5-2022

Posterior Predictive Model Checking of the Hierarchical Rater Model

Nnamdi Chika Ezike
University of Arkansas, Fayetteville

Posterior Predictive Model Checking of the Hierarchical Rater Model

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Educational Statistics and Research Methods

by

Nnamdi Chika Ezike
Enugu State University of Technology
Bachelor of Science in Statistics, 2010
Montana State University
Master of Science in Statistics, 2018

May 2022
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council

_____
Allison Ames Boykin, Ph.D.
Dissertation Chair

_____
Ronna C. Turner, Ph.D.
Committee Member

_____
Xinya Liang, Ph.D.
Committee Member

_____
Brandon L. Crawford, Ph.D.
Committee Member

Abstract

Fitting wrongly specified models to observed data may lead to invalid inferences about the model parameters of interest. The current study investigated the performance of the posterior predictive model checking (PPMC) approach in detecting model-data misfit of the hierarchical rater model (HRM). The HRM is a rater-mediated model that incorporates components of the polytomous item response theory (IRT) model, such as the partial credit model (PCM) and generalized partial credit model (GPCM), at the second level of the hierarchy, to model examinees' responses to performance assessments. To date, the HRM has not been rigorously evaluated using PPMC techniques. Monte Carlo simulations were employed to explore the effectiveness of 13 discrepancy measures in detecting model-data misfit of the HRM. Misfits were assessed at the test-, item-, and rater-level. Using the HRM-GPCM, data were generated by varying the rating design (fully-crossed and spiral), proportion of aberrant raters (no rater effects and 25% of the raters with rater effects), and number of examinees (250 and 500). Data generated were analyzed using the HRM-PCM and HRM-GPCM with eight raters and four items. Type I error and power rates were computed for each discrepancy measure.

The results indicate that the standard deviation of the total score was the only useful discrepancy measure at the test level. Furthermore, the item-total correlation and odds ratio were found to be powerful in detecting misspecification of the HRM-PCM at the item level. Of the three rater-level discrepancy measures, only the score-estimate correlation and rater-total correlation were adequate in detecting the misfit of the HRM-PCM. The performance of the discrepancy measures in detecting misfit of HRM-PCM differed by the magnitude of the item discrimination parameters. The impact of the simulation factors on detecting misfit of the HRM-PCM and implications are further discussed.

Keywords: Rater; performance assessments; hierarchical rater model; posterior predictive model checking; model-data fit

Acknowledgement

I am grateful to my advisor, Dr. Allison Ames Boykin, for her guidance and invaluable feedback that shaped this work. I would like to thank the faculty members in the Educational Statistics and Research Methods program at the University of Arkansas for laying the foundation for me to thrive in the program. I am also grateful to my dissertation committee members, Dr. Ronna Turner, Dr. Xinya Liang, and Dr. Brandon Crawford, for their support and feedback throughout my writing process.

There are several colleagues whose support and guidance helped me to come this far. I would like to give special thanks to my mentors at Educational Testing Service, Dr. Jodi Casabianca and Dr. Michael Walker, for their advice during the early stages of this work. I would also like to thank Dr. Wen-Juo Lo, Dr. Brian Primack, and Dr. Michael Hevel for their leadership and support during my graduate education. My graduate school journey is not complete without the support of fellow graduate students. I am grateful to Mary Margaret Hui, Aaron Myers, James Weese, Ji Li, and Nana Amma Asamoah for the opportunity to learn and collaborate with them.

The support of my friends and mentors cannot be overlooked. I am grateful to Alison Fridley, Ayodele Gilbert Ogunkoya, Beth Hopkins, Courtney Erikson, Ebele Nwosu, James Ely, Onyinye Onwukwe, Walker Hopkins, and Yemi Adebayo for their friendship and mentorship during this journey. My colleagues at Hanovia Limited played a significant role during this journey. I would like to give special thanks to Segun Oguntoyinbo, Remi Oguntoyinbo, Gbenga Adedayo, Funmi Adedayo, Omotola Joseph, Onyinyechukwu Onwuka du-Bruyn, Patrick Okonkwo, and Sandra Amoh for their support and encouragement.

The constant support, prayers, and encouragement from my family provided the

springboard to get this far. My mother, Uchenna, and my siblings, Chinonye, Okechukwu, Chioma, Ifeoma, Onyinye, Onyeka, Obianuju, and Kosisochukwu, share in my accomplishments. I love and cherish them.

Dedication

This dissertation is dedicated to my late father, Chukwuemeka Wilson Ezike, who is not here to see this day.

Table of Contents

List of Tables

List of Figures

CHAPTER 1:

INTRODUCTION

In education and psychology, measurement models describe characteristics and behaviors of people and objects, including the proficiency level of examinees, the difficulty level of items, and the behaviors of human raters. Often, human raters provide numerical scores for examinees, introducing another consideration for measurement models. Assessments requiring human raters are subject to measurement errors, which, among other factors, could be due to the subjectivity of raters or an unclear rating guide. To ensure that examinee characteristics or performances are accurately described, measurement models that compensate for rater effects are needed when human raters are employed. The misspecification of such measurement models may threaten the validity of inferences made using the model parameters. Consequently, model-checking is fundamental to ensure that valid and meaningful inferences are made. The main purpose of this study is to evaluate the fit of the hierarchical rater model (HRM; Patz et al., 2002) to data (i.e., absolute model-data fit) under varying conditions. The HRM is increasingly being applied in educational and psychological measurement fields to model data from performance assessments. This introductory chapter first introduces performance assessments followed by a description of approaches to model performance assessments. The latter parts of this chapter briefly describe the HRM, and the Bayesian techniques applied to evaluate absolute model-data fit of the HRM. Finally, the research questions are provided.

1.1 Overview of Performance Assessments

In educational testing, measuring examinees' mastery of a subject or learning objectives may come from exams, assignments, quizzes, or performing tasks such as conducting a laboratory experiment and participating in a debate (Lane & Stone, 2006). Many educators

traditionally rely on multiple-choice test items to assess examinee learning. Multiple-choice test items are graded as either correct or incorrect and include predefined lists of possible correct options. Items on multiple-choice tests can range from easy to hard in terms of the item difficulty level and from low to high with regards to how well the items discriminate between low achieving examinees and high achieving examinees.

Wiggins (1998) suggested that the essence of educative assessment is to give tests in which the examiner can learn whether the examinees can intelligently use what they were taught rather than selecting answers from easy-to-score questions. These types of assessments typically require some kind of *performance*, which could be in the form of writing or showcasing a skill. Assessments that require performance are generally referred to as *performance assessments*. Performance assessments can range from tasks on writing (Weigle, 2010), creativity (Hung et al., 2012; Primi et al., 2019), or musical performances (Wesolowski et al., 2015). For example, an examinee may be tasked with writing an essay to describe their experience abroad, or a more challenging task such as designing and implementing a laboratory experiment. These performances challenge the examinees to perform simple or complex tasks rather than selecting answers from predefined lists. Performance assessments provide examinees the opportunity to use their initiatives, judgments, and knowledge to tackle assigned problems (Wiggins, 1998). They also provide examinees the opportunity to collaborate with other examinees.

Literature has argued that performance assessments provide rich information that is not directly measured by multiple-choice items (Johnson et al., 2008; Priestley, 1982; Wiggins, 1990). Performance assessment asks the examinees to use their judgment in innovative ways to effectively solve problems. An examinee's strategies of formulating and solving the performance tasks provide the potential of diagnosing the strengths and weaknesses of the examinee (Lane,

2010). Hence, performance assessments provide information about what examinees know and how they apply their knowledge. It also provides an avenue to give direct feedback about the examinees.

There are drawbacks; performance assessments require recruiting and training human raters and are time-consuming and expensive to score (Lane & Stone, 2006; Myford & Wolfe, 2003; Popham, 2003; Wainer & Thissen, 1993). For example, Wainer and Thissen (1993) reported that scoring the constructed-response portion of the Advanced Placement (AP) Chemistry test costs $3 to $4 per item compared to 1¢ for the entire multiple-choice section. This implies that the costs associated with scoring performance assessments could be as high as $40 for each examinee for a test with 10 constructed-response items. Another important consideration is the number of raters required to score each constructed-response item. Typically, multiple human raters are employed to judge the quality of examinees' responses. The use of multiple raters to rate the same examinee is to facilitate the increase in reliability and reduce the error associated with using a single rater (Eckes, 2011; Johnson et al., 2000). Rating each examinee's work by more than one rater is especially crucial if the results have consequences for the examinees, especially in high-stake testing (Bock et al., 2002). Unfortunately, human raters are fallible, which may result in unwanted variability in scores between raters who rate the same examinee (Linacre, 1989; Wilson & Case, 2000). Raters may lack consistency in how they apply the rating scale. For example, two raters who received the same training may be differential in how they apply the scoring rubrics (Weigle, 1998). In some instances, some raters may have more experience in the content area than others. Therefore, failing to account for the characteristics of different raters may lead to biased inferences concerning examinees' proficiency in the skills being evaluated.

1.2 Modeling Performance Assessments

Psychometric modeling in educational measurement involves important considerations such as the amount of measurement error and reliability (AERA, APA, & NCME, 2014; Wang et al., 2000). Reliability captures the extent to which measurements are free of measurement error (Perron & Gillespie, 2015). Traditionally, inter-rater reliability has been used to measure the extent of agreement among raters. Several techniques for computing inter-rater reliability have been proposed, including "percentage agreement," Cohen's kappa (1960), and weighted kappa (Cohen, 1968). These approaches are often employed to gauge rating quality (Saal et al., 1980). Ratings awarded by multiple raters are typically assumed to be independent, and rating consensus is desirable. Assessments with high rater consensus will yield a high inter-rater reliability. Unfortunately, traditional inter-rater reliability coefficients, such as percentage agreement and weighted kappa, do not account for rater effects and other sources of variability such as items and occasions.

Consider a situation where 10 raters are employed to each rate 500 examinees on an ethical reasoning essay. Raters may become fatigued from long hours of rating and may introduce errors in the rating operation (Myford & Wolfe, 2003). There is also a possibility that some of these raters may have different interpretations of the scoring rubric, despite being trained and calibrated. A culmination of these rater effects (i.e., subjective interpretation of the rating scale and rater fatigue) may lead to raters awarding ratings that are less than or greater than the true scores of the examinees. Scholars have used the term *leniency* to refer to raters who rate above examinee true scores and *severity* to refer to raters who rate below examinee true scores (Eckes, 2011; Myford & Wolfe, 2003, 2004). Many traditional inter-rater reliability

coefficients may fall short in modeling the data resulting from this illustration. When using these inter-rater reliability indices, a high degree of agreement among raters does not necessarily imply a high degree of accuracy in the ratings (Eckes, 2011). This is especially true when raters have similar levels of leniency or severity. Therefore, it is important to employ measurement models that account for the differences in raters' characteristics.

