GPCM	HRM	GPCM	HRM	GPCM
Normal	0.043	0.043	0.033	0.031	0.055	0.053
Unreliable	0.062	0.115	0.034	0.033	0.071	0.120
Severe	0.052	0.166	0.034	0.032	0.064	0.170
Severe/Unreliable	0.051	0.185	0.034	0.032	0.061	0.188

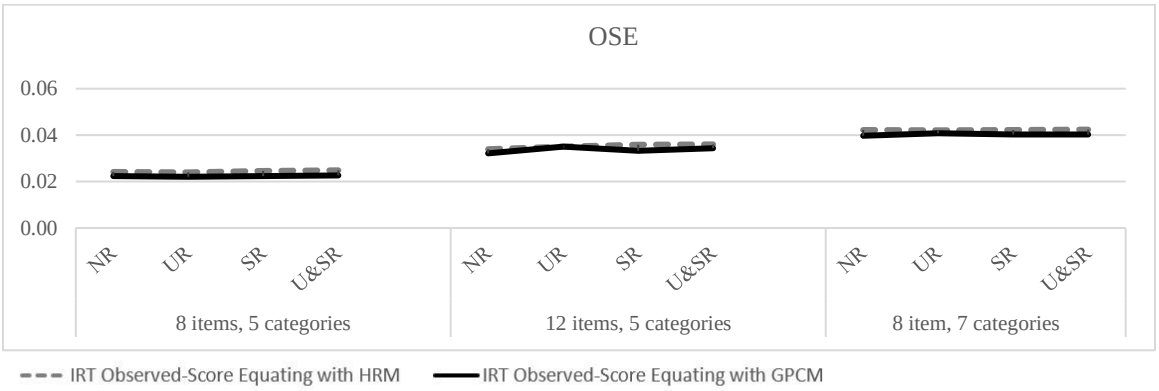
Figures 2.13 through 2.15 are provided for a deeper examination of the potential interaction of test length and the number of score category on equating accuracy. Specifically, Figures 2.13, 2.14, and 2.15 present the values of OSE, Obias, ORMSE, respectively, across the three test conditions, including the forms with eight five-category items, twelve five-category items, and eight seven-category items. In general, both equating methods lead to larger OSEs as the test becomes longer and as the number of score category increases, as seen in Figure 2.13. HRM consistently leads to marginally larger OSEs than GPCM.

The magnitude of Obias is similar between the two equating methods under the normal rater condition. Under the aberrant rater conditions, the increased accuracy associated with HRM is clear, which is invariant to the changes in test length and the number of score category.

An interesting finding identified in Figures 2.14 and 2.15 is that test length seems to have a larger impact on equating accuracy, in terms of Obias and ORMSE, for GPCM compared to HRM under certain rater conditions. For example, GPCM produces larger Obias for test forms with twelve items compared to those with eight items under the severe raters or severe/unreliable rater conditions. However, bias for HRM seems to be more consistent under these two rater conditions. Regarding the number of score category, it seems to have a little to no impact on Obias and ORMSE.

Figure 2.13

OSE over the Three Test Conditions



Note. NR, UR, SR, and U&SR refer to normal, unreliable, severe, and severe/unreliable raters, respectively.

Figure 2.14

Obias over the Three Test Conditions

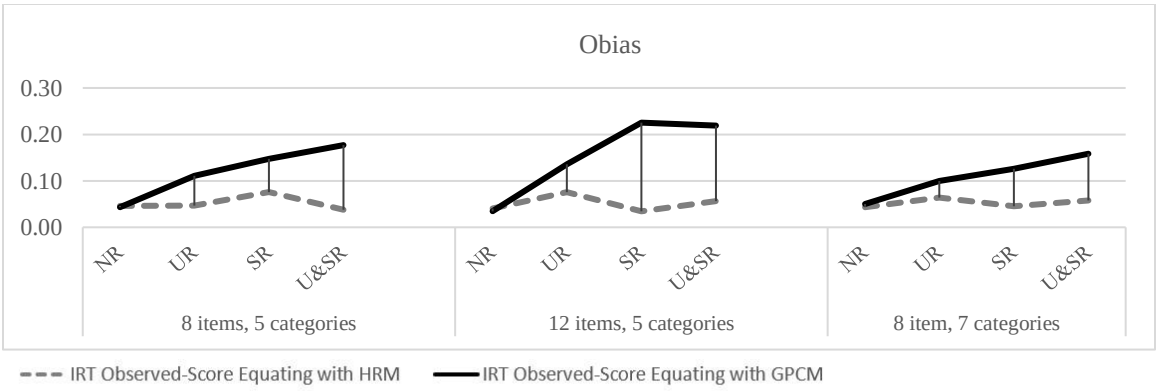
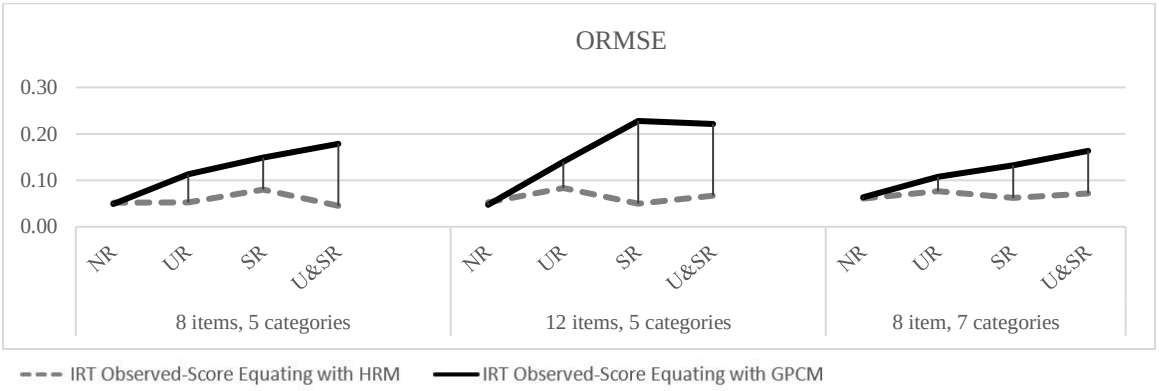


Figure 2.15

ORMSE over the Three Test Conditions



Discussion and Conclusion

Rater-mediated assessments that involve human raters in the scoring procedure inevitably introduces rater errors. The existing equating methods fail to account for rater errors in the equating process. The aim of this research is to address the gap in current literature by introducing an IRT observed-score equating method that employs HRM to account for rater errors and enhance test score comparability. The effectiveness of the newly suggested equating method is compared to that of the traditional IRT observed-score equating method with GPCM in a range of simulated scenarios using a random group design. Researchers and practitioners who are interested in maintaining comparability of high-stakes rater-mediated tests will benefit from the conclusions of the current study, which are presented below.

First, both equating methods demonstrated their robustness to a small rater effect by providing fairly accurate equating results. However, when aberrant raters were present, using the HRM equating method, in general, led to more accurate equating results of both bias and RMSE. Specifically, the equating errors from GPCM were largely impacted by the increase in rater errors whereas the proposed method tended to be more consistent in equating results regardless of the presence of unreliable or biased raters. This finding indicates that the use of HRM in the equating process enables errors caused by human raters to be accounted for such that equating accuracy can be maintained. Song and Lee (2022) noted that HRM tends to yield more precise ability estimates when rater effects are notably significant in ratings. By extending the use of HRM to the context of test equating, the current study identified a similar pattern. Boyer and Patz (2019) concluded that rater's severity (bias) seems to have a larger impact on test score comparability than rater's variability. This study confirmed the finding by observing significantly larger bias and RMSEs for the condition of severe raters than that of unreliable

raters when the traditional equating based on GPCM was implemented. However, this pattern was not found for HRM observed-score equating, indicating the new method effectively mitigates the impact introduced by severe raters. The effectiveness of the HRM equating method suggests that it is especially valuable for tests that include a large number of CR items, where there is a higher risk of high-impact rater error. In such cases, caution would require multiple ratings (e.g., Song & Lee, 2022), making the HRM equating method a particularly helpful tool to ensure the comparability of test scores.

Second, only a minor difference was found in SEs between the two equating methods. Although HRM was found to be associated with slightly more overall SEs, it produced smaller SEs near the ends of the score scale. Unlike bias and RMSE, SE was not impacted whether raters were normal or aberrant for both equating methods.

The examination of conditional results revealed that GPCM, under most of the conditions, tended to generate more errors at the ends of the score scale, in terms of the three evaluation criteria. This finding is in line with Tao and Cao (2016), who concluded that equating results are likely to fluctuate for score points with low frequency. The HRM observed-score equating method generally led to reduced bias and RMSE at both ends with only a few exceptions, suggesting that the new method can effectively reduce equating errors particularly for score ranges with limited observations by taking rater errors into account. Moreover, a larger discrepancy was found between the two methods in bias and RMSE when new forms were scored by severe raters or by severe/unreliable raters. Under these rater conditions, a much smaller proportion of examinees earned a maximum score on each item, such that information available for parameter estimation was limited. HRM could still generate relatively accurate results whereas the baseline method failed to do so.

In terms of the two simulation factors, more SEs were found to be associated with longer tests and an increased number of score categories for both equating methods. The two factors seem to have limited effects on bias and RMSE. As test length and score category increased, the errors associated with raters also increased, which is a potential reason for not obtaining more accurate results for longer tests with more score categories.

Limitations and Future Research

This study sets the stage for the development of a more sophisticated equating method that enables rater errors to be captured, ultimately enhancing test score comparability. As discussed earlier, each IRT rater model makes a different set of assumptions in regard to the relationship between persons, items, and raters. For some cases, the use of HRM may not adequately model the actual data structure and therefore not be justifiable. Future research could continue to explore the possibility of incorporating other rater models into the equating context for rater-mediated assessments. For example, GRM which models both the interactions between examinees and items, as well as between examinees and raters, can be applied to equating to address local dependence. Also, this study only suggested the observed-score equating but future research may consider, IRT true-score equating in conjunction with those rater-effect models (e.g, facet-models).

Several limitations should be noted before implementing the proposed equating method. First, the current study set the old forms to be scored by normal raters. In practices, however, the old forms can also be graded by aberrant raters, which are not considered in the study. Future work may investigate the impact of other rater conditions that also include a varying level of rater effects for the old form.

The sample size used in the study was fixed at 1,000. Although sample size was found to be less impactful than the number of items or score categories on ability estimation at certain conditions for HRM (Song & Lee, 2022), it will be important that future research investigates the impact of sample size in the context of test equating when rater effects are present. Moreover, other possible conditions of number of items and the number of score category can be examined to produce a more comprehensive picture of the performance of HRM observed-score equating. In this study, a marginal distribution of a test form was computed by averaging the multiple marginal distributions by the number of raters for each form. Raters are viewed as the source of measurement errors, so having a larger number of raters has the potential increase the accuracy of estimated marginal distributions for each form. Therefore, it is also worthy to investigate the impact on equating accuracy as more raters are involved in the scoring process.

Another major limitation has to do with defining a criterion equating relationship. It is worth noting that HRM was used for generating the data which might have offered unintended advantages to HRM observed-score equating over GPCM observed-score equating. Therefore, future research may use a different model in simulating data or simulate data using rater parameters based on rater statistics (i.e., quadratic weighted kappa) obtained from real data, as in the previous study (Boyer & Patz, 2019).

References

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1–36.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49-58.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, *72*(4), 567-572.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*(4), 410-421.

Boyer, M., & Patz, J. P. (2019, April). *Understanding and mitigating rater drift: When does rater drift threaten score comparability?.* Paper presented at the Annual meeting of the National Council on Measurement in Education, Toronto, Ontario.

Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. In *Handbook of Item Response Theory, Volume One* (pp. 477-494). Chapman and Hall/CRC.

Casabianca, J. M., Junker, B. W., Nieto, R., & Bond, M. A. (2017). A hierarchical rater model for longitudinal data. *Multivariate Behavioral Research*, *52*(5), 576-592.

Casabianca, J. M., & Wolfe, E. W. (2017). The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model. *Psychological Test and Assessment Modeling*, *59*(4), 471-492.