Significant advances have been made in modeling ratings of an examinee's work by multiple raters. The many-facet Rasch measurement (MFRM; Linacre, 1989) model is one of the foremost and popularly applied models that account for rater effects. The MFRM is an extension of the traditional Rasch model. The MFRM provides for measures [parameters] to account for potential sources of variability such as raters, scoring criteria, and rating methods (Eckes, 2011). These sources of variability are referred to as *facets*. For example, a four-facet MFRM model would be an assessment in which each examinee responds to different performance assessment items, and each item was judged by multiple raters using two different types of scoring rubrics. In this situation, examinees, items, raters, and scoring methods are the four facets of interest. In this example, the examinee ability, item, rater, and scoring method parameters can all be estimated using the MFRM model.

MFRM has often been criticized for its failure to account for the dependence structure of the ratings. The MFRM treats raters as locally independent experts such that each rater has a unique perspective of what the "true" rating is (Linacre, 2003). This implies that more measurement information is produced by increasing the number of ratings of the same person by multiple raters. The implication of this is that as the number of raters per item increases, the MFRM appears to give infinitely precise measurement of the examinee's latent proficiency, even though they answered no more items (Patz et al., 2002). Despite this criticism of the MFRM, the

model remains very popular among practitioners.

Other recent measurement models for calibrating performance assessments include HRM, rater bundle model (Wilson & Hoskens, 2001), hierarchical rater model-signal detection theory (HRM-SDT; DeCarlo, 2008, 2010; DeCarlo et al., 2011), and generalized rater model (Wang et al., 2014). One common theme across these models is their ability to model rater behaviors. At the kernel of most of these modeling approaches is the extension of item response theory (IRT; Hambleton & Swaminathan, 1985) models to capture rater effects. Unlike the MFRM, the HRM and HRM-SDT models are not limited to Rasch models.

This study focuses on the HRM which accounts for rater effects such as bias (i.e., award scores that are below or above the examinees' true scores) and variability (i.e., consistency/inconsistency in the use of the rating scale). The HRM models the hierarchy that exists in rating data by combining polytomous IRT-based and generalizability theory (Cronbach et al., 1972; Shavelson & Webb, 1991) modeling approaches. The HRM incorporates the dependence structure in ratings awarded to the same examinee by multiple raters. The HRM models the data from multiple ratings using a two-stage measurement process. The first level models the observed ratings given the *ideal* ratings using a discrete signal-detection-like model. In the second level, the HRM models ideal ratings given the examinee latent abilities using a polytomous IRT model.

In HRM, estimation of item, examinee ability, and rater parameters can be accomplished by employing a Bayesian Markov Chain Monte Carlo (MCMC; Patz, 1996) algorithm. Hombo and Donoghue (2001) have also estimated parameters of the HRM using marginal maximum likelihood (MML; Bock & Lieberman, 1970; Bock & Aitkin, 1981) estimation method. Bayesian modeling has remained attractive because of its capability to model complex designs comprising

complex dependency structures (Fox, 2010; Gilks et al., 1995). This makes Bayesian estimation a natural home for HRM because of its hierarchical structure and parameterization. Current literature that applied the HRM has employed Bayesian MCMC estimation techniques (see Casabianca & Wolfe, 2017; Casabianca et al., 2017; Nieto & Casabianca, 2019).

When applying the HRM, users are allowed to choose a polytomous IRT model at the second level of the HRM. The polytomous IRT models are described in the next chapter. Unfortunately, the effects of misspecifying the polytomous IRT component in the HRM have not been documented in literature. Because models are only approximations to reality, it is imperative to assess how well a posited model fits the observed data. In traditional IRT models, model misspecifications have been documented to lead to biased examinee latent trait estimates (Feuerstahler, 2018), and low statistical power of fit indices (Ames et al., 2020; Orlando & Thissen, 2000). In the context of the HRM, valid inferences about the examinees and accurate estimation of rater behaviors (i.e., severity/leniency and consistency) must be assured. To this end, model-checking is fundamental to ensure that valid and meaningful inferences are made.

1.3 Purpose of the Study

The conclusions and usefulness of the results from a measurement model are dependent on the extent to which the model accurately reflects the data (Orlando & Thissen, 2000). In IRT, adequate model-data fit has two main advantages: (1) parameter estimates of the items are not dependent on the samples of examinees drawn from the population of examinees for whom the test was designed for; (2) expected values of examinee ability estimates are not dependent on the choice of test items (Hambleton & Swaminathan, 1985). These two desirable features hinge on the adequate fit of the model to the data and should not be overlooked. Therefore, to ensure that the model parameters are adequately estimated, it is essential to evaluate various aspects of

model-data fit (Sheng, 2017).

The main purpose of this study is to evaluate absolute model-data fit of the HRM under varying conditions using posterior predictive model checking (PPMC; Gelman et al., 1996; Meng, 1994; Rubin, 1984) techniques. The PPMC is a Bayesian technique for assessing absolute model-data fit. Using a Bayesian paradigm, Gelman et al. (1996) summarized three approaches of evaluating model-data fit, which they gave as: (1) examining the sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable, given the substantive context of the model; and (3) checking that the model fits the data. This study will focus on the last approach of checking model-data fit.

The PPMC is implemented to evaluate whether certain aspects of the data are not captured by the model. In PPMC, data are replicated from the joint posterior distribution conditional on the observed data. The replicated data based on the posterior predictive samples are then compared to the observed data. Data replicated can be graphically illustrated or numerically quantified. Any systematic differences between features of the replicated and observed data indicate a failure of the model to explain those aspects of the data (Sinharay et al., 2006). One of the drawbacks of graphically comparing observed and replicated data is that some key features of the data may not be easily noticeable. Thus, more quantifiable techniques using discrepancy measures and posterior predictive *p*-values (PPP-values) allow for a more direct evaluation of the discrepancy between the observed data and posited model (Gelman et al., 1996; Meng, 1994). There is no limit to the number of discrepancy measures that can be applied in PPMC, however, the choice of the discrepancy measure is important to ensure that misfits are detected.

The PPMC approach has been used to assess model-data fit in traditional IRT models

(Ames, 2015, 2018; Hoijtink, 2011; Levy, 2006; Levy et al., 2009; Sinharay et al., 2006). The HRM has not benefited from rigorous assessment of absolute model-data fit using PPMC. Previous studies (Casabianca et al., 2017; Nieto & Casabianca, 2019) have applied PPMC to two extensions of the HRM: longitudinal-HRM (L-HRM) and multidimensional-HRM (M-HRM). These two studies applied only two discrepancy measures (total score and rater variability). In addition, none evaluated the utility of PPMC in detecting misfit of the HRM when the model is misspecified. In other words, these studies did not assess the power of the PPMC technique in detecting misfit if the HRM were misspecified. This study intends to explore the effectiveness of different discrepancy measures in detecting model-data fit when the HRM is correctly and incorrectly specified. Specifically, the study will attempt to answer the following research questions:

Research Question 1

What is the Type I error rate and power of the test-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?

Research Question 2

What is the Type I error rate and power of the item-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?

Research Question 3

What is the Type I error rate and power of the rater-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?

1.4 Definition of Key Terms

The definitions of the terms provided in this section is aimed at assisting the reader to understand the context of each term used in this study.

Discrepancy Measure. This is a statistic computed from the observed (or replicated) data. Discrepancy measures could be summary statistics such as measures of center, variability, and shape. It could also be measures such as correlation and reliability coefficients.

Hierarchical Rater Model. A measurement model for calibrating data from performance assessments. The hierarchical rater model models the hierarchy that exists in rating data by accounting for rater severity and variability. In the first stage, an IRT model describes the relationship between ideal ratings and observed ratings using signal-detection-like model, while the second stage describes the relationship between the ideal ratings and latent ability traits using a polytomous IRT model.

Ideal Rating. This is an examinee's true rating based on the quality of the examinee's work. The ideal rating is an examinee's true score if the examinee were rated by a rater who is unbiased and has high consistency in the use of the rating scale.

Item Score. The average rating given to an examinee by multiple raters on a particular item.

Item-Total Correlation. The correlation between a particular item's score (defined above) and the total score (which is the sum of the item scores). The item-total correlation is regarded as a discrimination index. It indicates how well items discriminate between high-performing and low-performing examinees.

Performance Assessment. This is an assessment of tasks performed by examinees to measure the degree to which the examinees apply their skills and knowledge to problems.

Posterior Predictive Distribution. This is a distribution of future observable data given the observed data.

Posterior Predictive Model Checking. This is a Bayesian technique used for assessing absolute model-data fit using replicated data simulated from the posterior predictive distribution.

Posterior Predictive P-value. The probability that the simulated or replicated data could be more extreme than the observed data as measured by the discrepancy measure.

Rater Leniency. A rater is considered a lenient rater if the rater rates well above the true scores of the examinees.

Rater Score. The average rating awarded by a particular rater to an examinee across all items.

Rater Severity. A rater is considered a severe rater if the rater rates well below the true scores of the examinees.

Rater-Total Correlation. The correlation between a rater's score and the total score, where the total score is the sum of the average rating given to an examinee by multiple raters across all items.

Score-Estimate Correlation. The correlation between a rater's score and the ability estimates of the examinees.

Total Score. The sum, across all items, of the average item ratings given to an examinee by multiple raters.

1.5 Chapter Summary

Chapter 1 began with an overview of performance assessments. Several examples of performance assessments were described in the early part of this chapter. In addition, some of the drawbacks of performance assessments were outlined. Traditional approaches of describing the

extent to which raters agree, using inter-rater reliability, were discussed. A brief discussion of some of the approaches of modeling performance assessments was provided, including MFRM and HRM. The main advantages of adequate model-data fit, and the purpose of the study were further described in this chapter. The latter part of this chapter provided the research questions in addition to the definition of the key terms used in this study.

CHAPTER 2:

LITERATURE REVIEW

The first part of this chapter highlights traditional IRT models for dichotomous and polytomous items and modeling approaches for performance assessments. This chapter also discusses rater effects and rating designs. The later part of this chapter introduces Bayesian estimation and model-data fit including an extensive description of the Bayesian techniques for assessing absolute model-data fit using PPMC.