Congdon, P. J. , & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163–178.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45.

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333-356.

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 261-288). Mahwah, NJ: Lawrence Erlbaum

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, *14*(1), 31-57.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*(2), 105-121.

Hanson, B. A. (1994). Extension of Lord-Wingersky algorithm to computing test score distributions for polytomous items. Retrieved from http://www.openirt.com/b-a-h/papers/note9401.pdf

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58.

Jin, K. Y., & Eckes, T. (2022). Detecting differential rater functioning in severity and centrality: The dual DRF facets model. *Educational and Psychological Measurement, 82*(4), 757-781.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*(7), 499-518.

Kim, S. Y., Lee, W. C., & Kolen, M. J. (2020). Simple-structure multidimensional item response theory equating for multidimensional tests. *Educational and Psychological Measurement*, *80*(1), 91-125.

Kim, S. Y., Lee, W. (2016). Composition of common items for equating with mixed-format tests. In Kolen, M. J., Lee, W.-C. (Eds.), Mixed-format tests: Psychometric properties with a primary focus on equating (Vol. 4; CASMA Monograph No. 2.4; pp. 7-46). Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, 1-11.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399-418.

Lee, H. K. (2009). Native and nonnative rater behavior in grading Korean students' English essays. *Asia Pacific Education Review*, *10*(3), 387-397.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Liu, C., & Kolen, M. J. (2011). A comparison among IRT equating methods and traditional equating methods for mixed-format tests. In M. J. Kolen & W. Lee (Eds.), Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1). (CASMA Monograph Number 2.1) (pp. 75–94). Iowa City, IA: CASMA, The University of Iowa.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score" equatings". *Applied Psychological Measurement*, *8*(4), 453-461.

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity*, *15*, 13-25.

Lunz, M. E., & O'Neill, T.R. (1997, March). *A longitudinal study of judge leniency and consistency*. Paper presented at the Annual meeting of the American Educational Research Association, Chicago, IL.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika 47*, 149–174.

McLaughlin, K., Ainslie, M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Medical Education*, *43*(10), 989-992.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement 16*, 159–176.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389.

Nieto, R., & Casabianca, J. M. (2019). Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response tests. *Journal of Educational Measurement*, *56*(3), 547-581.

Patz, R.J., Junker, B.W., Johnson, M.S., Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341-384.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, *8*(2), 137-156.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*, *30*(3), 253-268.

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling*, *60*(1), 101-138.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956-970.

Song, Y. A., & Lee, W. C. (2022). Effects of Using Double Ratings as Item Scores on IRT Proficiency Estimation. *Applied Measurement in Education*, 1-21.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201-210.

Su, Y. S., & Yajima, M. (2021). Package 'R2jags': Using R to Run "JAGS" (2021). Retrieved from https://cran.r-project.org/web/packages/R2jags/R2jags.pdf

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal Educational Measurement*, 36, 336-346.

Tate, R. L. (2000). Performance of a proposed method for linking of mixed format tests with constructed response and multiple choice items. *Journal Educational Measurement*, 37, 329-346.

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education*, *29*(2), 108-121.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and rater factors affecting rating behavior. *Group & Organization Management*, *30*(1), 89-98.

Wang, J., & Engelhard Jr, G. (2019). Conceptualizing rater judgments and rating processes for rater-mediated assessments. *Journal of Educational Measurement*, *56*(3), 582-609.

Wang, W. C., Su, C. M., & Qiu, X. L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement*, *51*(3), 260-280.

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296-318

Wexley, K. N., & Youtz, M. A. (1985). Rater beliefs about others: Their effects on rating errors and rater accuracy. *Journal of Occupational Psychology*, *58*(4), 265-275.

Wind, S. A. (2019). Nonparametric evidence of validity, reliability, and fairness for rater-mediated assessments: An illustration using Mokken scale analysis. *Journal of Educational Measurement*, *56*(3), 478-504.

Wolfe, E. W. (2020). Human scoring with automated scoring in mind. In *Handbook of Automated Scoring* (pp. 49-68). Chapman and Hall/CRC.

Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training context*. Pearson Assessments. Retrieved from https://www.researchgate.net/profile/Edward-Wolfe-2/publication/242785513_Rater_Effects_as_a_Function_of_Rater_Training_Context/links/00b7d52d431a19b61a000000/Rater-Effects-as-a-Function-of-Rater-Training-Context.pdf

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (1999). *Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model*. Paper presented at the Annual meeting of the American Educational Research Association, Montreal, Canada.

CHAPTER/ STUDY 3: EVALUATING COVARIATE BALANCE DIAGNOSTICS FOR PROPENSITY SCORE WEIGHTING WITH MULTILEVEL DATA

Introduction

Propensity score (PS) methods have long been employed by educational researchers and program evaluators to identify students' propensity for participating in educational programs, which is a crucial factor in achieving balanced covariates and detecting causal-effect relationships (e.g., Melguizo et al., 2011; Morgan et al., 2010; Yamada et al., 2018; Yamada & Bryk, 2016). A propensity score represents the probability of receiving a treatment assignment based on measured covariates (Rosenbaum & Rubin, 1983; 1984). It is assumed that, conditional on a true PS, if the covariates in treatment and control groups are equivalent, then the treatment will not be confounded with covariates (Rosenbaum & Rubin, 1983). Confirmation of group equivalence is predicated upon the researcher's ability to properly test for covariate balance; however, existing tests are rudimentary. Educational research designs now extensively utilize multilevel structures to account for the inherent nesting of the education delivery model (Raudenbush & Bryk, 2002), and there is a pressing need for advanced covariate balance testing to be used for these complex data structures, as only a few currently exist.

Matching based on PS is one of the most widely used PS methods to reduce selection bias. Thoemmes and Kim (2011) conducted a systematic review of social science studies that employed PS approaches and found that matching was used more often (n = 55, 64%) than three other PS methods: PS weighting, stratification, and covariate adjustment. However, a large sample size is often needed for conducting PS matching if used in conjunction with a specific matching strategy (Guo & Fraser, 2015). This may not be realistic for studies evaluating interventions targeted primarily at small groups of participants or studies with limited funding

for data collection. Moreover, when conducting PS matching, there is a potential to discard some part of the sample either in the control or treatment group (Austin, 2014; Guo & Fraser, 2015; Olmos & Govindasamy, 2015), which can lower the validity of treatment effects.

Compared to matching, PS weighting has several advantages. It allows researchers to use the full sample in control and treatment groups for outcome analysis, which helps to maintain statistical power in identifying a treatment effect (Guo & Fraser, 2015; Stone & Tang, 2013). Unlike PS matching, which requires constructing comparison groups, PS weighting is easier to implement because the estimated PS can be used directly (Leite et al., 2015). Furthermore, weighting methods can be used to produce estimates not only for the average treatment effect (ATE), but also for the average treatment effect on treated participants (ATT) with slight changes in equations (Schafer & Kang, 2008; West et al., 2014). Additionally, weighting can be performed using many available computer software, enabling researchers from various fields using different software to implement PS methods (Olmos & Govindasamy, 2015).

After applying a PS method to a sample, a crucial step in the analysis routine is to perform a covariate balance diagnostic to check the equivalence of the covariates between the two comparison groups. The covariate balance diagnostic is used to examine if the ignorability assumption of treatment assignment has been met (Bai & Clark, 2018; Rosenbaum & Rubin, 1983). Ignorability of treatment assignment is required for using PS methods, and it assumes that the assignment to treatment or control groups is independent of outcomes when accounting for covariates. Achieving balanced covariates is necessary for meeting the ignorability assumption, allowing one to confidently draw the conclusion that selection bias is reduced by applying PS (Bai & Clark, 2018). In addition, assessing covariate balance is important to detect whether the PS model has been adequately defined (Austin, 2011). If balance diagnostics indicate an

imbalance, evaluators would need to select a different PS model or reconsider the covariates to obtain more balanced covariates between the two groups.

Research on covariate balance diagnostics is mostly conducted in the context of PS matching (Ali et al., 2014; Austin, 2009; Burnett 2019; Jacovidis et al., 2017; Zhang et al., 2019) whereas only a small amount of research has examined how to assess covariate balance for PS weighting. In a systematic review of 29 studies that employed inverse probability of treatment weighting (IPTW) as the weighting approach, Austin and Stuart (2015) found that many studies failed to assess the balance of weighted samples in two comparison groups.

While PS weighting is a simpler approach than matching, it is vulnerable to the impact of misspecification of a PS model, leading to biased estimates (Kang & Schafer, 2007; Schafer and Kang, 2008). Thus, it is crucial to assess covariate balance for PS weighting to check the adequacy of model specification. Austin and Stuart (2015) proposed a formal set of covariate balance diagnostics that can be used for IPTW, but this study only discussed covariate balance diagnostics for single-level data. However, in educational contexts, data are often hierarchically structured, with students nested in classrooms, schools, and/or districts, which requires the use of multilevel models to estimate PS and treatment effects. Variations may exist in the selection mechanism as well as the level of treatment among clusters (e.g., classrooms) and, to account for such complexity. Multilevel models have been used for estimating PS and treatment effects in previous literature. It is possible that the way to assess covariate balance for multilevel data is different from what has been used for single-level data.

There is a limited understanding about which measure leads to the most precise assessment and how to address the multilevel structure in evaluating covariates. The current study was carried out to investigate the performance of different measures for evaluating

covariate balance when PS weighting is used for multilevel data, with an aim to provide guidance to researchers and practitioners on selecting the most appropriate covariate balance diagnostics with complex data.

Covariate Balance Diagnostics

Various covariate balance diagnostics have been proposed to assess the extent to which the propensities of treatment and control groups are balanced, such as numerical diagnostics, graphical methods, and inferential tests (Burnett, 2019). Ali et al. (2014) found that the standardized bias (SB) method consistently outperformed the Kolmogorov-Smirnov distance, the Lévy distance, and the overlapping coefficient methods in the context of PS matching. The SB method is easy to compute, and it is not likely to be influenced by sample size (Ali et al., 2014; Ali et al., 2015). Specific to IPTW, Austin and Stuart (2015) suggested a set of defined procedures for assessing covariate balance diagnostics, one of which is to use a weighted SB method to compare means and assess the prevalence of covariates between comparison groups in weighted samples. Rather than examining the mean difference between the two samples, using variance ratio (VR) can assess the difference in the dispersion of the distribution of the covariates in the two comparison groups (Austin, 2009; Imai et al., 2008). A moderate difference of the covariates between the groups was found in the values of VR when the PS model was mis-specified, but when the PS model was correctly specified, the value of VR was close to 1, indicating that the two groups have similar variances.

Percent bias reduction (PBR) is another way to assess covariate balance (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1984). This statistical method reflects the degree to which the bias is reduced after applying the PS method. Although this method has been used in previous studies (e.g., Baser, 2006; Drake, 1993), limited research has focused on its

effectiveness. A few inferential tests such as t-tests and the Kolmogorov-Smirnov distance have been frequently used in previous research (Ali et al., 2015; Granger et al., 2020; Thoemmes & Kim, 2011). However, these tests should be used with caution as they are based on population parameters and the resulting estimates may be influenced by sample size (Ali et al., 2015; Ho et al., 2007).

Education researchers have employed various covariate balance diagnostics to ensure group equivalency when using PS methods on multilevel data (Arpino & Cannas, 2016; Arpino & Maelli, 2011; Leite et al., 2021; Rickles & Seltzer, 2014; Yamada et al., 2018). While some scholars have evaluated covariate balance with (weighted) standardized mean difference or a ratio of standard deviations using a pooled method after PS matching/weighting (Leite et al., 2015; McCormick et al., 2013), this method treats cluster-level covariates in the same way as individual-level covariates and ignores the clustering effect. To address this limitation, Arpino and Maelli (2011) proposed an additional step of averaging values of absolute standardized bias (ASB) to obtain an overall ASB between matched groups. Another approach used is the within-cluster method, where ASBs or other bias indices are calculated for individual-level covariates within each cluster separately (Kim & Seltzer, 2007; Yamada et al., 2018; Yamada & Bryk, 2016), and then each covariate from different clusters is averaged (Rickles & Seltzer, 2014).