2.1 Models for Polytomous Responses

In educational testing, items can be dichotomously scored in a binary format (i.e., scored as either correct or incorrect) or polytomously scored (i.e., three or more score points). When items are scored using a binary format, latent trait measurement models such as the one-parameter logistic (1PL; Rasch, 1960), two-parameter logistic (2PL; Birnbaum, 1958, 1968), and three-parameter logistic (3PL; Birnbaum, 1968) models are frequently applied to describe the probability of an examinee correctly responding to an item given the item parameters and the examinee's latent ability. In the 1PL model, only the item difficulty and examinee latent ability parameters are estimated. The 1PL assumes that all items have the same discrimination value (i.e., equal slope) and no guessing parameter. The 2PL model relaxes the restrictive assumption of the 1PL model by allowing the estimation of the discrimination parameter of each item. Finally, the 3PL model estimates the difficulty, discrimination, and examinee latent ability parameters. In addition, the 3PL model includes a guessing parameter to model the effect of examinees selecting the correct response option based on guessing.

Items with three or more response categories are referred to as polytomous items. These types of items are common in educational assessments or psychological measures. For example,

a school administrator might administer a teacher satisfaction survey. The items on the survey may elicit a teacher's responses to several items by endorsing each item on a 5-point Likert-type scale (e.g., 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree). When items are scored polytomously, the probability of an examinee (or subject) attaining a score category can be modeled by polytomous IRT models. The partial credit model (PCM; Masters, 1982), the generalized PCM (GPCM; Muraki, 1992), and the graded response model (GRM; Samejima, 1969) are some of the commonly applied polytomous IRT models. These three models are characterized by how they define their step functions using either the adjacent category approach or the cumulative approach and the constraints placed on the parameters (Penfield, 2014).

Partial Credit Model

The PCM is designed to model responses where partial credit (or full credit) is awarded to examinees based on the quality of work. The PCM is also appropriate for calibrating psychological measures where respondents endorse their beliefs on a given scale. The PCM belongs to the family of Rasch models (Rasch, 1960) and can be thought as an extension of the 1PL model. In the PCM, the discrimination parameter for all items is constrained to 1.

The PCM specifies the probability of examinee *i* endorsing item *j*'s *k*th category given the examinee's latent ability, and item step difficulty parameters. The item response function (IRF) for the PCM can be mathematically expressed as

$$P\left(X_{ij} = k \middle| \theta_i,\ \delta_{jv}\right) = \frac{\exp\left[\sum_{v=0}^{k}\left(\theta_i - \delta_{jv}\right)\right]}{\sum_{c=0}^{K-1}\exp\left[\sum_{v=0}^{c}(\theta_i - \delta_{jv})\right]}, \tag{2.1}$$

where δ_{jv} is the step difficulty parameter of item *j* with $\sum_{v=0}^{0}\left(\theta_i - \delta_{jv}\right) \equiv 0$, K is the number of category of item *j*, and θ_i represents the latent ability of examinee *i*. In terms of log odds, the PCM is given by

$$\ln\left[\frac{P\left(X_{ij}=k\middle|\theta_i,\delta_{jv}\right)}{P\left(X_{ij}=k-1\middle|\theta_i,\delta_{jv}\right)}\right]=\ \theta_i-\delta_{jv}. \tag{2.2}$$

The PCM is defined using the adjacent category approach where the step function uses adjacent pair of score categories. Therefore, δ_{jv} is the relative difficulty parameter of each step. In essence, δ_{jv} indicate where on the examinee latent trait continuum the score of one category is more likely than the previous category. An important distinction of the PCM is that the δ_{jv} parameters do not necessarily have to be ordered. In the PCM, we can express (and obtain) the probabilities associated with each score point. For example, assuming examinees were scored on a scale of 0 to 2, then the probabilities of obtaining scores of 0, 1, and 2 are presented in Equations 2.3, 2.4, and 2.5, respectively. The resulting probabilities depend on the examinees' latent ability and item step difficulty parameters.

$$P\left(X_{ij}=0\middle|\theta_i,\delta_{jv}\right)=\ \frac{1}{1+exp\left[\left(\theta_i-\delta_{j1}\right)\right]+exp\left[\left(2\theta_i-\delta_{j1}-\delta_{j2}\right)\right]}, \tag{2.3}$$

$$P\left(X_{ij}=1\middle|\theta_i,\delta_{jv}\right)=\ \frac{exp\left[\left(\theta_i-\delta_{j1}\right)\right]}{1+exp\left[\left(\theta_i-\delta_{j1}\right)\right]+exp\left[\left(2\theta_i-\delta_{j1}-\delta_{j2}\right)\right]}, \tag{2.4}$$

$$P\left(X_{ij}=2\middle|\theta_i,\delta_{jv}\right)=\ \frac{exp\left[\left(2\theta_i-\delta_{j1}-\delta_{j2}\right)\right]}{1+exp\left[\left(\theta_i-\delta_{j1}\right)\right]+exp\left[\left(2\theta_i-\delta_{j1}-\delta_{j2}\right)\right]}. \tag{2.5}$$

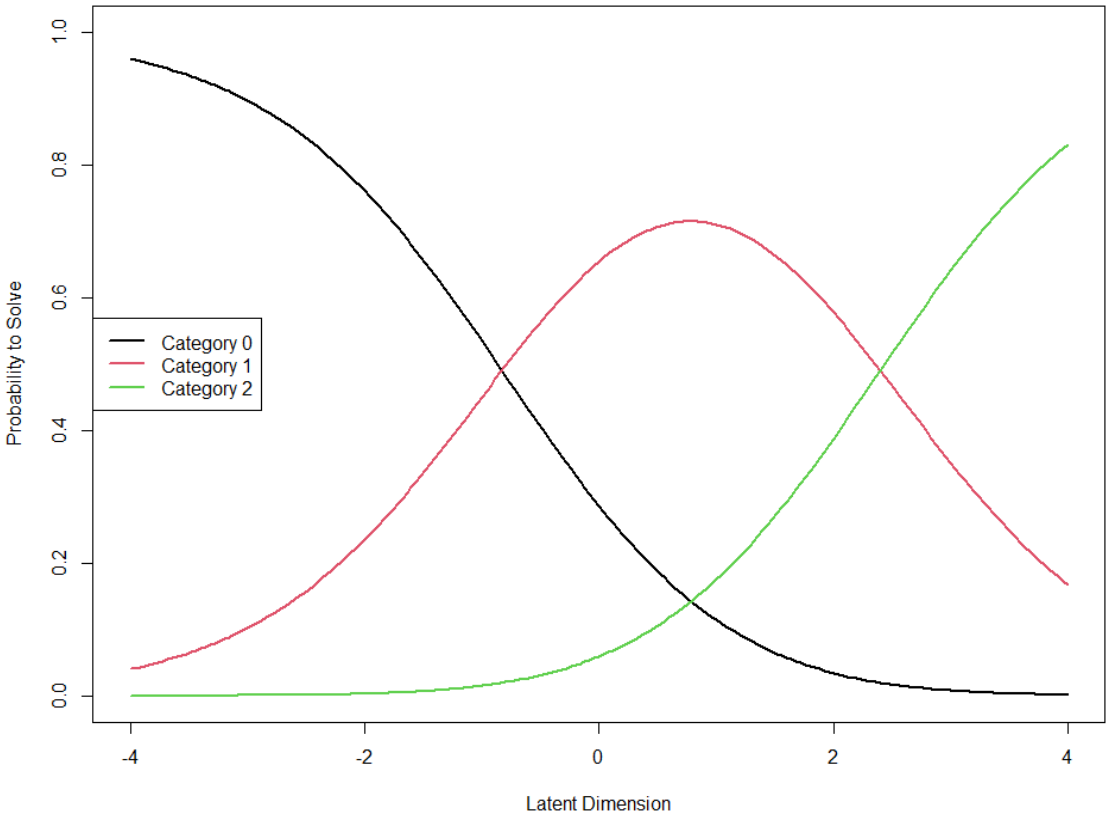

Figure 1. Category response curves for a hypothetical example using partial credit model

Continuing with this example, an example of the category response curves for a hypothetical item scored on a scale of 0 to 2 is illustrated in Figure 1. The step parameters for this example are $\delta_{j1} = -0.867$ and $\delta_{j2} = 2.202$. From Figure 1, it can be seen that an examinee with a latent ability of $\theta = 3.00$ is most likely to receive a score of 2 and an examinee with a latent ability of $\theta = -3.00$ is most likely to receive a score of 0. Specifically, the probability of an examinee with a latent ability of $\theta = -3.00$ obtaining a score of 2 is 0.005 (computed using Equation 2.5), however, the probability of this examinee with latent ability of $\theta = -3.00$ obtaining a score of 0 is 0.894 (computed using Equation 2.3).

Generalized Partial Credit Model

As the name implies, the GPCM (Muraki, 1992) is a generalization of the PCM. Under the GPCM, items within a scale are allowed to have different discrimination parameters. Thus, a discrimination parameter is estimated for each item. Like the PCM, the GPCM is also defined using the adjacent category approach. The IRF for the GPCM can be mathematically expressed as

$$P\left(X_{ij} = k \middle| \theta_i, \alpha_j, \delta_{jv}\right) = \frac{\exp\left[\sum_{v=0}^{k} \alpha_j\left(\theta_i - \delta_{jv}\right)\right]}{\sum_{c=0}^{K-1} \exp\left[\sum_{v=0}^{c} \alpha_j(\theta_i - \delta_{jv})\right]}, \tag{2.6}$$

where α_j is the discrimination parameter of item *j*, δ_{jv} is the step difficulty parameter of item *j* with $\sum_{v=0}^{0}\left(\theta_i - \delta_{jv}\right) \equiv 0$, K is the number of categories of item *j*, and θ_i is the latent ability of examinee i. Just like the PCM, the δ_{jv} represents the difficulty of the step of moving from one response category to another response category. The δ_{jv} can be decomposed into item threshold parameters (τ_{jv}) and item location parameter (β_j). The decomposition is given as $\delta_{jv} = \beta_j - \tau_{jv}$.

Graded Response Model

The GRM is applied to polytomous items with two or more ordered categorical responses. The GRM models the probability of an examinee obtaining a score at or above each item score category. The GRM is an extension of the 2PL model. Samejima (1969) developed a two-step process for computing the probability that an examinee receives a certain score given the item characteristics and examinee latent ability.

In the first step, the IRF of an examinee receiving a score of *k* or higher is mathematically expressed as

$$P_{jk}^{*}\left(X_{ij} \geq k \middle| \theta_i, \alpha_j, \delta_{jk}\right) = \frac{\exp\left(\alpha_j\left(\theta_i - \delta_{jk}\right)\right)}{1 + \exp\left(\alpha_j\left(\theta_i - \delta_{jk}\right)\right)}, \tag{2.7}$$

where α_j is the discrimination parameter, δ_{jk} is the threshold parameter for category *k* of item *j*, and θ_i is examinee *i*'s latent ability. The expression in Equation 2.7 is the 2PL model for a series of dichotomies. For example, for an item with three ordered response categories, the possible series of dichotomies are 0 vs. 1, 2 and 0, 1 vs. 2.