The comparison of the pooled and within-cluster methods for evaluating covariate balance reveals that both approaches have their own strengths and weaknesses, as discussed by Burnett (2019). The pooled method provides a single summary statistic for each covariate that can be evaluated based on a criterion directly. However, this method does not offer sufficient information about within-cluster covariates. In contrast, the within-cluster method considers covariate balance within each cluster, but it may not be feasible when the cluster size is small.

Moreover, when there are many clusters in a dataset, computing covariate statistics within each cluster requires substantial effort. In addition, statistics for each covariate within each cluster cannot be used directly, and instead, they need to be averaged to obtain an overall statistic for each covariate that can be evaluated based on a criterion (Burnett, 2019).

To further examine these methods, Burnett (2019) conducted a simulation study to identify the optimal diagnostic for evaluating covariate balance with multilevel data after PS matching is applied, comparing pooled ASB, within-cluster ASB, pooled VR, and within-cluster VR. The results showed that pooled ASB had more accurate predictions for treatment effect bias, while within-cluster ASB was more sensitive to model specification. Although this study provided valuable information, it only focused on comparing balance diagnostics when PS matching was used. PS weighting has recently received increasing attention (Austin & Stuart, 2015), but there is no existing literature that has examined the performance of various diagnostics when PS weighting is used with multilevel data.

The current study aims to provide a comprehensive understanding of the effectiveness of several covariate balance diagnostics when PS weighting is used with hierarchical data. Six diagnostics were examined, including a) pooled ASB, b) within-cluster ASB, c) pooled VR, d) within-cluster VR, e) pooled PBR, and f) within-cluster PBR. The study uses IPTW because it is the most used weighting procedure for adjusting samples and has been demonstrated to remove more bias compared to alternatives such as marginal mean weighting (Leite et al., 2019). For IPTW, weights are assigned to individuals based on the inverse of their probability of receiving treatment, as estimated by PS (Rosenbaum & Rubin, 1983). Based on the study design in Burnett's (2019) research, this study expands the analysis to include the performance of six

diagnostics in the context of weighting. The study also investigates the correlations between the covariate balance diagnostics and the ATE bias under a set of conditions.

This study aims to find a diagnostic that is sufficient at identifying model misspecification. In addition, this study intends to find a diagnostic that can provide additional insight into the bias in the treatment effect. The following research questions will be addressed:

1. Among pooled ASB, within-cluster ASB, pooled VR, within-cluster VR, pooled PBR, and within-cluster PBR, which balance diagnostic performs best in detecting a correct PS model when using IPTW to work with multilevel data?
2. How different factors, including cluster size, ICCs of individual-level covariates, treatment prevalence, and covariate imbalance, affect the performance of each diagnostic?
3. Among the six covariate balance measures, which one demonstrates the strongest correlation with the bias in the treatment effect when utilizing IPTW with multilevel data?

Method

Simulation Factors

This Monte Carlo study used a simulation design with an intent to approximate the real-world scenarios by varying different factors. The choice of factors and levels was based on an examination of previous methodological research (Fuentes et al, 2021; Kush et al., 2022). In addition, the authors' previous research experiences in quasi-experimental studies guided the study design to produce contextualized implications of the findings.

Cluster size

Previous studies have highlighted the importance of the number of individuals within each cluster when dealing with multilevel data (Austin & Leckie, 2018; Leite et al., 2015; Moineddin et al., 2007). In this study, cluster size varied across three levels: 30, 50, and 100. The choice of the cluster size was based on an authors' previous research project in which we evaluated the effectiveness of an intervention in a higher education context (Westine et al., 2023). In that project, classes were considered as clusters, which often consist of 30 to 100 students. Thus, the number of clusters in this study was fixed at 30.

ICC of Individual-level Covariate X (ICC_X)

The intraclass correlation coefficient (*ICC*) measures the degree of clustering within groups. Previous research has demonstrated that the ICC of Individual-level Covariate X (ICC_X), had an impact on balance measures in identifying a PS model (Thoemmes & West, 2011). In this study, ICC_X varied at three levels: 0.1, 0.3, 0.5. With respect to the ICC of the treatment indicator in the PS model, it was fixed at 0.2 (Fuentes et al., 2021).

Treatment Prevalence

Treatment prevalence, representing the ratio of individuals assigned to the treatment group, was another factor manipulated in this study. Kush et al. (2022) found that the treatment effect estimation was more accurate when prevalence level was 0.5, compared to lower or higher treatment ratios. In this study, three levels of prevalence (0.2, 0.5, and 0.8) were simulated.

Baseline Imbalance

Baseline covariate balance is often checked before propensity score estimation to assess the degree of overlap between the two comparison groups. More bias and mean standard error in the estimation of treatment effect were found as the baseline imbalance increased (Kush et al.,

2022). This study used the average ASB of both the individual-level and cluster-level covariates to measure baseline imbalance. Two levels of baseline covariate balance, including 0.3-0.5 and 0.5-0.7, were examined. Since multiple conditions were simulated, a range, rather than a specific number, was assigned to each of the two levels to provide some flexibility for data generation. Table 3.1 summarizes the investigated factors.

Table 3.1

Factors and Levels of Simulation Study

Factors	Levels
Cluster Size	30, 50, 100
No. of clusters	30
ICC of the individual-level covariate X (ICC_X).	0.1, 0.3, 0.5
Baseline Imbalance	0.3-0.5, 0.5-0.7
Treatment Prevalence	0.2, 0.5, 0.8

Data Generation

The data simulation process followed the steps used in Burnett's (2019) study and is outlined in Figure 3.1. The steps for data generation include: (1) specifying a PS model for data generation, (2) simulating data for individual- and cluster-level covariates, (3) calculating the propensity score for each individual based on the specified model, (4) assigning individuals to the treatment or control group based on their propensity scores, and (5) generating outcome data based on a specified outcome model.

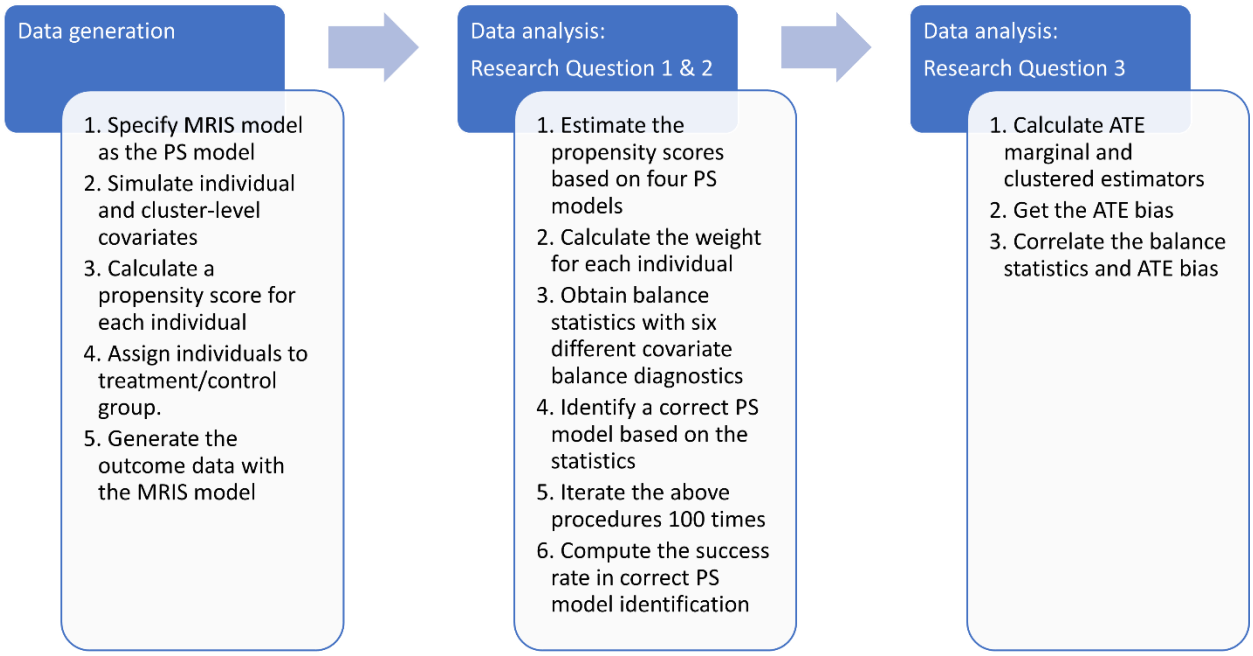
Specifically, propensity score was generated based on a multilevel logistic regression model with both random intercept and slope components (Fuentes et al., 2021; Leite et al., 2015), referred to as the MRIS model,

$$logit\left(P_{ij}(T_{ij}=1)\right)=\alpha_0+\alpha_X X_{ij}+\alpha_Z Z_j+\mu_{0j}+\mu_{1j}X_{ij}+\varepsilon_{ij}, \quad (3.1)$$

where T_{ij} is the binary value indicating if individual i in cluster j is assigned to the treatment group. P_{ij} is the probability of assignment to the treatment group. The intercept is represented by α_0, while α_X and α_Z are the fixed effects of individual-level covariate X_{ij} and cluster-level covariate Z_j, respectively. The random slope of individual-level covariate X_{ij} is denoted by μ_{1j}, and μ_{0j} represents the random effect on the intercept of cluster j. ε_{ij} is the residual at the individual-level.

Figure 3.1

Flowchart of Data Generation and Analysis



In this simulation, X_{ij} and Z_j were set to follow a standard normal distribution - $N(0,1)$. μ_{0j} and μ_{1j} were drawn from a bivariate normal distribution with a mean of zero, whereas the variance of μ_{1j} was equal to half of the variance of μ_{0j}. ε_{ij} was drawn from a logistic distribution with a mean of 0 and a variance of $\frac{\pi^2}{3}$. Both α_X and α_Z were set to 0.5 (Fuentes et al.,

2021). Each individual was assigned either to the treatment or control group based on a random draw from a binomial distribution using their propensity score (Kush et al., 2022). If a value of 1 was generated, the individual was assigned to the treatment group, and if a value of 0 was obtained, the individual was assigned to the control group.

The data were generated iteratively to achieve the required level of treatment prevalence, with a tolerance range of 0.05. To accomplish this, the intercept α_0 of the PS model was adjusted until the desired prevalence levels were reached (prevalence = 0.2, α_0 = -0.8; prevalence = 0.5, α_0 = 0; prevalence = 0.8, α_0 = 0.8). The multilevel pseudo R^2 was computed using the values of ICC_X and the intercept α_0 following the method proposed by Snijders and Bosker (2012), as presented in Appendix B. Specifically, the pseudo R^2 consisted of two components: the individual-level R^2_{L1} and the cluster-level R^2_{L2}. Across the conditions, R^2_{L1} and R^2_{L2} ranged from about 0.06 to 0.12. The pseudo R^2 was approximately at a value of 0.18.

The outcome model incorporates the individual-level and cluster-level covariates, and is expressed as follows:

$$Y_{ij} = \beta_0 + \tau T_{ij} + \beta_X X_{ij} + \beta_Z Z_j + v_{0j} + v_{1j} X_{ij} + e_{ij}. \quad (3.2)$$

Y_{ij} represents the outcome of individual i at cluster j. β_0 is the intercept. τ is the treatment effect. β_X and β_Z are the fixed effects of the individual-level covariate X_{ij} and the cluster-level covariate Z_j, respectively. v_{1j} is the random slope of the individual-level covariate X_{ij}. v_{0j} represents the random effect on the intercept of the j^{th} cluster. e_{ij} is the residual at the individual-level. β_0 was set to 0. v_{0j}, v_{1j}, and e_{ij} were generated from the standard normal distribution. β_X and β_Z were both set to 0.5. The treatment effect (ATE) τ was set to 0.3.

Data Analysis

Once data was generated, propensity scores were estimated by using four different PS models. In addition to the MRIS model, which was used to simulate data, three additional models were used to estimate PS: a single-level PS model (SL model), a logistic regression model with fixed-cluster effects (FC model), and a multilevel logistic model with random intercept only (MRI model). The SL model did not consider variations in assigning individuals to different clusters, which can result in biased estimates (Arpino & Mealli, 2011; Thoemmes & West, 2011). The FC model can produce unstable estimates of PS when the data involve a large number of small clusters due to the large number of free parameters (Li et al., 2013). Random slope is not considered in the MRI model, which ignores the cluster variations in the individual-level covariate. These three models were considered as incorrect models in this study (see Table 3.2). Further details about these three models can be found in previous studies (Arpino & Mealli, 2011; Burnett, 2019; Leite et al., 2015; Li et al., 2013).

Table 2.2

PS Models of Simulation Study

Model for data generation	Multilevel logistic regression model with both random intercept and slope (MRIS)
Models for data analysis	Single-level model (SL)
	Logistic regression model with fixed-cluster effects (FC)
	Multilevel logistic model with random intercept (MRI)
	Multilevel logistic regression model with both random intercept and slope (MRIS)

Once the propensity scores were calculated using the four models, the weight ω_{ij} for individual *i* in cluster *j* was calculated as the inverse of the probability of the individual being assigned to the treatment or control group:

$$\omega_{ij} = \begin{cases} \frac{1}{1-\hat{\gamma}_{ij}}, & if\ T_{ij} = 0 \\ \frac{1}{\hat{\gamma}_{ij}}, & if\ T_{ij} = 1 \end{cases}, \tag{3.3}$$

where $\hat{\gamma}_{ij}$ is the estimated PS for individual *i* in cluster *j*. A weight of $\frac{1}{\hat{\gamma}_{ij}}$ is calculated if the participant is in the treatment group. In contrast, a weight of $\frac{1}{1-\hat{\gamma}_{ij}}$ is used if the participant belongs to the control group.

Research Questions 1 and 2

To answer research questions 1 and 2, the following procedures were conducted (Burnett, 2019). Firstly, the statistics of pooled ASB, VR, and PBR were obtained by averaging the covariate balance values for both the individual-level and cluster-level covariates, without considering the variations among clusters. On the other hand, the three within-cluster measures were derived by first obtaining the balance value for each cluster and then averaging across clusters for the individual-level covariate. The equations for the six diagnostics can be found in Appendix A.

The second step involved identifying the correct PS model based on the balance statistics obtained in the first step. The PS model that resulted in the lowest ASB and the highest PBR was selected as the one that achieved the most balanced groups. For the VR method, the PS model producing the value closest to 1 was selected. This procedure was repeated 100 times.

The success rate of each balance diagnostic in correctly identifying the MRIS model was recorded in the last step. The evaluation was based on the frequency of successfully selecting the MRIS model for each balance diagnostic, with the assessment that had the highest frequency being considered the best. In this study, this evaluation method was referred to as the “balance statistics” method.

In practice, a threshold is often used with a covariate balance diagnostic to assess covariate balance. The threshold is used to determine the level of difference between the treatment and control groups that is considered acceptable. The results of using this evaluation method with different threshold values were presented in Appendix C.

Research Question 3

To answer the third research question, the study examined the correlation between the covariate balance diagnostic statistics and bias in the treatment effect. After applying IPTW to the sample, two estimators of ATE, namely marginal and cluster estimators, were used to estimate ATE. The study investigated both estimators since they handle variations among clusters differently. The marginal estimator calculates the difference of the weighted overall means of outcomes between the treatment and control groups, ignoring the existence of clusters (Li et al., 2013). In contrast, the clustered estimator of ATE is obtained by first calculating the sum of the weighted ATE within each cluster and dividing the total by the sum of the weights in each cluster. Given the hierarchical structure of the data, the cluster estimator appears more appropriate in this study context than the marginal estimator. For more information on ATE estimators, refer to Li et al. (2013).

The bias of ATE is the absolute difference between the ATE estimates and the true ATE value defined previously. After assigning weights to individuals in both the control and treatment groups, ATE estimates were generated by calculating the difference between the average weighted observations of the two groups. The ATE estimates were calculated in two ways, including a marginal and a cluster estimator. The correlation between the covariate balance diagnostics and the bias in the treatment effect was then calculated for each condition based on 100 replications. In cases where the VR statistic was greater than 1, the reciprocal was used

before calculating the correlations. After calculating the correlations, the correlation coefficients from the VR and PBR methods were multiplied by -1 for easier comparison among the six diagnostics. The diagnostic with the highest correlation was considered the best method.

Software Implementation

The data were generated and analyzed using the R program (R Core Team, 2021). Specifically, the R package *lme4* (Bates et al., 2015) was utilized to construct the four PS models and to estimate propensity scores. The *glm* function was used for the SL and FC models, while the *glmer* function was employed for the MRI and MRIS models.

Results

Overall Performance of Covariate Balance Diagnostics on Model Selection

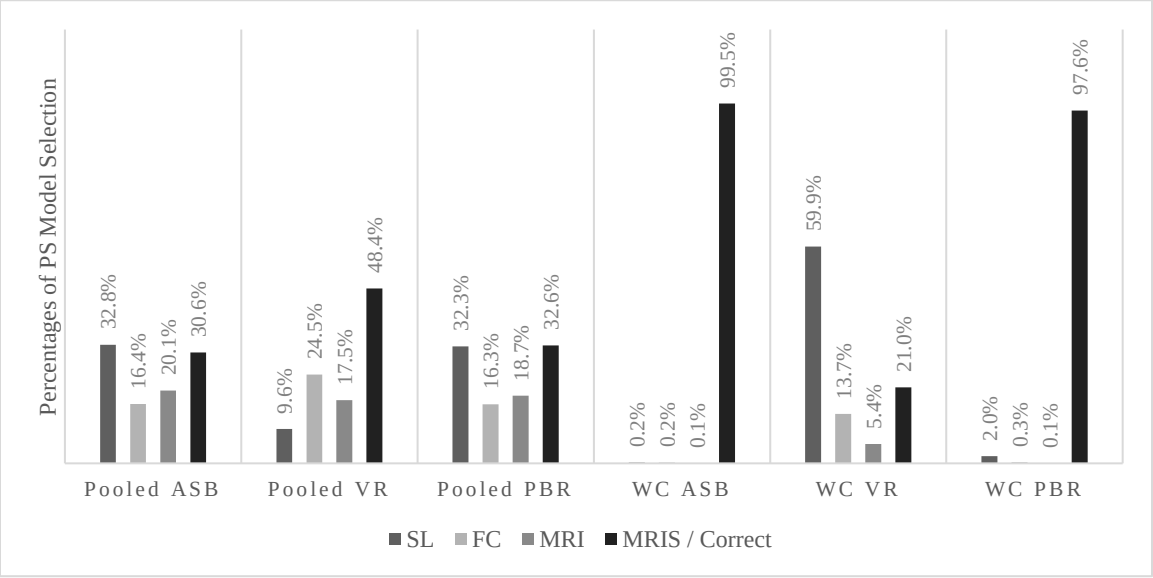
Figure 3.2 presents the aggregated results of the PS model selection for each of the six covariate balance diagnostics, which include pooled ASB, within-cluster ASB, pooled VR, within-cluster VR, pooled PBR, and within-cluster PBR, across all simulated conditions. Specifically, the percentages of each of the four models that were selected by each of the six balance diagnostics were displayed in the figure.

According to Figure 3.2, within-cluster ASB (99.5%) yielded the highest frequency of selecting the MRIS model as the PS model. Within-cluster PBR (97.6%) was associated with the second-highest percentage for getting the correct model selected. In other words, by using these two diagnostics, researchers are more likely to identify a correct PS model that generates accurate weights. This may be due to the fact that variations exist across clusters in the simulated data, making the within-cluster diagnostics more accurate for multilevel data. However, this pattern was not found in VR, where pooled VR generated more accurate results than within-cluster VR. Across the four models, the MRIS model had the highest frequency of selection by

using the pooled VR (48.4%) diagnostic. However, the percentage of the correct choice was not as dramatic as those of within-cluster ASB and PBR. The other three diagnostics, pooled ASB and PBR, and within-cluster VR, did not perform well in detecting the correct PS model. For pooled ASB and PBR, the probability of identifying a MRIS model was less than one-third, similar to that of the SL model. Within-cluster VR had the lowest probability of opting for the MRIS model, suggesting that the diagnostic may not work well for reducing selection bias.

Figure 3.2

Percentages of PS Model Selection of Covariate Balance Diagnostics with Balance Statistics



Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

The Impact of Simulation Factors on Model Selection

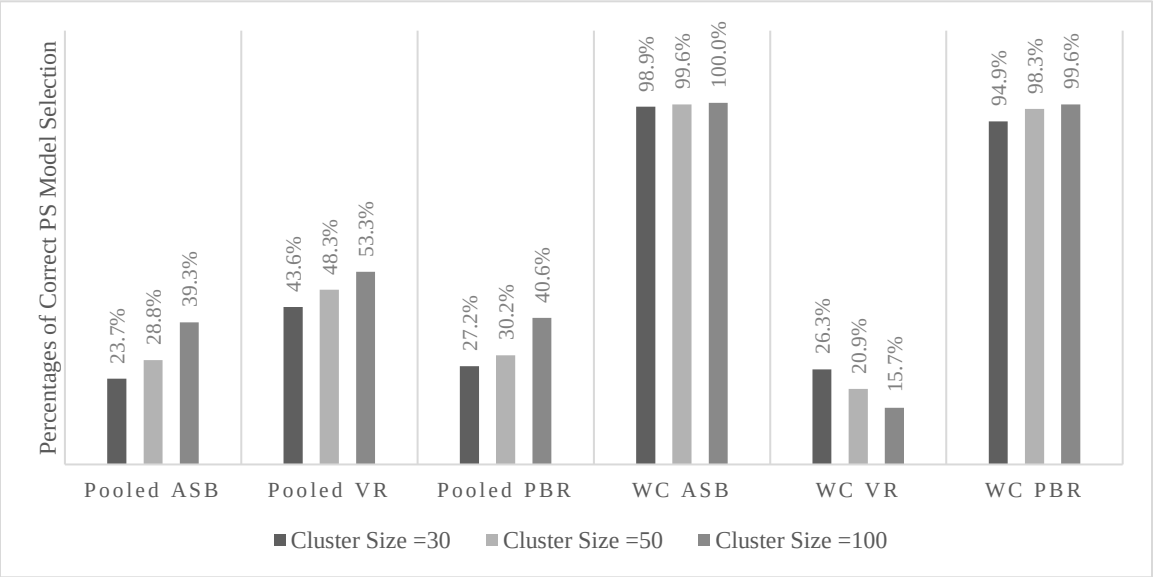
The performance of the six covariate balance diagnostics was compared across different factors. The results were displayed in Figures 3.3 to 3.6.

Cluster size

Most of the covariate balance diagnostics, except for within-cluster VR, showed a similar trend in using balance statistics for model comparison. Figure 3.3 reveals that as the cluster size increased, they led to more accurate model selection results. However, the degree of improvement varied. Within-cluster ASB and PBR only showed a slight improvement due to the ceiling effect, whereas the other three diagnostics exhibited an improvement of about 10% to 15% as cluster size changed from 30 to 100. In contrast, within-cluster VR showed a different pattern, in which the probability of identifying the correct PS model decreased with an increase in cluster size.