In the second step, the marginal probabilities are computed by taking the difference between the cumulative probabilities of the adjacent categories. This is mathematically expressed as

$$P_{jk}(\theta) = P^*_{jk}(\theta) - P^*_{j(k+1)}(\theta). \tag{2.8}$$

Equation 2.8 is equivalent to expressing the GRM as:

$$P\left(X_{ij} = k \middle| \theta_i, \alpha_j, \delta_{jk}\right) = \frac{\exp\left(\alpha_j\left(\theta_i - \delta_{jk}\right)\right)}{1 + \exp\left(\alpha_j\left(\theta_i - \delta_{jk}\right)\right)} - \frac{\exp\left(\alpha_j\left(\theta_i - \delta_{jk}\right)\right)}{1 + \exp\left(\alpha_j\left(\theta_i - \delta_{j(k+1)}\right)\right)} \tag{2.9}$$

For the three ordered response category example, the probabilities of responding in the three categories can be computed using:

$$\begin{cases} P_{j0} = 1 - P^*_{j1}(\theta) \\ P_{j1} = P^*_{j1}(\theta) - P^*_{j2}(\theta) \\ P_{j2} = P^*_{j2}(\theta) - 0 \end{cases} \tag{2.10}$$

2.2 Models for Performance Assessments

Performance assessments are typically polytomously scored. For instance, examinees may be asked to respond to an essay about their first trip abroad. Scoring rubrics employed to assess essay quality are usually designed to communicate expectations about the features of the essay. These rubrics could be on a scale of 0 to 3 (e.g., 0 = poor essay quality, 1 = fair essay quality, 2 = good essay quality, 3 = excellent essay quality). In performance assessments, multiple human raters typically judge the quality of an examinee's work.

Most traditional IRT models assume that an examinee's response to an item is

independent of the examinee's response to any other item given the examinee's ability, an assumption referred to as local independence (Hambleton & Swaminathan, 1985). Mathematically, the assumption of local independence is expressed as:

$$P\left(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J | \theta_i\right) = \prod_{j=1}^{J} P\left(X_j = x_j \middle| \theta_i\right), \tag{2.11}$$

where θ_i is the latent ability of examinee i and x_j is the response of the item response variable X_J. The expression in Equation 2.11 is referred to as strong local independence (McDonald 1994; Stout, 2002). In traditional IRT models, the presence of local item dependence could lead to the overestimation of item discrimination parameters and the accuracy of the parameter estimates (Mislevy et al., 2012).

The statistical property of local independence is the reason why the sum of the item information functions is the test information function (Verhelst & Verstralen, 2001). This implies that the test information function is a result of the independent contributions of the item information functions. The test information function plays an important role when a test administrator wants to select items that can provide a desired level of accuracy at specific regions of the examinee ability scale (Hambleton & Swaminathan, 1985; Park, 1997).

Although the polytomous IRT models discussed in Section 2.1 (i.e., PCM, GPCM, and GRM) are specifically designed to calibrate polytomous items; analyzing data from performance assessments with these traditional polytomous IRT models may lead to biased estimates due to the violation of local independence (Bock et al., 2002; Verhelst & Verstralen, 2001; Yens, 1993). This is due to the dependency in the rating structure of multiple ratings on the same examinee assessment artifact. In particular, multiple ratings of the same examinee will be correlated (Verhelst & Verstralen, 2001). Yen (1993) highlighted 10 possible sources of local item

dependency, which include (1) interference or external assistance (e.g., assistance from an instructor), (2) speededness, (3) fatigue, (4) practice, (5) item or response format, (6) passage dependence, (7) item chaining, (8) explanation of previous answer, (9) scoring rubrics or raters, and (10) content, knowledge, and abilities. The aspect that relates to performance assessments the most is (9) scoring rubrics or raters. Yen (1993) explained that items scored with the same scoring rubric may exhibit local dependence when specific demands are placed on the raters or because the items measure common constructs in the examinee. Furthermore, some performance assessment scoring designs only require specific raters to score a subsection of the test. When multiple ratings are elicited and there are uncontrolled rater effects, then the local independence assumption may be violated (Yen, 1993).

One potential approach to correcting the issue of local dependence is to consider each rater's ratings as separate items. For example, consider a test with three items (Items 1, 2, and 3) rated by two raters (Raters 1 and 2). This will result in six ratings (2 raters x 3 items). These six ratings could be treated as separate items. However, as observed by Bock et al. (2002), the resulting standard error of estimates would be underestimated if multiple ratings were treated as separate items.

Another approach could be to sum or average multiple item ratings of an examinee to create a single item score. Thereafter, the single item scores are analyzed using polytomous IRT models (e.g., PCM or GPCM). Consider the hypothetical illustration in Table 1. In Table 1, three raters rated five examinees on two items. Columns three to five of Table 1 gives the ratings of the three raters. For example, Examinee 5 was rated 4, 4, and 3 on Item 1 resulting in an average score of 3.67. Unfortunately, polytomous IRT models only allow for discrete values. As shown in Table 1, the average scores for some examinees are non-discrete values, which makes it

impossible to directly use any of the polytomous IRT models to fit the data in this situation. Even if the average scores were discretized, the resulting standard errors of the examinee ability estimates will still be underestimated (Song, 2019). Most importantly, summing across raters or the use of the average would result in loss of rater consistency information. Consequently, using an appropriate modeling approach is necessary to account for the dependency between multiple ratings of the same examinee. The MFRM and HRM are examples of measurement models that account for rating structures where multiple raters judge the same examinee's work.

Table 1. Hypothetical Distribution of Ratings by Rater and Examinee

Examinee	Item	Rater			Sum Score	Average Score
		1	2	3		
1	1	5	4	4	13	4.33
2		3	3	3	9	3.00
3		3	4	5	12	4.00
4		4	4	4	12	4.00
5		4	4	3	11	3.67
1	2	5	5	5	15	5.00
2		4	4	3	11	3.67
3		4	4	5	13	4.33
4		4	4	4	12	4.00
5		4	4	4	12	4.00

Many-Facet Rasch Measurement Model

Linacre (1989) introduced the MFRM, an extension of the traditional Rasch model, to account for other *facets* in rater-mediated assessments. These facets are other potential sources of variability such as raters, scoring criteria, rating methods, and many others (Eckes, 2011). The MFRM models the sources of variability on an additive logit scale. Essentially, the MFRM is a PCM that incorporates other facets. For instance, in a situation where the interest is to account for rater bias such as severity/leniency, the MFRM model can be expressed as

$$\ln\left[\frac{P\left(X_{ijr}=k\middle|\theta_i\right)}{P\left(X_{ijr}=k-1\middle|\theta_i\right)}\right]=\ \theta_i-\delta_{jk}-\ \phi_r\ , \tag{2.12}$$

where X_{ijr} is the observed ratings awarded to examinee *i* on item *j* by rater *r*, ϕ_r is the severity of rater *j* in how they rate the examinees, θ_i is the latent ability of examinee *i*, δ_{jk} is the step difficulty parameter, $P\left(X_{ijr}=k\middle|\theta_i\right)$ is the probability of examinee *i* being rated in category *k* on item *j* by rater *r*, and $P\left(X_{ijr}=k-1\middle|\theta_i\right)$ is the probability of examinee *i* being rated in category $k-1$ on item *j* by rater *r*.

One of the strengths of the MFRM is that the model allows for the estimation of interaction effects. For instance, the MFRM model can include rater-by-item interaction, which allows for investigating rater behavior across items. Also, if raters applied different types of rating methods, a rater-by-method interaction can be investigated (e.g., analytical versus holistic rubric). One of the limitations of the MFRM model is that it assumes that all items have the same discriminatory value. Also, the MFRM does not capture the consistency of raters in their ratings. In other words, the MFRM assumes that raters have the same consistency level. These assumptions are too restrictive and may not be satisfied in practice (Uto & Ueno, 2020).

As previously noted in this section, the multiple ratings awarded to examinees are not locally independent. Ignoring this assumption violation can downwardly bias the standard error (Mariano, 2002; Verhelst & Verstralen, 2001; Wilson & Hoskens, 2001). Although MFRM is one of the most widely applied performance assessment models, one of the criticisms of the MFRM is its failure to account for the dependence structure of the ratings. The MFRM treats raters as locally independent experts such that each rater has a unique perspective of what the "true" rating is (Linacre, 2003). This means that more measurement information is produced by increasing the number of ratings of the same person by multiple raters. Patz et al. (2002)

discussed that as the number of raters per item increases, the MFRM appears to give infinitely precise measurement of the examinee's latent proficiency even though they answered no more items. Patz et al. (2002) proposed the HRM as an attempt to address the dependencies between multiple ratings using the hierarchical structure of performance assessment data.

Hierarchical Rater Model

The HRM incorporates the dependence structure in ratings awarded to the same examinee by multiple raters. The model corrects the problem of downward bias of standard errors in the MFRM by breaking the data generation process into two stages (Patz et al., 2002). In the first stage, a hypothetical rating describing an examinee's performance on a particular item is given. In the second stage of the model, raters evaluate the quality of the examinee's work using prescribed rubrics. Hierarchically, the first level of the HRM models the distribution of ratings awarded given the ideal ratings (i.e., $X_{ijr}|\xi_{ij}$). The second level of the hierarchy models the examinees' responses given the latent ability traits (i.e., $\xi_{ij}|\theta_i$). The hierarchy of the HRM is expressed as

$$\theta_i \sim Normal(\mu, \sigma^2), \qquad i = 1, \dots, N \tag{2.13}$$

$$\xi_{ij}|\theta_i \sim \text{a polytomous IRT model}, \qquad j = 1, \dots, J \qquad \text{for each } i \tag{2.14}$$

$$X_{ijr}|\xi_{ij} \sim \text{a polytomous signal detection model}, \qquad r = 1, \dots, R. \qquad \text{for each } i, j \tag{2.15}$$

In the first expression above (Equation 2.13), the proficiency levels of the examinees are assumed to be normally distributed with mean (μ) and variance (σ^2). The distribution of examinee latent trait may not necessarily follow a normal distribution. ξ_{ij} represents the ideal rating for examinee *i*'s response to item *j*. The ideal ratings ξ_{ij} given the examinee latent trait θ_i (i.e., Equation 2.14) may be modeled using polytomous IRT models such as PCM, GPCM, and GRM. X_{ijr} represents the observed ratings awarded to examinee *i* on item *j* by rater *r*. The

observed ratings given the ideal ratings is modeled using a discrete signal detection model (i.e., Equation 2.15). The signal-detection-like model allows using a matrix of probabilities of the ratings to model the relationship between observed and ideal ratings. Table 2 presents an example of a matrix of probabilities between observed and ideal ratings for a five-category scale.