Figure 3.3

Percentages of Correct PS Model Selection of Covariate Balance Diagnostics with Balance Statistics by Cluster Size



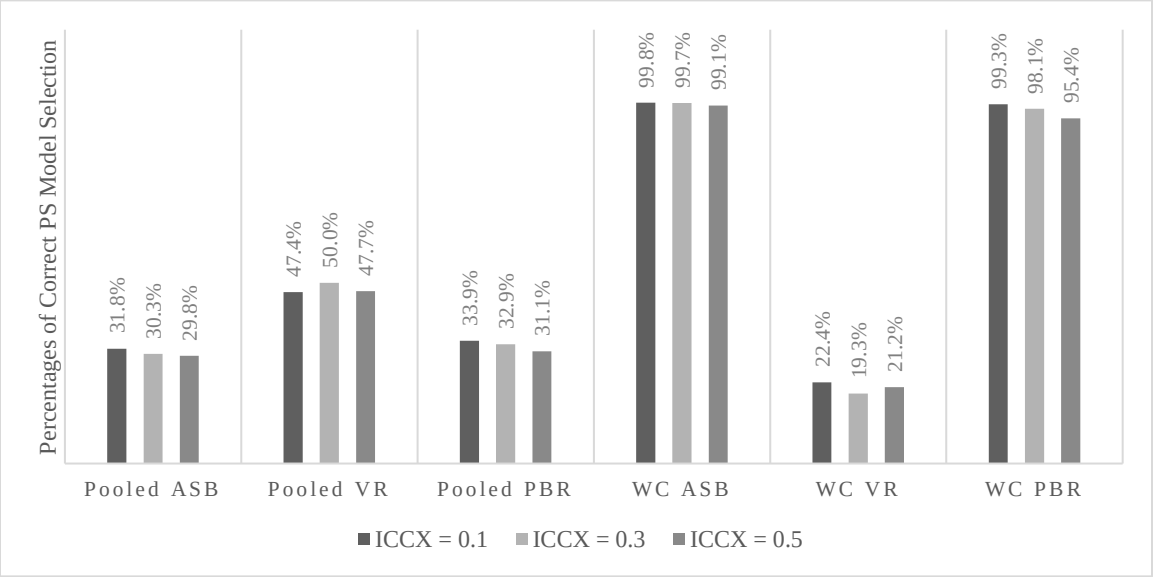
Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

ICC of Individual-level Covariate X (ICC_X)

According to Figure 3.4, the within- and clustered- ASB and PBR balance diagnostics were associated with a slightly decreased frequency in identifying the correct model as ICC_X increased. The range of the decreased percentage was within 4% across the three levels of ICC_X. However, the effect of ICC_X on pooled VR did not follow the same pattern as the other diagnostics. Specifically, when ICC_X was set to 0.3, the percentage of correct model identification appeared to be slightly higher than the other two levels. Overall, the individual-level covariate X across clusters had little to no impact on the performance of most balance diagnostics.

Figure 3.4

Percentages of Correct PS Model Selection of Covariate Balance Diagnostics with Balance Statistics by ICC_X



Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

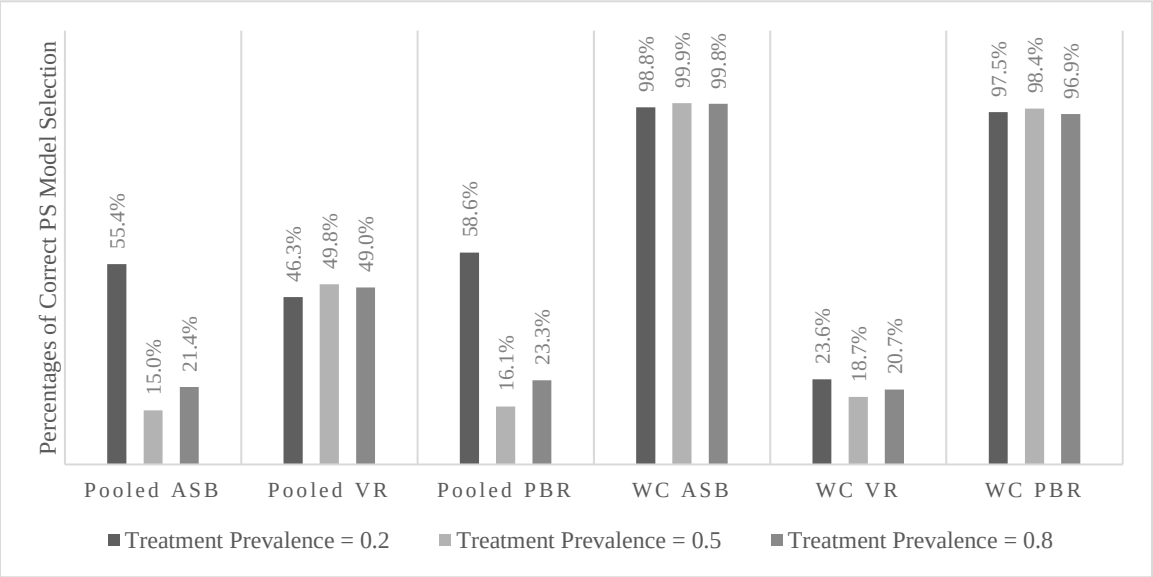
Treatment Prevalence

Figure 3.5 displays that the effect of the ratio of the sample exposed to the intervention varied across the covariate balance diagnostics. Treatment prevalence appeared to have a

minimal impact on within-cluster ASB and PBR, and pooled VR. The change among different levels of treatment prevalence was within 3%. It was observed that when the treatment prevalence was set to 0.5, these approaches produced marginally more accurate results in identifying the correct PS model. However, the impact of treatment prevalence on pooled ASB and PBR, and within-cluster VR was more apparent, following a completely different pattern compared to the other three methods. The lowest percentage of correct model selection appeared at a treatment prevalence of 0.5, followed by 0.8 and 0.2. Specifically, the percentage of correct model identification reduced from over 55% to about 15% when using pooled ASB and PBR as treatment prevalence changed from 0.2 to 0.5.

Figure 3.5

Percentages of Correct PS Model Selection of Covariate Balance Diagnostics with Balance Statistics by Treatment Prevalence



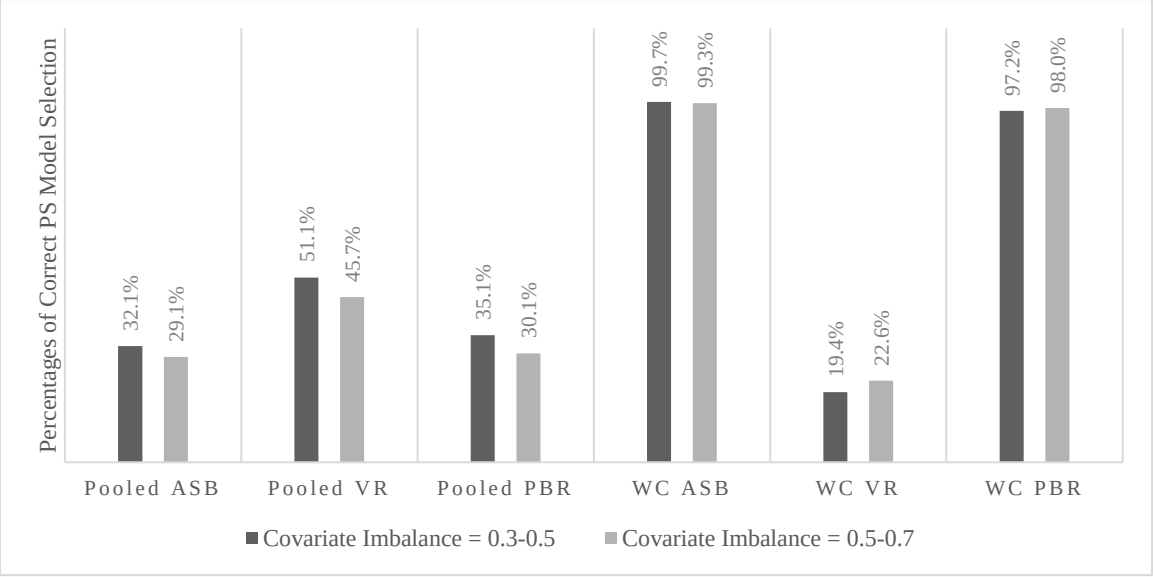
Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

Baseline Imbalance

Figure 3.6 illustrates that the impact of baseline covariate imbalance on the performance of the diagnostics was limited, with a fluctuation between -5.4% and 3.2%. Only within-cluster VR and PBR showed a slightly improved frequency of successfully identifying the correct PS model as the baseline imbalance increased. In contrast, the other four diagnostics exhibited a reverse trend. As the difference between the covariates of the two groups increased, the success rate of correct PS model detection slightly decreased.

Figure 3.6

Percentages of Correct PS Model Selection of Covariate Balance Diagnostics with Balance Statistics by Baseline Imbalance



Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

Association between Balance Statistics and ATE Bias

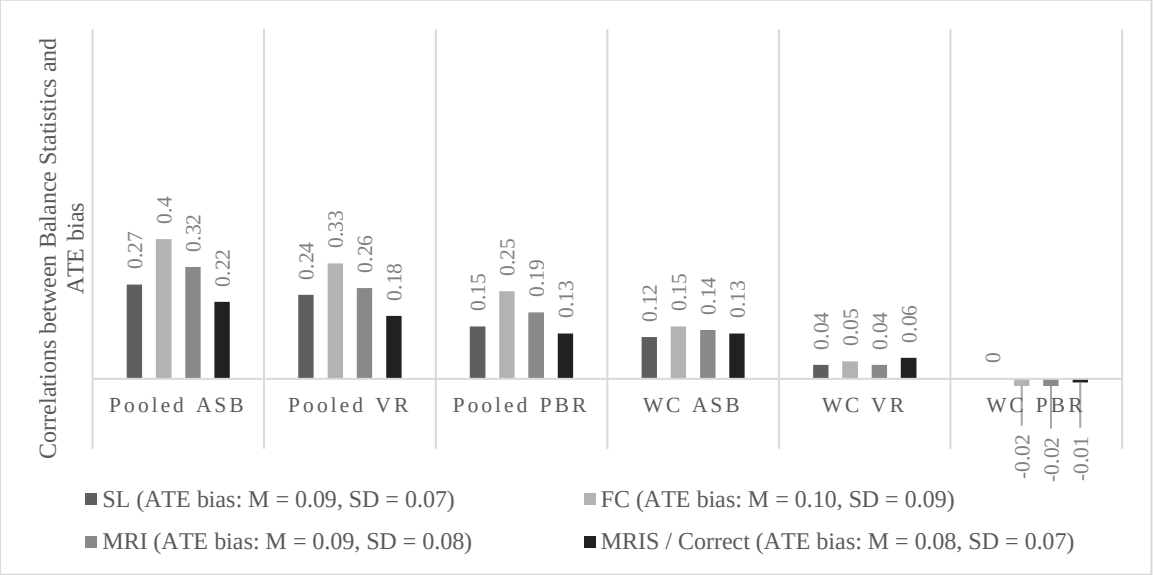
The marginal and clustered ATE bias were calculated before examining their associations with balance statistics. The values of the ATE bias were displayed in the model name legend in Figures 3.7 and 3.8. Overall, the mean and standard deviations of the clustered ATE bias were

smaller than those of the marginal bias using the same PS model. Notably, regardless of the type of ATE estimator, the smallest bias resulted from the correctly specified MRIS model.

Figure 3.7 shows the association between balance statistics of each of the six diagnostics and the marginal ATE bias. Interestingly, the highest correlation was found in pooled ASB, and, in general, the pooled balance statistics tended to have higher correlations with marginal ATE bias than within-cluster balance statistics. Within-clustered ASB produced the highest correlations among the three within-cluster measures.