Table 2. Hypothetical Rating Probabilities Matrix for a Five-Category Scale

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}

The rating probability is defined as $p_{\xi_{kr}} \equiv P(rater\ r\ rates\ k\ |ideal\ rating\ \xi)$. For example, p_{20r} is the probability of rater *r* awarding a score of 0 given an ideal rating of 2. It is expected that the probabilities along the diagonal (i.e., $p_{00r}, p_{11r}, p_{22r}, \ldots, p_{kkr}$) will be largest. This probability (i.e., $p_{\xi_{kr}}$) is proportional to the normal density with mean of $\xi + \phi_r$ and standard deviation of ψ_r. This parameterization allows for the estimation of rater effects such as rater bias and variability. The rater probability proportional to the normal density is expressed as

$$p_{\xi_{kr}} \equiv P\left(X_{ijr} = k \middle| \xi = \xi_{ij}\right) \propto \exp\left\{-\frac{1}{2\psi_r^2}\left[k - (\xi + \phi_r)\right]^2\right\}, \tag{2.16}$$

where ϕ_r indicates the severity/leniency of rater *r*. This captures the rater *r*'s deviation from the ideal rating. Patz et al. (2002) termed ϕ_r as the rater's bias. Following the guideline provided by Patz et al. (2002), rater *r* is judged to exhibit *severity* relative to the ideal rating category if $\phi_r < -0.5$, which implies that rater *r* awarded ratings in some categories less than the ideal rating

category (i.e., $k < \xi$), whereas, rater *r* is exhibiting *leniency* if $\phi_r > 0.5$ (i.e., $k > \xi$). It is not clear if these suggested guidelines hold for narrower or wider rating scale categories (e.g., three-option or seven-option scales). A rater is considered more probable to award ideal ratings when $\phi_r = 0$. The parameter ψ_r represents the variability or lack of reliability of rater *r*. The inverse of a rater's lack of reliability provides the rater's precision (i.e., $\tau_r = 1/\psi_r$). Essentially, ψ_r indicates the rater's level of consistency in the use of the rating scale. A more consistent rater will have ψ_r value close to zero. Rater bias and variability statistics are to be interpreted together. For example, a rater with $\phi = -0.67$ and $\psi = 0.03$ is indicative of a consistently severe rater. This rater will be judged to consistently award ratings in some category less than the ideal rating. However, a rater with $\phi = 0.04$ and $\psi = 0.10$ is indicative of a rater who is likely to award ratings close to the true scores of the examinees in a consistent manner.

As the application of the HRM continues to rise, the model has been further expanded to cater to other facets or dimensions. To further understand how covariates influence rater bias and variability, Mariano and Junker (2007) extended the HRM to include other covariates of the rating process. These covariates could be random or fixed effects involved in the rating process. For instance, the rating occasion or the time (e.g., seconds) it took to complete the rating may be of interest to a testing program. Understanding the effects of these covariates on rater bias and variability may prove useful for adapting and improving features of rating designs (Casabianca et al., 2016).

The HRM has also been expanded to model performance assessments from longitudinal designs. The L-HRM models examinees' latent traits using an autoregressive time series model. The L-HRM estimates a common set of item parameters for all timepoints under the assumption that, over time, the items hold the same properties and relationship with the trait (Casabianca et

al., 2017).

Assessing the dimensionality of a construct is a crucial component when validating IRT models. This is especially important in performance assessments where more complex skills may be of interest. Factors such as the characteristics of the scoring schemes and characteristics of the stimulus may impact the dimensionality of the assessment (Lane & Stone, 2006). When these factors are present, analyzing the data using unidimensional performance assessment models may not be appropriate; instead, a multidimensional framework needs to be considered. Multidimensional IRT (MIRT) is used to model the relationship between two or more unobservable variables and the probability of the examinee correctly answering any particular test item (Ackerman et al., 2003). A multidimensional extension of the HRM was developed by Nieto and Casabianca (2019) to model multidimensional performance assessments with three or more response categories and allow for estimation of multiple latent ability simultaneously. The formulation of the M-HRM assumes a multidimensional latent structure and that rater behavior is not consistent across dimensions. However, Nieto and Casabianca (2019) also provided modeling options for a situation where rater behavior is consistent across dimensions.

The HRM-SDM is another extension of the HRM. In Patz et al.'s (2002) version of HRM, only rater severity/leniency and variability effects are modeled. Restriction of range or central tendency effect is another potential source of rater error (Myford & Wolfe, 2003, 2004) that is not captured in the original formation of the HRM. DeCarlo et al. (2011) expanded the HRM to model raters' tendency to favor middle categories by incorporating a latent class model that is motivated by the signal detection theory into the first level of the HRM.

2.3 Rater Effects

Assessments that require human judgments are subject to errors. Engelhard (1994), Myford and Wolfe (2003; 2004), Saal et al. (1980), and Wolfe (2014) described different ways that human raters can introduce errors in performance assessments. This includes effects due to (a) central tendency, (b) halo, (c) restriction of range, (d) severity or leniency.

Central tendency refers to when raters overuse the middle category of a scale avoiding the use of extreme categories (Myford & Wolfe, 2003). For example, on a 5-point rating scale, a rater might assign scores of 3 to most examinees and only few examinees getting low or high scores. Evidence of a rater exhibiting effect due to central tendency may be seen in the narrow standard deviation of the ratings award by this rater. When a rater fails to discriminate among conceptually distinct aspects of an examinee's behavior, then the rater is said to exhibit a *halo* effect (Saal et al., 1980).

Restriction of range is often grouped with central tendency effect. Saal et al. (1980) discussed these two sources of rater errors separately. Restriction of range refers to when raters' ratings are clustered or grouped around a part of the rating scale. As earlier described, with an example, the overuse of middle categories (e.g., "3") depicts raters' exhibiting central tendency. However, when the raters restrict their ratings around a certain score point (e.g., "1") then these raters could be exhibiting restriction of range effects.

Raters may also exhibit severity and leniency effects. Leniency refers to raters who rate above the true scores of the examinees and severity refers to raters who rate below the true scores of the examinee true scores.

Review of HRM Literature on Rater Effects

Rater effects have been extensively discussed in psychometric literature spanning different types of models for calibrating performance assessments. This section focuses on findings regarding rater effects in literature that applied the HRM. Patz et al. (2002) used simulation to compare the HRM and MFRM. Patz and colleagues generated data with the HRM using a fully-crossed rating design with 500 examinees, 3 raters, and 5 items. They employed the PCM as the IRT generating component of the HRM. In this instance, it is logical that the PCM was used as the IRT component of the HRM since the MFRM assumes that all items have equal discriminating power. The generated data were analyzed using the HRM and MFRM. As expected, the HRM and MFRM performed differently. First, the MFRM appeared to underestimate the variance of the examinee ability estimates. Patz and colleagues attribute this to the quality of the raters. Specifically, two out of the three raters in their simulation had large rater variability parameters. These large ψ_r parameters resulted in their simulated ratings to tend towards middle categories when the ideal ratings have extreme values. In addition, their results show that difficulty parameter estimates for the MFRM shrunk toward zero. This was especially true for items with extreme difficulty parameters such as -2 and $+2$. Specifically, of the five items, only one of the true item difficulty parameters was captured within the 95% confidence interval for the model estimated with MFRM. However, four of the five difficulty parameters were contained within the 95% credible interval for the model estimated with HRM. Most interestingly, the rater variability parameter for Rater 3 who had a true variability value of 0.06 was least recovered by the HRM. The posterior median for this rater was 0.01, although the 95% credible interval was found to capture the true parameter (i.e., 0.06). Patz and colleagues further illustrated the utility of the HRM using real data. The dataset they used consisted of 11

constructed-response items, 38 raters, and different rating designs from the Grade 5 Florida Math Assessment. Indeed, they showed that the HRM can detect rater behaviors such as rater bias and variability.

Casabianca et al. (2017) conducted two simulation studies to examine the L-HRM under varying conditions. In study one, they investigated parameter recovery of the L-HRM with varying sample sizes (250 and 500 subjects), number of raters (3 and 6 raters), and number of time points (3 and 7 time points). Using only five items, parameter recovery of the item, rater, and longitudinal model parameters and latent traits were investigated. Their findings of the rater parameters show small absolute biases of rater bias and rater variability parameters. Interestingly, they found that recovery of the rater variability parameter was most difficult when the rater variability parameters were less than 0.25. This finding parallels Patz et al. (2002) that also revealed difficulty in recovery of rater variability parameter when the true value is small. In study two, Casabianca and colleagues examined parameter recovery of the L-HRM by varying the growth (0.25, 0.50, and 0.75) and autocorrelation (0.00, 0.30, 0.60, and 0.90). The number of raters (10 raters), sample size (400 subjects), number of timepoints (4 time points), and number of items (10 items) were fixed. As in study 1, parameter recovery of item, rater, and longitudinal model parameters and latent traits were examined. Again, the results show that the rater bias and variability parameters were adequately recovered.

Nieto and Casabianca (2019) evaluated parameter recovery of the M-HRM under varying conditions. The authors simulated data using a double-scored design with 1000 subjects and 25 raters. The number of items (3 and 6 items), number of dimensions (2 and 4 dimensions), and correlation between dimensions (0.00, 0.40, and 0.80) were varied. Data for their simulations were generated with the M-HRM. Three analyses models were employed: M-HRM,

multidimensional GPCM (MGPCM), and consecutive HRM (Consecutive-HRM). The MGPCM does not account for rater effects and the Consecutive-HRM is akin to fitting unidimensional HRMs. The root mean squared error (RMSE) show that raters with small variability parameters and severity parameters close to zero had the smallest RMSE estimates. Noteworthy is that in their design, each subject was randomly assigned to two raters. In other words, each subject only received two ratings instead of ratings from all 25 raters. The implications of this type of rating design are presented next.

2.4 Rating Designs

One of the important aspects of performance assessments is the rating design. A crucial consideration is the number of raters needed to accurately estimate an examinee's proficiency level. In addition, the costs of employing multiple raters and other resources such as the time it takes for multiple raters to rate the same examinee's tasks are considered relative to how much precise measurement information is anticipated. It is often expected that higher information about the examinee will result in more precise measurement. The complete and incomplete rating designs are techniques that can be employed to obtain data in performance assessments (Eckes, 2011). In the *complete rating design*, ratings are awarded to every examinee by every rater on every item. An illustration of the complete rating design is shown in Table 3. This example depicts a scenario with 4 raters, 10 examinees, and 2 items. In this design, every rater is connected to every examinee and item. Although this fully crossed type of design is desirable, it is often expensive to implement especially in large-scale assessments with many examinees (e.g., the Graduate Records Examinee (GRE) and Test of English as a Foreign Language (TOEFL)) and potentially unrealistic due to time constraints.

In the *incomplete rating design*, raters only award ratings to a subset of examinees or

items. While incomplete rating designs are more practical, it is important that the design has enough links between raters and items to adequately estimate the model parameters. Linacre and Wright (2002) note that MFRM does not require a fully-crossed design but it is necessary that ratings are designed to create a network through which every item can be directly or indirectly linked to every other item. Incomplete rating designs could lead to *connected* or *disconnected* datasets. Connected designs provide sufficient systematic links between facets, whereas disconnected designs provide insufficient links between facets. The spiral rating design (Hombo et al., 2001) is an example of a connected design in which every rater rates every examinee only on a subset of items. An example of the spiral rating design is presented in Table 4. In this design, Raters 1 and 3 rated every examinee on Item 1, while Raters 2 and 4 rated every examinee on Item 2. This ensured that all items are systematically connected. Table 5 shows an incomplete rating design with disjointed subsets (Linacre, 2020). This structure consists of two subsets. In the first subsets, Raters 1 and 2 rated every item for a subset of examinees (Examinees 1 to 5), while Raters 3 and 4 rated every item for another subset of examinees (Examinees 6 to 10). In the disconnected incomplete rating design shown in Table 5, there is no overlap in the examinees rated by raters in the two subsets.

Deficient designs such as the disconnected incomplete rating design may introduce bias in parameter estimation and may pose issues with model-data fit. Linacre (2020) attribute the lack of connectedness to the accidental or deliberate manner in which the data was collected. For example, a practitioner who does not know the implications of these variants of incomplete rating design may design a rating process that is disjointed.

Table 3. An Example of a Fully Crossed Rating Design

Rater	Item	Examinee									
		1	2	3	4	5	6	7	8	9	10
1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: ✓ indicates that a rater rated an examinee on a particular item

Table 4. An Example of a Spiral Rating Design

Rater	Item	Examinee									
		1	2	3	4	5	6	7	8	9	10
1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2										
2	1										
	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2										
4	1										
	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: ✓ indicates that a rater rated an examinee on a particular item

Table 5. An Example of a Disconnected Rating Design

Rater	Item	Examinee									
		1	2	3	4	5	6	7	8	9	10
1	1	✓	✓	✓	✓	✓					
	2	✓	✓	✓	✓	✓					
2	1	✓	✓	✓	✓	✓					
	2	✓	✓	✓	✓	✓					
3	1						✓	✓	✓	✓	✓
	2						✓	✓	✓	✓	✓
4	1						✓	✓	✓	✓	✓
	2						✓	✓	✓	✓	✓

Note: ✓ indicates that a rater rated an examinee on a particular item

Studies on Rating Designs in Performance Assessments

The effects of rating designs have been explored by a number of researchers. Patz et al. (2002) explored the effects of the rating modality design using real data with 11 constructed-response items and 38 raters on the estimation of parameters using the HRM. The rating comprised of three rating designs referred to as Modality One, Modality Two, and Modality Three. In Modality One, raters scored the entire booklet of 11 items, whereas one item was assigned to each rater in Modality Two. In Modality Three, each rater was assigned to rate blocks of three to four items. Patz and colleagues found that Modality One had the smallest absolute modality bias in comparison to Modalities Two and Three. The 95% credible interval associated with the modality bias of Modality One did not overlap with the 95% credible intervals of Modalities Two and Three suggesting that Modality One, in which raters scored 11 items, was distinctly different in terms of the absolute modality bias.

Hombo et al. (2001) examined the impact of rating designs on examinee ability estimates. They compared three types of rating designs: fully crossed (which they called *crossed*), nested, and spiral. In the crossed design, 16,000 examinees were each rated by four raters on 20 items. This resulted in 64,000 ratings per item. The nested and spiral designs are examples of incomplete rating designs. Hombo and colleagues explored three and four variations of the nested and spiral designs, respectively. In each of the three nested designs, Hombo and colleagues assigned a single rater to rate all 16,000 examinees on each item. Thus, there were only 16,000 ratings per item, which is only 25% of the ratings observed under the crossed design. Also, in each variation of the spiral design, each rater was assigned a subset of the 20 items. Like the nested design, there were 16,000 item ratings under each variation of the spiral design that they explored. Data used in their study were generated using 1PL and 2PL models

that account for rater severity effects.

Hombo et al. (2001) found that spiral designs produced estimates that were reasonably accurate, and this design was robust to rater tendencies. Among the three designs (i.e., crossed, nested, and spiral), the nested design produced substantially larger biases in the examinee latent trait abilities. Noteworthy is that all three rating designs that Hombo and colleagues employed had sufficient links between items, examinees, and raters.

Wind et al. (2019) explored the impact of different types of rating designs for the classification of musical performances. Wind and colleagues employed four rating designs using simulations. Data were simulated for 250 performances (examinees) using the MFRM model with 25 raters. In the first design, all the raters scored all the performances (i.e., fully-crossed rating design). The second design was an incomplete design with each rater scoring one set of 10 performances in common with three other raters. The third design also had each rater score one set of 10 performances; however, the ratings were common with one other rater. Finally, in the fourth design, 24 raters did not score any performances in common, but their ratings were connected through one rater. Thus, all four rating designs had sufficient connections. Using a *decision consistency index* to classify performances, their findings suggest that classification consistency of the performances were highest in the first design, which was a fully-crossed design, and lowest in the fourth design, which ratings of other raters were only connected through one rater.

The lack of connectedness makes it difficult to compare examinees and raters who are in different subsets (Engelhard, 1997). For example, it is difficult to compare examinees and raters in the two subsets shown in Table 5. This issue could be mitigated by applying the group-anchoring technique when calibrating the data. Using the group-anchoring approach, Wind and

Stager's (2019) simulation found that disconnected designs impacted the precision of examinee estimates. Particularly, the correlations between the generating and estimated parameters of examinee latent ability were less than 0.40 in most instances. Their findings suggest that rating design is crucial to accurately estimate examinee proficiency levels.

2.5 Parameter Estimation of Hierarchical Rater Model

Accurately estimating parameters that reflect rater behaviors, item characteristics, and examinee proficiency levels are essential components of model building with regards to the HRM. There are four commonly used techniques for the estimation of parameters of IRT models: Bayesian estimation, MML, joint maximum likelihood (JML), and conditional maximum likelihood. The parameters of the HRM have been estimated with MML (Hombo & Donoghue, 2001) and Bayesian estimation with MCMC (Casabianca & Wolfe, 2017; Casabianca et al., 2017; Nieto & Casabianca, 2019; Patz et al., 2002). Bayesian estimation is the estimation method applied in this study and is introduced here.

Bayes' Theorem. Bayesian statistics involves the use of Bayes' Theorem to combine observed data and prior information to make inferences about parameters. Bayes' Theorem gives a framework for computing conditional probabilities. Assuming there are two events (A and B), by Bayes' Theorem, the conditional probability of A given B can be expressed as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2.17}$$

where $P(A|B)$ is the probability of event A given B, $P(B|A)$ is the probability of event B given A, $P(A)$ is the probability of event A, and $P(B)$ is the probability of event B.

Extending the Bayes' Theorem to modeling unknown model parameters, which are of interest in educational testing, we can consider the inference about model parameters based on observed data to be conditionally given as

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}, \tag{2.18}$$

where $P(\theta|x)$ is the probability of the parameter θ given the observed data x, $P(x|\theta)$ is the probability of observing the data given the θ parameter values, $P(x)$ is the "unconditional" probability of the data, and $P(\theta)$ is the "unconditional" probability of the parameters. The $P(\theta)$ is referred to as the prior distribution. The prior distribution expresses the uncertainty about θ. The denominator of Equation 2.18 is considered to be the marginal likelihood since it does not depend on the unknown parameters. Essentially, $P(x)$ sums or integrates over all possible values of the parameter θ. In a case of discrete parameters, $P(x)$ can be expressed as $\sum P(\theta)P(x|\theta)$ and $\int P(\theta)P(x|\theta)d\theta$ if the parameters are continuous. These expressions are referred to as a *normalizing constant*, which essentially makes the probability distribution integrate to 1 in the continuous case and sum to 1 in the discrete case. Consequently, Equation 2.18 can be rewritten as

$$P(\theta|x) \propto P(x|\theta)P(\theta)\,. \tag{2.19}$$

The conditional distribution of the parameter θ given the observed data, $P(\theta|x)$, is the posterior distribution. The posterior distribution captures the prior information and the likelihood of the data. Essentially, the posterior distribution reflects the plausibility of the parameter values given some prior information or knowledge and the observed data.

Bayesian techniques offer several advantages over frequentist estimation methods. Tsutakawa and Soltys (1988) and Tsutakawa and Johnson (1990) mentioned that frequentist estimation approaches such as MML ignore the uncertainty due to the error in calibration. Tsutakawa and colleagues noted that ignoring this could lead to inferential errors, especially with a small sample size. This uncertainty is captured in the Bayesian paradigm, which makes a Bayesian estimation method very attractive. Frequentist method such as MML estimation makes distributional assumptions about the ability parameter. In reality, this assumption may not be met. Woods (2014) summarized that simulation studies have shown that nontrivial biased item parameter estimates are produced when the probability density of the examinee latent ability is non-normal and nonsymmetric. Bayesian estimation methods have been shown to perform well when latent ability distributions are non-normal (Kieftenbeld & Natesan, 2012). The benefits of Bayesian estimations in the presence of small sample sizes were also demonstrated by Finch and French (2019). One of the offerings of Bayesian estimation methods is the ability to estimate parameters of highly parameterized models, although this comes with some costs. When working with complex models, it becomes difficult to directly obtain samples from the posterior distribution. Fortunately, this can be resolved using MCMC methodology.

Markov Chain Monte Carlo. It is often difficult to numerically derive the posterior distribution, $P(\theta|x)$, which is usually high-dimensional and may not be analytically feasible. For example, the computation of the posterior distribution may require integration over several parameters (e.g., item and ability) and complex parameter spaces that require constraining parameters. The MCMC methodology, a simulation-based approach, can handle the dimensionality problems of numerical integration (Fox, 2010).