Figure 3.7

Correlations between Balance Statistics and Bias in Marginal ATE



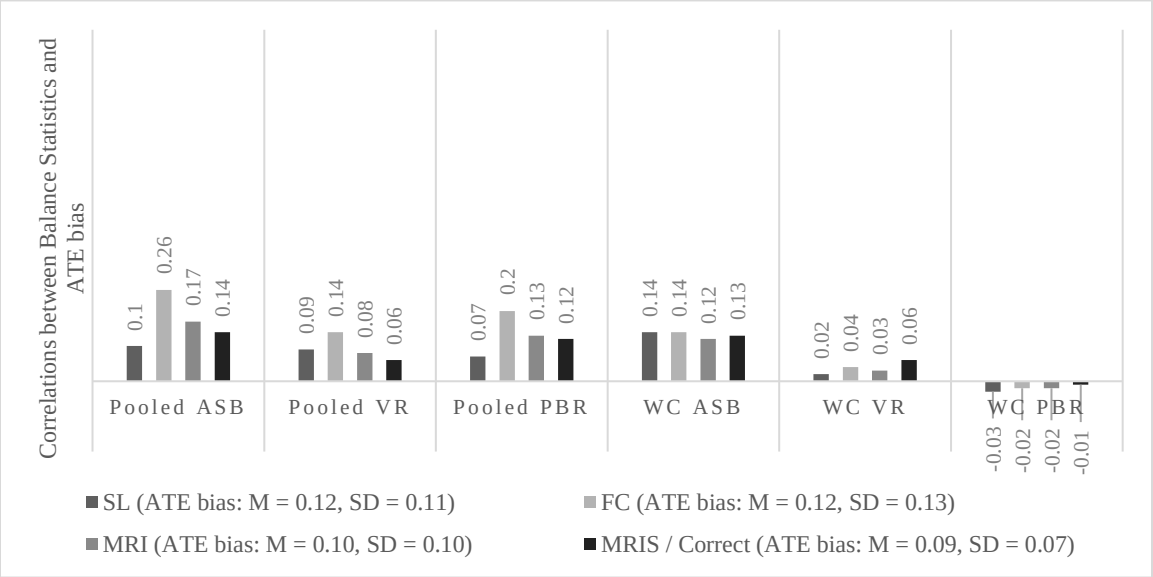
Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

Figure 3.8 displays the results of using the clustered ATE estimator. In general, the clustered ATE bias using the pooled balance diagnostics tended to have lower correlations with balance statistics than the marginal ATE bias, whereas both ATE estimators produced similar results using within-cluster balance measures. Most of the patterns associated with the six

balance diagnostics were consistent between using the clustered and marginal ATE estimators. It is interesting to note that across the PS models, the FC model was associated with the highest correlations.

Figure 3.8

Correlations between Balance Statistics and Bias in Within-cluster ATE



Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

Discussion

Propensity score methods are commonly used to adjust the covariate balance between the treatment and control groups to reduce selection bias. One of the challenges faced by researchers is choosing the most appropriate balance diagnostics to ensure the accurate assessment of the covariates between the two groups. Without a well-performed diagnostic, the correct PS model may not be identified, resulting in biased treatment effect estimation. This study aimed to examine the performance of six covariate balance diagnostics when PS weighting was employed with multilevel data. The study intended to provide general guidance for researchers on the use

of balance diagnostics in a real-world context to minimize the differences between the two comparison groups.

The overall results indicated a relatively higher probability for within-cluster ASB to identify the correct PS model for multilevel data when IPTW was used. Within-cluster ASB performed better than within-cluster VR, but pooled VR produced more accurate results than pooled ASB. These findings were consistent with Burnett's study (2019) on the use of PS matching for multilevel data. It is also noteworthy that within-cluster PBR, which has not been explored in the multilevel context, was found to be associated with robust performance in model selection. This method led to more accurate results than pooled PBR, consistent with the pattern observed in the two ASB measures. This shared pattern between PBR and ASB may be related to the fact that both methods involve calculating the difference between the means of two groups, whereas VR focuses on the variances of each group. Further research is necessary to provide a clearer explanation for this pattern.

The majority of the covariate balance diagnostics were influenced by the size of the clusters, with the accuracy of model selection being positively related to cluster size. Previous studies have also demonstrated a similar pattern in multilevel logistic regression model parameter estimation and treatment effect estimation (Leite et al., 2015; Moineddin et al., 2007). Moineddin et al. (2007) noted that a small cluster size might result in overestimation of the random slope and intercept of the multilevel model. Regarding ICC_X, ASB and PBR for both pooled and within-cluster calculations produced less accurate results as the value of ICC_X increased, but the impact was insignificant. This indicates that the variations across clusters of covariate X had only a slight negative impact on ASB and PBR. With respect to the treatment prevalence, the influence varied by each diagnostic. The prevalence of 0.5 either led to the best

or the worst performance across the diagnostics, which is consistent with the study conducted by Kush et al. (2022). Additionally, it was observed that there was a negative impact of baseline imbalance on the bias in treatment effect (Kush et al., 2022). In this study, most of the covariate balance diagnostics were associated with less accurate results as the baseline imbalance between the two groups increased, except for within-cluster VR and PBR, but the overall impact was not significant.

Regarding the association between each balance diagnostic and the treatment effect bias, pooled ASB had the highest correlations. This finding is consistent with the conclusions presented in Burnett's study (2019). Although within-cluster ASB and PBR were the most successful in identifying the appropriate PS model, they did not yield the highest correlations among the six measures. The inconsistency in model identification and the prediction in treatment bias might be related to the fact that the model showing the least error in fitting the data of treatment assignment is not necessarily the same as the model that generates the least bias in estimating the treatment effect (Burnett, 2019; Schafer & Kang, 2008).

Conclusion and Future Research

The objective of this study was to enhance researchers' comprehension of the use of covariate balance diagnostics when employing PS weighting with multilevel data. Overall, within-cluster ASB and PBR were found to produce relatively robust results in choosing the correct PS model, regardless of the simulation conditions. Another noteworthy finding is that, across all conditions, pooled VR was associated with the least precise results. The four factors examined may have different levels of impact on the success rate of selecting the correct PS model. Regarding the prediction of the bias in the treatment effect, pooled ASB had the highest accuracy among the six covariate balance diagnostics.

Although this study examined several factors and levels, they were not exhaustive. For instance, the simulated number of clusters was fixed at 30. Moineddin et al. (2007) found that valid estimates for multilevel logistic regression models can be generated using a minimum cluster size of 50 with at least 50 clusters. Additionally, in an educational context, the number of classes/schools may vary depending on the scope of the study. Future research could vary this factor to explore the impact on the precision of balance diagnostics. Moreover, only one weighting method, IPTW, was explored in the study. Researchers have suggested that IPTW may produce extreme weights for individuals when low propensity scores are assigned to treatment groups or high propensity scores are assigned to control groups, leading to inaccurate effect estimation (Austin & Stuart, 2017). Other weighting approaches, such as trimming inverse probability of treatment weighting (Stürmer et al., 2010) or overlap weighting (Li et al., 2019), could be further studied. Furthermore, researchers may consider exploring the association between the use of covariate balance diagnostics and reduction of the bias in treatment effect using different methods. This would provide additional insights into the balance diagnostics and their relationship with the accuracy of treatment effect estimation.

References

Ali, M. S., Groenwold, R. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C., ... & Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology*, *68*(2), 122-131.

Ali, M. S., Groenwold, R. H., Pestman, W. R., Belitser, S. V., Roes, K. C., Hoes, A. W., ... & Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiology and Drug Safety*, *23*(8), 802-811.

Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine*, *35*(12), 2074-2091.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770-1780.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083-3107.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399-424.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*(6), 1057-1069.

Austin, P. C., & Leckie, G. (2018). The effect of number of clusters and cluster size on statistical power and Type I error rates when testing random effects variance components in

multilevel linear and logistic regression models. *Journal of statistical computation and simulation*, *88*(16), 3151-3163.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661-3679.

Austin, P. C., & Stuart, E. A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical methods in medical research*, *26*(4), 1654-1670.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, *67*(1), 1–48. doi:10.18637/jss.v067.i01.

Bai, H., & Clark, M. H. (2018). *Propensity score methods and applications*. Sage Publications.

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*, *9*(6), 377-385.

Burnett, A. (2019). *The Performance of Balance Diagnostics for Propensity-Score Matched Samples in Multilevel Settings* (Doctoral dissertation).

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417-446.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 1231-1236.

Fuentes, A., Lüdtke, O., & Robitzsch, A. (2022). Causal inference with multilevel data: A comparison of different propensity score weighting approaches. *Multivariate Behavioral Research*, *57*(6), 916-939.

Granger, E., Watkins, T., Sergeant, J. C., & Lunt, M. (2020). A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Medical Research Methodology*, *20*, 1-9.

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed). SAGE Publications.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199-236.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, *171*(2), 481-502.

Jacovidis, J. N., Foelber, K. J., & Horst, S. J. (2017). The effect of propensity score matching method on the quantity and quality of matches. *The Journal of Experimental Education*, *85*(4), 535-558.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523-539.

Kim, J., & Seltzer, M. (2007). *Causal inference in multilevel settings in which selection processes vary across schools* (CSE technical report 708). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California.

Kush, J. M., Pas, E. T., Musci, R. J., & Bradshaw, C. P. (2022). Covariate Balance for Observational Effectiveness Studies: A Comparison of Matching and Weighting. *Journal of Research on Educational Effectiveness*, 1-24.

Leite, W. L., Aydin, B., & Gurel, S. (2019). A comparison of propensity score weighting methods for evaluating the effects of programs with multiple versions. *The Journal of Experimental Education*, *87*(1), 75-88.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research*, *50*(3), 265-284.

Leite, W. L., Jing, Z., Kuang, H., Kim, D., & Huggins-Manley, A. C. (2021). Multilevel mixture modeling with propensity score weights for quasi-experimental evaluation of virtual learning environments. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 964-982.

Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, *188*(1), 250-257.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*(19), 3373-3387.

McCormick, M. P., O'Connor, E. E., Cappella, E., & McClowry, S. G. (2013). Teacher–child relationships and academic achievement: A multilevel propensity score model approach. *Journal of School Psychology*, *51*(5), 611-624.

Melguizo, T., Kienzl, G. S., & Alfonso, M. (2011). Comparing the educational attainment of community college transfer students and four-year college rising juniors using propensity score matching methods. *The Journal of Higher Education*, *82*(3), 265-291.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC medical research methodology, 7*(1), 1-10.

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.

Olmos, A., & Govindasamy, P. (2015). A practical guide for using propensity score weighting in R. *Practical Assessment, Research, and Evaluation, 20*(13), 1-8.

Powell, M. G., Hull, D. M., & Beaujean, A. A. (2020). Propensity Score Matching for Education Data: Worked Examples. *The Journal of Experimental Education, 88*(1), 145-164.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage Publications.

Rickles, J. H., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics, 39(6),* 612-636.

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. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3), 169-188.

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

Snijders, T. A., & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). Sage Publishers.

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research, and Evaluation*, *18*(13) 1-12.

Stürmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *American journal of epidemiology*, *172*(7), 843-854.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90-118.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514-543.

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology*, *82*(5), 906-919.

Westine, C., Wu, T., Maher, D., Kim, Stella., & Dong, N. (2023, April 13-16). *Effectiveness of a university learning assistant program* [Conference presentation]. The Annual Meeting of American Educational Research Association, Chicago, IL.

Yamada, H., Bohannon, A. X., Grunow, A., & Thorn, C. A. (2018). Assessing the effectiveness of Quantway®: A multilevel model with propensity score matching. *Community College Review*, *46*(3), 257-287.

Yamada, H., & Bryk, A. S. (2016). Assessing the first two years' effectiveness of Statway®: A multilevel model with propensity score matching. *Community College Review*, *44*(3), 179-204.

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, *7*(1):16.

Appendix A: Covariate Balance Measures

In PS weighting, the pooled ASB is calculated as an absolute standardized bias over all the individual on weighted individual-level and cluster-level covariates (Arpino & Maellie, 2011; Burnett, 2019), which can be defined as,

$$ASB_{weight} = \left| \frac{\bar{X}_{T_{weight}} - \bar{X}_{C_{weight}}}{\sqrt{\frac{S^2_{T_{weight}} + S^2_{C_{weight}}}{2}}} \right| \times 100\% \quad \text{((a)}$$

where $\bar{X}_{T_{weight}}$ and $\bar{X}_{C_{weight}}$ are the weighted sample means on the covariate of the treatment and control groups, respectively, which can be calculated as in Equations (b) and (c).

$$\bar{X}_{C_{weight}} = \frac{\sum \omega_{C_i} x_{C_i}}{\sum \omega_{C_i}}, \quad \text{(b)}$$

where ω_{C_i} and x_{C_i} are the weight and covariate for the i^{th} individual in the control group *C*, respectively. ω_{C_i} can be obtained by using IPTW.

$$\bar{X}_{T_{weight}} = \frac{\sum \omega_{T_i} x_{T_i}}{\sum \omega_{T_i}}, \quad \text{(c)}$$

where ω_{T_i} and x_{T_i} are the weight and covariate for the i^{th} individual in the treatment group T, respectively. The value of ω_{Ti} can be obtained by using IPTW.