The MCMC methodology combines Monte Carlo with Markov chain techniques. Monte Carlo uses simulation to estimate the characteristics of the distribution by drawing large random samples from a probability distribution, called a proposal distribution. On the other hand, a Markov chain is a stochastic process in that the probability of transitioning to the next state depends on the current state. The key is for the Markov process to converge to a stationary distribution. This is achieved by running a simulation long enough that the distribution of the current draws is close to the stationary distribution (Gelman et al., 2004). This process can be time-consuming for datasets with large number of examinees, items, and raters. Some commonly applied MCMC algorithms include the Metropolis-Hastings sampler (Hastings, 1970), Gibbs sampler (Geman & Geman, 1984), and the No-U-Turn sampler (Hoffman & Gelman, 2014). The features of these algorithms are extensively discussed in Gelman et al. (2004) and Hoffman and Gelman (2014).

2.6 Model-Data Fit for Traditional IRT Models

A measurement model is of little importance if it fails to communicate the story that the data is expected to convey. An accurate story is only told when all necessary model assumptions are satisfied, including that the model fits the data. Typically, the fit of a model is assessed by evaluating the discrepancy between the model-implied data structure and the observed data (Levy, 2006). Model-data fit indices are specific to the type and complexity of the model. From a measurement perspective, a good fitting model is a model that accurately estimates the true item parameters and the true abilities of the examinees. Essentially, the degree to which the properties of an IRT model are attained depends on whether the appropriate model is used (Hambleton & Swaminathan, 1985; Stone & Hansen, 2000). Traditional IRT fit statistics can be classified into person-fit statistics, test-fit statistics, and item-fit statistics. Item-fit statistics are useful in

assessing the fit of individual items, whereas test-fit statistics are used in assessing the fit of all the items as a whole (Stone & Zhu, 2015). On the other hand, person-fit statistics are examinee-specific.

An item may produce a poor fit when the wrong IRT model is employed to calibrate the item (Yen & Fitzpatrick, 2006). Most widely used fit indices adopt either the chi-square or likelihood ratio approach (Ames & Penfield, 2015). The Yen's Q_1 (Yen, 1981) and Yen's Q_3 (Yen, 1984), G^2(McKinley & Mills, 1985), $S - X^2$ (Kang & Cohen, 2008; Orland & Thissen, 2000), $S - G^2$ (Orland & Thissen, 2000), Bock's χ^2 (Bocks, 1960), *OUTFIT* and *INFIT* (Wright & Panchapakesan, 1969) are some of the popular model-data fit indices for both dichotomous and polytomous (the generalized forms of some of the indices) items. The sensitivity and Type-I error rates of many of these fit indices have been extensively investigated and their shortcomings have also been documented (Kang & Cohen, 2007, 2008; Orland & Thissen, 2000).

Yen's Q_1 Statistic

Yen (1981) proposed the Yen's Q_1 statistic to detect items that are poor fitting. The Q_1 is based on a chi-square approach. The Yen's Q_1 statistic is expressed as

$$Q_{1j} = \sum_{g=1}^{10} N_g \frac{\left(O_{jg} - E_{jg}\right)^2}{E_{jg}(1 - E_{jg})}, \tag{2.20}$$

where O_{jg} is the observed proportion of correct responses to item *j* for the examinees in group g, E_{jg} is the predicted proportion of correct responses to item *j* for the examinees in group g, N_g is the number of examinees in group g. From the formulation, it is seen that the Yen's Q_1 divides examinee ability scale into 10 groups. If the model is true, the Q_1 is distributed approximately as a χ^2 with degrees of freedom of 10 – *m*, where *m* the number of parameters estimated for item *j*. One of the drawbacks of the Yen's Q_1 statistic is that it is sensitive to sample size. A large

sample size will tend to increase the Q_1 statistic. In addition, Yen's Q_1 is dependent on the estimates of the examinee latent ability for group assignment. Studies have shown that Yen's Q_1 statistic results in high Type I error rates for short test length (see Ames et al., 2020; Kang & Chen, 2008; Orlando & Thissen, 2000). Fitzpatrick et al. (1996) provided a generalized form of the Yen's Q_1 for polytomous items.

Yen's Q_3 Statistic

Yen (1984) proposed the Yen's Q_3 statistic as a measure for detecting local dependence of items after accounting for the latent ability of the examinees. For dichotomously scored items, Yen (1984) defined, d_j, the difference between examinee's observed score on item *j* and the predicted score based on the IRT model as

$$d_j = X_j - E\left(X_j \middle| \hat{\theta}\right). \tag{2.21}$$

The Yen's Q_3 statistic is computed as the correlation between the deviation scores of two items. This statistic is given as

$$Q_{3jj'} = r_{d_j d_{j'}}, \tag{2.22}$$

where $r_{d_j d_{j'}}$ is the Pearson product moment correlation between the examinees' deviation scores on item *j* and item j'. Items *j* and j' are a pair of items of the same scale. Correlating the deviation scores essentially takes into account the examinees' abilities. Under the null assumption that sets of items are locally independent, if some sets of items are locally dependent, then they will have significantly high residual correlations (Yen, 1993).

G^2 Statistic

McKinley and Mills (1985) proposed the G^2 item-fit statistic, which is based on a likelihood ratio. The G^2 statistic is mathematically expressed as

$$G_j^2 = 2\sum_{g=1}^{10} N_g \left[O_{jg} \ln\left(\frac{O_{jg}}{E_{jg}}\right) + \left(1 - O_{jg}\right) \ln\left(\frac{1 - O_{jg}}{1 - E_{jg}}\right)\right], \quad (2.23)$$

where O_{jg} is the observed proportion of correct responses to item *j* for the examinees in group g, E_{jg} is the predicted proportion of correct responses to item *j* for the examinees in group g, N_g is the number of examinees in group g. Like the Yen's Q_1 statistic, the original formulation provided by McKinley and Mills (1985) divides individuals into 10 groups. If the model is true, the G^2 is distributed approximately as a χ^2 with degrees of freedom of *10 – m*, where *m* is the number of parameters estimated for item *j*. Orlando and Thissen (2000) note that the approach of grouping examinees into equal-size groups is sample dependent, which tends to affect the fit statistic.

$S - X^2$ and $S - G^2$ Statistics

Orlando and Thissen (2000) proposed $S - X^2$ and $S - G^2$ statistics based on the chi-square and likelihood ratio approaches, respectively. One distinct difference of their approach from Yen's Q_1 and G^2 is how Orlando and Thissen defined the groups. The groups for their proposed statistics are defined based on number of correct scores (i.e., observed test scores) instead of latent ability estimates. The $S - X^2$ is expressed as

$$S - X_j^2 = \sum_{g=1}^{P-1} N_g \frac{\left(O_{jg} - E_{jg}\right)^2}{E_{jg}(1 - E_{jg})}, \quad (2.24)$$

and $S - G^2$ is expressed as

$$S - G_j^2 = 2\sum_{g=1}^{P-1} N_g \left[O_{jg} \ln\left(\frac{O_{jg}}{E_{jg}}\right) + \left(1 - O_{jg}\right) \ln\left(\frac{1 - O_{jg}}{1 - E_{jg}}\right)\right], \quad (2.25)$$

where O_{jg} is the observed proportions of correct responses to item *j* for the examinees in group g, E_{jg} is the expected proportions of correct responses to item *j* for the examinees in group g, N_g

is the number of examinees in group g. For dichotomously scored items with P items, the proportion of examinees scoring zero is zero and the proportion of examinees scoring with correct responses on all P items is 1. Hence, the summation in the $S - X^2$ and $S - G^2$ formulations exclude $g = 0$ and $g = P$.

Another key distinction of Orlando and Thissen's (2000) $S - X^2$ and $S - G^2$ statistics is how the expected proportions of correct responses is computed. In the formulation provided in Equation 2.25, E_{jg} is computed as

$$E_{jg} = \frac{\int P_j(\theta) f^{*j}(g-1|\theta)\phi(\theta)\partial\theta}{\int f(g|\theta)\phi(\theta)\partial\theta}, \tag{2.26}$$

where $\phi(\theta)$ is the population distribution of the examinee ability parameter θ, $f(g|\theta)$ is the posterior distribution of the number of correct responses for group g, $f(g-1|\theta)$ is the posterior distribution of the number of correct responses for group g excluding item j. Approximating the integral in Equation 2.26 can be achieved using rectangular quadrature over equally spaced increments of θ from -4.5 to 4.5 (Orlando & Thissen, 2000). Kang and Chen (2008) described a generalized $S - X^2$ index for polytomous items.

2.7 Model-Data Fit for Performance Assessment Models

In performance assessments, the discrepancies in ratings awarded to a particular examinee by multiple raters could be due to the different levels of severity/leniency exhibited by the raters or other types of rater effects such as centrality/extremity, halo, or inaccuracy. Eckes (2011) summarized three steps to analyzing data from performance assessments: (1) forming hypotheses on the facets that are likely to be relevant, (2) specifying a measurement model that is ideal to incorporate each facet, (3) applying the model to account for each facet's impact in the best possible fashion. The adequacy of the measurement model used is an important ingredient in model building. If the wrong model is specified in Step 2 of Eckes' summary, then the validity of

the estimates is questionable. The MFRM uses a chi-square statistic for the evaluation of the overall (global) absolute model-data fit. This index is based on a log-likelihood chi-square and may be more useful in assessing the practical utility of the model (Eckes, 2011; Linacre, 2020).

The model specification of the MFRM also allows for the examination of fit for all facets in the model such as item, examinee, and rater using the INFIT and OUTFIT statistics (Eckes, 2011; Linacre, 2020). Specifically, the INFIT and OUTFIT measures are used to characterize the deviations in ratings that were observed and ratings that were expected. In the context of raters, large deviations between the observed and expected ratings for an individual rater might be an indication of the existence of rater effects (Myford & Wolfe, 2003). For example, a rater OUTFIT statistic can be expressed as

$$OUTFIT_r = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J} z_{ijr}^2}{N * J}, \tag{2.27}$$

where N is the number of examinees, J is the number of items, and z_{ijr} is the standardized residual of rater *r*'s rating of examinee *i* on item *j*. The standardized residual is computed using

$$z_{ijr} = \frac{X_{ijr} - E_{ijr}}{\sqrt{V_{ijr}}}, \tag{2.28}$$

where E_{ijr} is the expected rating of examinee *i* rated by rater *r* on item *j*, X_{ijr} is the observed rating of examinee *i* rated by rater *r* on item *j*, and V_{ijr} is the variance of the observed rating X_{ijr} around its expectation. The OUTFIT statistic is an unweighted mean-squared fit statistic, hence, it is sensitive to unusually deviant ratings from a rater who is considered a consistent rater.