$S^2_{T_{weight}}$ and $S^2_{C_{weight}}$ are the weighted variances on the covariates of the treatment and control groups, respectively, which can be calculated using Equations (d) and (e) (Austin & Stuart, 2015).

$$S^2_{T_{weight}} = \frac{\sum \omega_{T_i} x_{T_i}}{\left(\sum \omega_{T_i}\right)^2 - \sum \omega^2_{T_i}} \sum \omega_{T_i} \left(x_{T_i} - \bar{X}_{T_{weight}}\right)^2, \quad \text{(d)}$$

$$S^2_{C_{weight}} = \frac{\sum \omega_{C_i} x_{C_i}}{\left(\sum \omega_{C_i}\right)^2 - \sum \omega^2_{C_i}} \sum \omega_{C_i} (x_{C_i} - \bar{X}_{Cweight})^2. \quad \text{(e)}$$

The pooled method evaluates covariate balance for multilevel data as if it is single-level data, so ASB_{weight} can be directly used as the pooled ASB_{weight}. However, when the variants across clusters are considered, the ASB_{weight} needs to be adapted for calculating the value for within-cluster *ASB.*

The within-cluster ASB can be obtained by calculating the ASB for each cluster, then averaging the ASBs across all clusters, as shown in equation (f),

$$ASB_{within-cluster} = \frac{1}{J}\sum_{j=1}^{J} ASB_{weight_j}, \quad \text{(f)}$$

where *J* is the number of clusters and ASB_{weight_j} is the weighted ASB of the j^{th} cluster. The value of ASB_{weight_j} can be calculated by using Equation (a), but it should be noted the ASB varies across clusters.

The pooled variance ratio, VR_{weight}, for each covariate is defined as the ratio of the sample variances of the treatment and the control groups on the selected covariates (Austin, 2009; Burnett, 2019), which is expressed in equation (g),

$$VR_{weight} = \frac{S^2_{T_{weight}}}{S^2_{C_{weight}}} . \quad \text{(g)}$$

This index indicates the dispersion of the covariates between the two groups instead of examining the means. The distributions between the two groups are similar if VR_{weight} is close to 1.

The within-cluster variance ratio, $VR_{within-cluster}$, can be calculated by averaging the VRs for each cluster. The equation is shown as follows (Burnett, 2019),

$$VR_{within-cluster} = \frac{1}{J}\sum_{j=1}^{J} VR_{weight_j}, \quad \text{(h)}$$

Pooled PBR is calculated as the percent reduction in the absolute difference of means between treatment and control groups before and after applying the weights to covariates.

$$PBR_{pooled} = \frac{|\bar{X}_T - \bar{X}_C| - \left|\bar{X}_{T_{weight}} - \bar{X}_{C_{weight}}\right|}{|\bar{X}_T - \bar{X}_C|} \times 100\%, \quad \text{(i)}$$

where $\bar{X}_T$ and $\bar{X}_C$ are the sample means on the covariates of the treatment and control groups without applying the weights, respectively, which can be calculated as Equations (j) and (k).

$$\bar{X}_T = \frac{\sum x_{T_i}}{N_T}, \quad \text{(j)}$$

where N_T is the number of individuals in the treatment group, and

$$\bar{X}_C = \frac{\sum x_{C_i}}{N_C}, \quad \text{(k)}$$

where N_C is the number of individuals in the treatment group.

Within-cluster PBR can be obtained by calculating the PBR for each cluster, then averaging the PBRs across all clusters, as shown in equation (l)

$$PBR_{within-cluster} = \frac{1}{J}\sum_{j=1}^{J} \frac{\left|\bar{X}_{T_j} - \bar{X}_{C_j}\right| - \left|\bar{X}_{T_{weight_j}} - \bar{X}_{C_{weight_j}}\right|}{\left|\bar{X}_{T_j} - \bar{X}_{C_j}\right|} \times 100\%, \quad \text{((l)}$$

where $\bar{X}_{T_j}$ and $\bar{X}_{C_j}$ are the sample means on the covariate of the treatment and control groups in the j^{th} cluster, respectively. $\bar{X}_{T_{weight_j}}$ and $\bar{X}_{C_{weight_j}}$ are the weighted sample means on the covariates of the treatment and control groups in the j^{th} cluster, respectively.

Appendix B: Multilevel Pseudo-R square

Snijders and Bosker (2012) proposed a method to obtain multilevel pseudo-R square.

$$R^2 = R_{L1}^2 + R_{L2}^2, \quad \text{(m)}$$

where R^2 is the total R square value, R_{L1}^2 is the R square value at the individual-level, and R_{L1}^2 is the R square value at the cluster-level.

$$R_{L1}^2 = \frac{(\alpha_X^2 + \sigma_{u_1}^2)(1 - ICC_X)}{\alpha_X^2 + \alpha_Z^2 + \sigma_{u_0}^2 + ICC_X(\sigma_{u_1}^2 + \frac{\pi^2}{3})}, \quad \text{(n)}$$

$$R_{L2}^2 = \frac{ICC_X(\alpha_X^2 + \sigma_{u_1}^2) + \alpha_Z^2}{\alpha_X^2 + \alpha_Z^2 + \sigma_{u_0}^2 + ICC_X(\sigma_{u_1}^2 + \frac{\pi^2}{3})}, \quad \text{(o)}$$

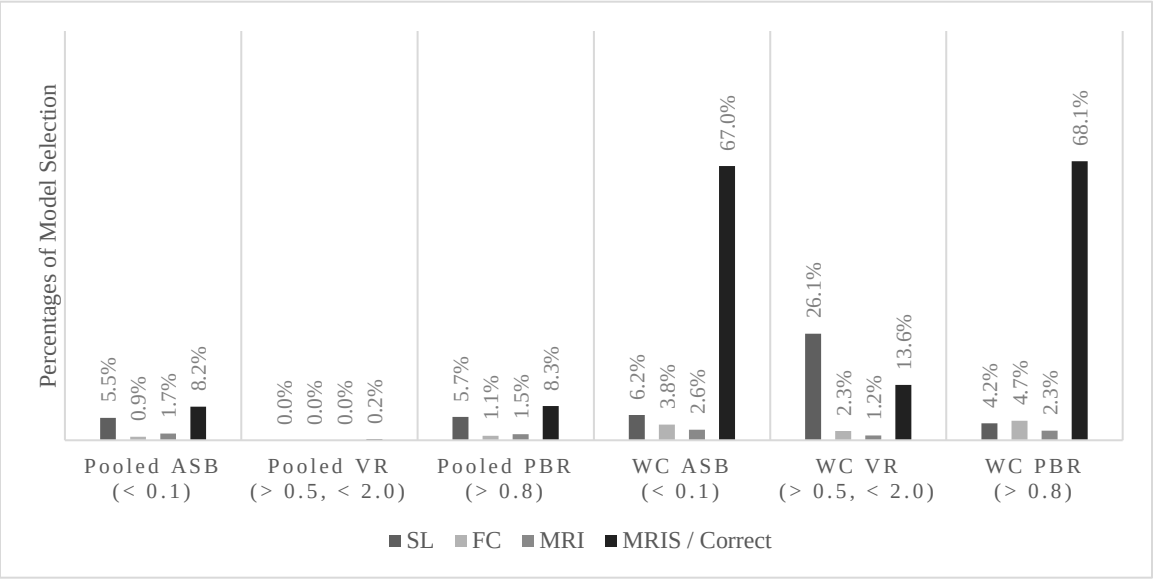
where ICC_X is the ICC of the individual-level covariate X, $\sigma_{u_0}^2$ is the variance of the variable u_0, and $\sigma_{u_1}^2$ is the variance of the variable u_1.

Appendix C

A second method to evaluate covariate balance diagnostics – a "balance threshold" method – was included in the study by replacing step 2 under the data analysis section with a different procedure. Specifically, the statistics generated from step 1 were compared against the respective thresholds. The common thresholds for balance for each statistic include: ASB (< 0.1) (Rubin 2001), VR (0.5 to 2.0) (Rubin, 2001), and PBR (> 0.8) (Cochran & Rubin, 1973). Specifically, a binary score (0, 1) was assigned based on the comparison between the results of the balance measure and the threshold (Burnett, 2019). If the statistic indicated a balance of the covariates between the two groups, a score of 1 was recorded. Otherwise, a 0 was recorded. For pooled balance measures, this binary-check mechanism was applied for each covariate before averaging the binary results across the covariates. For within-cluster measures, the balance statistics calculated within each cluster was compared to the threshold and the binary results were averaged across clusters. The resulting values associated with the four PS models were then compared. If there was a tie in the binary results between two models, no model was selected. Then the process repeated 100 times and the percentage of each covariate balance diagnostics selecting the correct model was recorded for further comparison. Results are presented in Figure 3.A.

Figure 3.A

Percentages of PS Model Selection of Covariate Balance Diagnostics with Balance Thresholds



Note. WC refers to within-cluster. SL, FC, MRI, MRIS/Correct indicate single-level model, logistic regression model with fixed-cluster effects, multilevel logistic model with random intercept, and multilevel logistic regression model with both random intercept and slope, respectively.

OVERALL CONCLUSION

Findings and Implications

Researchers from various fields constantly face challenges for maintaining comparability in different contexts. . This dissertation aims to address three commonly encountered issues, each of which was addressed using simulation techniques. Specifically, by taking various simulation conditions into consideration, this dissertation intends to provide insights into selecting an appropriate method to make sure that the comparability between test scales, test forms, and treatment and control groups is achieved. If comparability is not maintained for a test, test results validity can be undermined and test score interpretation will be compromised. In addition, without obtaining two equivalent groups, the estimated treatment effect of an intervention of a quasi-experimental study can be biased, potentially leading to misjudgments, and hindering the development of effective policies. The three obstacles that impede the attainment of comparability and results of each study are discussed below.

1. Which missing data handling method should be used to achieve accurate scale transformation for maintaining comparability among examinees taking different test forms?

Different calibration methods can be used by testing programs for estimating item parameters and examinees' proficiency. For concurrent calibration, when two test forms are calibrated simultaneously using computer software, there is no need to conduct scale transformation because the parameters are automatically placed on the same scale. Scale linking is also not required for separate calibration when two groups are equivalent. However, in the case of separate calibration where two groups may not be equivalent, scale transformation is a necessary step. For example, the 2018 Programme for International Student Assessment (PISA)

(OCED, 2019) used linear scale transformation to align ability and item parameters estimated from two groups on the common scale. Previous research has extensively discussed how to address missing responses when estimating item and ability parameters (e.g., Robitzsch, 2021; Xiao & Bulut, 2020). However, it is still an open question whether these methods for handling missing data affect the accuracy of the scale transformation procedure.

In the first study of the dissertation, how to maintain the comparability between IRT scales under the CINEG design with the presence of missing data was examined. In particular, six missing data handling approaches were investigated in scale linking when missing data were observed in common items. It was revealed that RF, MI, and FIML were associated with less errors as compared to the other methods. With the use of these methods, the comparability between two test scales could be improved. It is worth noting that the rapid advancement of statistical software and packages make relatively complex missing data handling approaches (RF, MI, and FIML) more accessible and less time-consuming. In contrast, LWD was found to be associated with the worst performance when dealing with missing data, and therefore not recommended to use in practice.

This study emphasized the need for testing programs that involve the scale linking procedure to consider not only the effect of missing data handling approaches on parameter estimation but also the association between these approaches and scale linking performance. The utilization of incorrect linking coefficients in the scaling process can result in incorrect conclusions about examinees' abilities, potentially putting some test takers at a disadvantage. Moreover, the use of improper linking coefficients can result in inaccurate item parameter estimation, leading to future test administrations incorporating these biased item parameters, and

increased costs. This perpetuates the negative impact and potentially can cause long-term harm to the validity of the test.

2. How can test score comparability be maintained if responses on different test forms are graded by raters with varying levels of severity and variability?

Focusing on maintaining comparability across multiple test forms, the second study aims to develop an innovative observed-score equating method for rater-mediated assessments. With this type of assessment, errors may come not only from the differences in difficulty levels on two forms, but also from the differences in raters who assign scores to responses on both forms. This study develops and illustrates a process of integrating the IRT observed-score equating method and the HRM model with an intent to take rater effects into account in the equating procedure.

The newly proposed HRM observed-score equating method produced less errors in bias and RMSE than the results for the GPCM method when the responses on the new forms were graded by aberrant raters (severe raters, unreliable raters and severe/unreliable raters). When the new form was scored by normal raters, both the GPCM and the HRM observed-score equating methods generated comparable results. The conclusions of the study indicated that with the use of the suggested equating method, the differences between two test forms can be better adjusted, ensuring a valid comparison and accurate interpretations of test results.

The proposed HRM method may have many practical applications. Standardized testing programs, such as the Advanced Placement Exam, and many clinical certification programs rely on human raters to evaluate examinees' skills in a particular domain. The HRM equating method can serve as an additional gatekeeper against rater error to produce reliable equating results by taking into account different types of rater effects. In addition, testing programs often invest significant funds in rater training and calibration to maintain rater quality. The utilization of the

HRM equating method will allow resources to be allocated and used more efficiently and effectively, while at the same time maintaining the score comparability between different forms of a test. Moreover, by considering the rater effects for each test administration, the impact of rater drift can be adjusted with the use of the HRM equating method. In sum, this new equating method has the potential to maintain score scales over the long term, ensuring consistent interpretations of test scores. This, in turn,t increases the likelihood of making valid educational policies or decisions.

3. Which covariate balance diagnosis is the most effective for selecting a more accurate propensity score model and constructing comparable groups when weighting is used with multilevel data?

The third study explores the performance of six covariate balance diagnostics when PS weighting is used with multilevel data. Whether the two groups are comparable or not should be evaluated based on proper balance diagnostics. Results of this study showed that within-cluster ASB and within-cluster PBR produced more accurate results in specifying a correct PS model as compared to other diagnostics. On the contrary, pooled and within-cluster VR were not recommended.

In educational settings, conducting multilevel studies is a common practice due to the nested nature of data structures. For example, in the learning assistant study conducted by Westine et al. (2023), students were nested in class sections and class sections were nested in each course subject. There could be a great degree of variations of students within each section and each course, such as student' gender, ethnicity, geographic region, and socio-economic status. These factors potentially lead to selection bias, if not properly addressed, thereby hindering accurate estimation of the treatment effect. However, whether the within-cluster

variations should be considered in covariate balance diagnostics was not clear. In particular, which covariate balance diagnostic leads to more valid results when PS weighting is used with nested data was not clear.

The findings of this research should increase understanding of covariate balance diagnostics, particularly those receiving relatively less attention in the literature including within-cluster ASB and within-cluster PBR. Proper assessment of covariate balance increases the likelihood of selecting a correct PS model, thereby reducing the risk of selection bias. When the two comparison groups are as similar as possible, then the estimated effect of an intervention can become more accurate. This will enable educational policymakers to make informed decisions based on results obtained from research that uses precise covariate balance diagnostics methods.

Future Research

The intriguing results of this dissertation have opened up various possibilities for future research. First, all of the three studies in this dissertation are based on simulated data. Simulation studies allow researchers to conduct computer experiments by generating data through random sampling from a known probability distribution, which is time-saving and cost-efficient compared to collecting and analyzing real data (Feinberg & Rubright, 2016; Morris et al., 2019). In simulation studies, the true parameters are known, enabling researchers to evaluate the performance of different statistical methods based on the differences between the estimated parameters and the parameters (Morris, 2019). Researchers have highlighted several important steps for conducting simulation studies, including resampling or simulating from parametric models, identifying methods to be evaluated, listing performance measures to be estimated and providing the rationales, analyzing the data, and reporting the results (Carsey & Harden, 2013; Morris et al., 2019). The three studies follow these steps to contribute to methodology in

educational research. The parameters used in the simulations are selected based on previous research or a common practice with an intent to better approximate realities. Although simulation studies have numerous advantages, they cannot replace empirical studies based on real data (Feinberg & Rubright, 2016). Thus, researchers need to be aware of the limitations when trying to generalize the results of this dissertation. Future research is needed to examine if the results from the simulation studies can be reproduced with real data.

Second, the selection of the methods evaluated in the dissertation was justified with rationales. For instance, promising results were noted by previous research or suitability with a specific type of data. In practice, however, there are other methods used in the three research areas. Future research can be conducted to evaluate additional methods for providing more comprehensive information. For example, the impact of using the IRtree method and Expected-Maximization (EM) algorithm to address missing data problem on scale linking is worth exploring. With respect to the second study, previous research has examined how rater drift is addressed within the use of Generalizability Theory (Harik et al., 2009). Future research may also extend the use of the theory in the equating context and compare it against with the use of the HRM equating method. The conclusion of the third study suggested the two within-cluster diagnostics perform well, but this does not mean they are the only methods that should be used for the assessment of covariate balance. The performance of graphical diagnostics, such as QQ plots and scatter plots, are also worth exploring with multilevel data. Also, their effectiveness with the use of different PS methods should be investigated.

This dissertation only focuses on three factors that may affect comparability at various stages of the research process. Although the foci of comparability considered here are specific situations related to the handling missing data, equating test-forms, and the design of propensity

score weighting techniques, concerns of comparability are buried throughout the entire research process. Therefore, improving comparability of data and groups (or other cases) in applied research will be an ever-evolving challenge facing the research team.

Novel problems of comparability can emerge from pedagogical and technological advances which alter the educational context, or methodological innovations which ultimately extend the frontiers of data collection, design, and analysis. For example, since the year 2020, more testing programs have started to offer remote/at-home tests. How to achieve comparability between remote or at-home tests and onsite tests is a critical topic to investigate (Puhan & Kim, 2022). Furthermore, growing evidence suggests that socio-economic disparities may also exacerbates the digital divide in the technological era, having impacts not only for comparability of measures, but for the delivery and evaluation of educational interventions. Students with limited digital literacy skills may be at a significant disadvantage with a variety of unknown impacts as compared to those who have greater access to resources.

Simulation research plays an important role in helping to answer these and related types of questions. Thus, it is vital for simulation researchers to work closely with policymakers and other consumers of research to properly define and inform the design of their comparability studies, as well as to interpret their findings to provide evidence-based answers that address these and other critical societal needs. Additionally, researchers seeking to conduct studies on comparability in educational settings must continue to adapt their designs in response to emerging innovations. As computing technology advances so must the scope of simulation studies to inform applied research. Future research should extend the present simulation studies to refine strategies to suit different contexts, address new applications, and respond to more complex models.

REFERENCES

Adams, R. (2007). Cross-moderation methods. In P. E. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 212-245). London, UK: Qualifications and Curriculum Authority.

Ali, M. S., Groenwold, R. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C., ... & Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology*, *68*(2), 122-131.

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

Anthony, C., Styck, K., Cooke, E., Martel, J., & Frye, K. (2020). Evaluating the impact of rater effects on behavior rating scale score validity and utility. *School Psychology Review*, 1-15.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083-3107.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*(6), 1057-1069.

Bai, H., & Clark, M. H. (2018). *Propensity score methods and applications*. Sage Publications.

Berger, V. W. (2006). A review of methods for ensuring the comparability of comparison groups in randomized clinical trials. *Reviews on Recent Clinical Trials*, *1*(1), 81-86.

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). *Comparability of Large-Scale educational assessments: Issues and recommendations*. Washington, DC: National Academy of Education.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*(3), 321-364.

Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo simulation and resampling methods for social science*. Sage Publications.

Congdon, P. J., & McQueen, J. (2000). Unmodeled rater discrimination error. *Objective measurement: Theory into Practice*, *5*, 165-180.

Dorans, N. J., Pommerich, M. E., & Holland, P. W. (2007). Linking and aligning scores and scales. In *Linking and Aligning Scores and Scales, Jun, 2005, Princeton University, Princeton, NJ, US; The aforementioned conference provided raw material for this volume.*. Springer Science+ Business Media.

Elliott, G. (2013). A guide to comparability terminology and methods. *Research Matters, Cambridge, special*, (2), 9-19.

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Evans, C. M., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. *Educational measurement: issues and practice*, *36*(3), 24-34.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245.

Green, B. F. (1995). Comparability of scores from performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 13-15.

Greenland, S., & Neutra, R. (1980). Control of confounding in the assessment of medical technology. *International Journal of Epidemiology*, *9*(4), 361-367.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*(1), 29-46.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144-149.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods*, *5*(1), 64-86.

Hung, S. P., Chen, P. H., & Chen, H. C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal*, *24*(4), 345-357.

Jacovidis, J. N., Foelber, K. J., & Horst, S. J. (2017). The effect of propensity score matching method on the quantity and quality of matches. *The Journal of Experimental Education*, *85*(4), 535-558.

Kassim, N. L. A. (2011). Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online® Journal of Language Studies*, *11*(3), 179-197.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*(4), 371-397.

Kim, S., & Kolen, M. J. (2022). Scale Linking for the Testlet Item Response Theory Model. *Applied Psychological Measurement*, *46*(2), 79-97.

Kolen, M. J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, *6*(2), 73-96.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Linacre, J. M. (1989). Many-faceted Rasch measurement. Chicago: MESA Press.

Loyd, B., Engelhard, Jr, G., & Crocker, L. (1995). Achieving form-to-form comparability: Fundamental issues and proposed strategies for equating performance assessments of teachers. *Educational Assessment*, *3*(1), 99-110.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems 1. *ETS Research Bulletin Series*, *1977*(1), i-41.

Mislevy, R. J. (2017). Missing responses in item response modeling. *Handbook of Item Response Theory, Volume Two: Statistical Tools*, 171-194.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074-2102.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Newgard, C. D., Hedges, J. R., Arthur, M., & Mullins, R. J. (2004). Advanced statistics: the propensity score—a method for estimating treatment effect in observational research. *Academic Emergency Medicine*, *11*(9), 953-961.

Nieto, R., & Casabianca, J. M. (2019). Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response tests. *Journal of Educational Measurement*, *56*(3), 547-581.

OECD (2019) *PISA 2018 technical report*. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341-384.

Powell, M. G., Hull, D. M., & Beaujean, A. A. (2020). Propensity score matching for education data: Worked examples. *The Journal of Experimental Education*, *88*(1), 145-164.

Puhan, G., & Kim, S. (2022). Score comparability issues with at-home testing and how to address them. *Journal of Educational Measurement*, *59*(2), 161-179.

Qiu, X. L., Chiu, M. M., Wang, W. C., & Chen, P. H. (2021). A new item response theory model for rater centrality using a hierarchical rater model approach. *Behavior Research Methods*, 1-15.

Robitzsch, A. (2021). On the treatment of missing item responses in educational large-scale assessment data: An illustrative simulation study and a case study using PISA 2018

mathematics data. *European Journal of Investigation in Health, Psychology and Education*, *11*(4), 1653-1687.

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. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, *79*(387), 516-524.

Sijtsma, K., & Van Der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, 38, 505– 528.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201-210.

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research, and Evaluation*, *18*(13), 1-12.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*(4), 333-344.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3*(3), 115.

Westine, C., Wu, T., Maher, D., Kim, Stella., & Dong, N. (2023, April 13-16). *Effectiveness of a university learning assistant program* [Conference presentation]. The Annual Meeting of American Educational Research Association, Chicago, IL.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (1999). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. Paper presented at the Annual meeting of the American Educational Research Association, Montreal, Canada.

Wolfe, E. W. (2020). Human scoring with automated scoring in mind. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Boca Raton, FL: Chapman & Hall/CRC.

Xiao, J., & Bulut, O. (2020). Evaluating the performances of missing data handling methods in ability estimation from sparse data. *Educational and Psychological Measurement*, *80*(5), 932-954.