The INFIT statistic is a weighted mean-square statistic. The INFIT is less sensitive to unexpected ratings. The rater INFIT statistic is given as

$$INFIT_r = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J} z_{ijr}^2 V_{ijr}}{\sum_{i=1}^{N}\sum_{j=1}^{J} V_{ijr}}. \tag{2.29}$$

Since raters' INFIT and OUTFIT statistics provide information on rater consistency in the use of the scales, large values may indicate inconsistency or other rater biases in the use of the rating scales. As earlier noted, the INFIT and OUTFIT measures can be extended to evaluate the fit of items and examinees.

Wolfe and McVay (2010) suggested the use of score-estimate correlation for evaluating rater effects. The score-estimate correlation, also referred to as point-measure correlation, is useful in detecting rater inaccuracy because this statistic depicts "the consistency between the rank ordering of the examinees by a particular rater and the rank ordering of those examinees by composite scores assigned by all other raters" (Wolfe & McVay, 2010, p. 7). The score-estimate correlation is analogous to the Pearson Product Moment Correlation since the scores and ability estimates are assumed to be continuous. A rater's score-estimate correlation (r_r) is computed using

$$r_r = \frac{\sum_{i=1}^{N}(\bar{X}_{i.r} - \bar{X}_{..r})(\hat{\theta}_i - \hat{\bar{\theta}})}{\sqrt{\sum_{i=1}^{N}(\bar{X}_{i.r} - \bar{X}_{..r})^2}\sqrt{\sum_{i=1}^{N}\left(\hat{\theta}_i - \hat{\bar{\theta}}\right)^2}}, \tag{2.30}$$

where $\bar{X}_{i.r}$ is observed average rating of rater *r* to examinee *i* across all items, $\bar{X}_{..r}$ is the observed average rating of rater *r* across all items and examinees, $\hat{\theta}_i$ is ability estimate of examinee *i,* and $\hat{\bar{\theta}}$ is the average ability estimate of all examinees. Boone and Staver (2020) indicated that the score-estimate correlation may be useful in detecting misfitting or miscoded items.

Another version of what looks like the score-estimate correlation is the Single Rater-Rest of the Raters (SR/ROR) correlation (Linacre, 2003). The SR-ROR reflects the extent to which a rater's ratings are consistent with the ratings of the rest of the raters. Myford and Wolfe (2003) provided some guidelines in interpreting SR/ROR correlations. They suggested that:

SR/ROR correlations less than .30 are considered to be somewhat low, while correlations

> greater than .70 are considered to be high for a rating scale composed of several categories. However, as the number of rating scale categories decreases, these rule-of-thumb values should be relaxed. For example, it is not uncommon to see SR/ROR correlations no higher than 0.20 in dichotomous ratings. If a SR/ROR correlation is near zero or negative for a given rater, then that rater rank orders ratees in a manner different from the other raters' rank ordering. (p. 410).

The SR/ROR correlation is computed using

$$r_{r,rest} = \frac{2NJ\left[\sum_{j=1}^{J}\sum_{i=1}^{N} R_{j,ijr}X_{ijr} + R_{i,ijr}X_{ijr}\right] - 2\sum_{j=1}^{J}\sum_{i=1}^{N} X_{ijr}\left[\sum_{j=1}^{J}\sum_{i=1}^{N} R_{j,ijr} + R_{i,ijr}\right]}{\sqrt{\left(2NJ\left[\sum_{j=1}^{J}\sum_{i=1}^{N} R_{j,ijr}^2 + R_{i,ijr}^2\right] - \left[\sum_{j=1}^{J}\sum_{i=1}^{N} R_{j,ijr} + R_{i,ijr}\right]^2\right)\left(4\left[NJ\sum_{j=1}^{J}\sum_{i=1}^{N} X_{ijr}^2\right] - \left[\sum_{j=1}^{J}\sum_{i=1}^{N} X_{ijr}\right]^2\right)}}, \quad (2.31)$$

where X_{ijr} is the observed rating of examinee *i*, rated by rater *r*, on item *j* and *NJ* is the total number of ratings awarded by rater *r*. The average scores of the items and examinees, excluding the rater *r* rating are given as $R_{j,ijr}$ and $R_{i,ijr}$, respectively. Specifically, $R_{j,ijr}$ and $R_{i,ijr}$ are computed, respectively, using Equations 2.32 and 2.33.

$$R_{j,ijr} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N} X_{ijr} - X_{ijr}}{N \cdot R - 1}, \quad (2.32)$$

$$R_{i,ijr} = \frac{\sum_{r=1}^{R}\sum_{j=1}^{J} X_{ijr} - X_{ijr}}{J \cdot R - 1}. \quad (2.33)$$

All the rater fit statistics detailed so far have been widely applied using MFRM (see Eckes, 2011; Wolfe & McVay, 2010). Gaps still exist in literature about the best approach of evaluating model-data fit for the HRM. There are no documented absolute fit measures for the HRM. Patz et al. (2002) used Schwarz's (1978) Bayesian Information Criteria (BIC) to evaluate relative model-data fit between the HRM and the MRFM. A recent *R* package by Robitzsch and Steinfeld (2018), which uses the *immer_hrm* function to fit the HRM, reports seven fit indices including BIC, Akaike Information Criteria (AIC; Akaike, 1973, 1974), and the AIC corrected

for bias (AICc; Sugiura, 1978). All these fit indices are typically employed when assessing relative model-data fit between competing models with smaller values indicating better fit. For example, Nieto and Casabianca (2019) used AIC, BIC, and deviance information criteria (DIC; Spiegelhalter et al., 2002) to compare the M-HRM to the MGPCM. Given the preponderance of Bayesian studies, there is a need for Bayes absolute fit. One of those methods is PPMC.

2.8 Posterior Predictive Model Checking

2.8.1 Overview of PPMC

Model-checking from a Bayesian perspective can be approached using Bayesian residuals and predictive diagnostic checks (Fox, 2010). This study was designed to evaluate the effectiveness of several discrepancy measures in assessing model-data fit of the HRM using PPMC. The PPMC is used for assessing whether there are aspects of the data not captured by the model. The PPMC is built on the posterior predictive methodology (Rubin, 1984) on the principle that data replicated or simulated from the posterior distribution should bear a resemblance to the observed data. The predictive posterior methodology accounts for the parametric uncertainty in the posterior distribution and the sampling uncertainty in the data. The posterior predictive distribution for replicated data, as defined by Rubin (1984), is a distribution of future observable quantity conditioned on the observed data. The posterior predictive distribution of X_{rep} is given by

$$p\left(X_{rep}\middle|X_{obs}\right) = \int p\left(X_{rep}\middle|\theta\right)p(\theta|X_{obs})d\theta. \quad (2.34)$$

In Equation 2.34, $p(\theta|X_{obs})$ specifies the posterior distribution of the unknown parameter, θ. This quantity gives the representation of the uncertainty in the unknown parameter. The symbol X_{rep} denotes the replicated or simulated data, which is drawn from the posterior predictive

distribution, and X_{obs} represents the observed data. A model is considered a good fitting model if the replicated data closely resemble the observed data. As described by Equation 2.34, the posterior predictive distribution is intuitive and computationally straightforward using Monte Carlo methods. Obtaining the replicated samples is achieved by randomly drawing values of the parameter from the joint posterior distribution. Drawn parameter values are then used to simulate the data based on the fitted model. This process is repeated many times as shown in Figure 2. As described in this figure, $\{X_{rep}^{[1]}, X_{rep}^{[2]}, \ldots, X_{rep}^{[S]}\}$ are independent and identically distributed samples from $p(X_{rep}|X_{obs})$ which represents the posterior predictive distribution. Typically, 1,000 or more datasets are replicated.

sample $\theta^{[1]} \sim p(\theta\|X_1, \ldots, X_n)$,	sample $X_{rep}^{[1]} \sim p(X_{rep}\|\theta^{[1]})$
sample $\theta^{[2]} \sim p(\theta\|X_1, \ldots, X_n)$,	sample $X_{rep}^{[2]} \sim p(X_{rep}\|\theta^{[2]})$
sample $\theta^{[3]} \sim p(\theta\|X_1, \ldots, X_n)$,	sample $X_{rep}^{[3]} \sim p(X_{rep}\|\theta^{[3]})$
sample $\theta^{[4]} \sim p(\theta\|X_1, \ldots, X_n)$,	sample $X_{rep}^{[4]} \sim p(X_{rep}\|\theta^{[4]})$
$\vdots$	$\vdots$
sample $\theta^{[S]} \sim p(\theta\|X_1, \ldots, X_n)$,	sample $X_{rep}^{[S]} \sim p(X_{rep}\|\theta^{[S]})$

Figure 2. Posterior Predictive Distribution Sampling Algorithm

Under the PPMC approach, model misfit could be assessed graphically or numerically. Plots such as scatterplots, bar charts, boxplots, and histograms can be employed to graphically compare the observed data and replicated data. Any systematic differences observed between aspects of the observed data set and those of the replicated data sets are indicative of the failure of the model to explain those aspects of the data (Sinharay et al., 2006). One of the drawbacks of graphically comparing observed and replicated is that some key features of the data may not be

easily noticeable.

Ames et al. (n.d.) illustrated the difficulty of only using graphical checks alone to evaluate model-data fit using PPMC. Here, I present a similar example using simulated data generated from a 1PL model. Data were simulated for 1000 examinees with 5 items (simulation parameters are presented in Appendix 1). Figure 3 presents the total score distribution of the simulated data. From Figure 3, it can be seen that 227 examinees obtained a total score of zero (i.e., incorrect responses on all five items), whereas 17 examinees obtained a total score of five (i.e., correct responses on all five items). The simulated data were further calibrated with 1PL and 2PL IRT models using Bayesian estimation methods with MCMC.

Using PPMC approaches, 1000 datasets were replicated using the joint posterior distributions of the 1PL and 2PL models. Figure 4 displays the total score distributions for two replicated datasets: one for the 1PL model and one for the 2PL model. In Figure 4, the left panel shows the total score distribution under the 1PL model, and the right panel shows the total score distribution under the 2PL model. The number of examinees scoring a total score of one, two, and three under the PPMC for the 1PL closely resemble the observed data. As shown in the right panel of Figure 4, the data replicated using 2PL did not closely resemble the observed data. However, it can be seen that 17 examinees obtained a total score of 5 in the observed total score distribution and 16 examinees obtained a total score of 5 in the replicated dataset for the 2PL model. This makes it subjective to only judge how well the replicated data resembles the observed data using graphical checks alone. To this end, more quantifiable techniques using discrepancy measures and PPP-values allow for a more direct evaluation of the discrepancy between the observed data and posited model (Meng, 1994; Gelman et al., 1996).

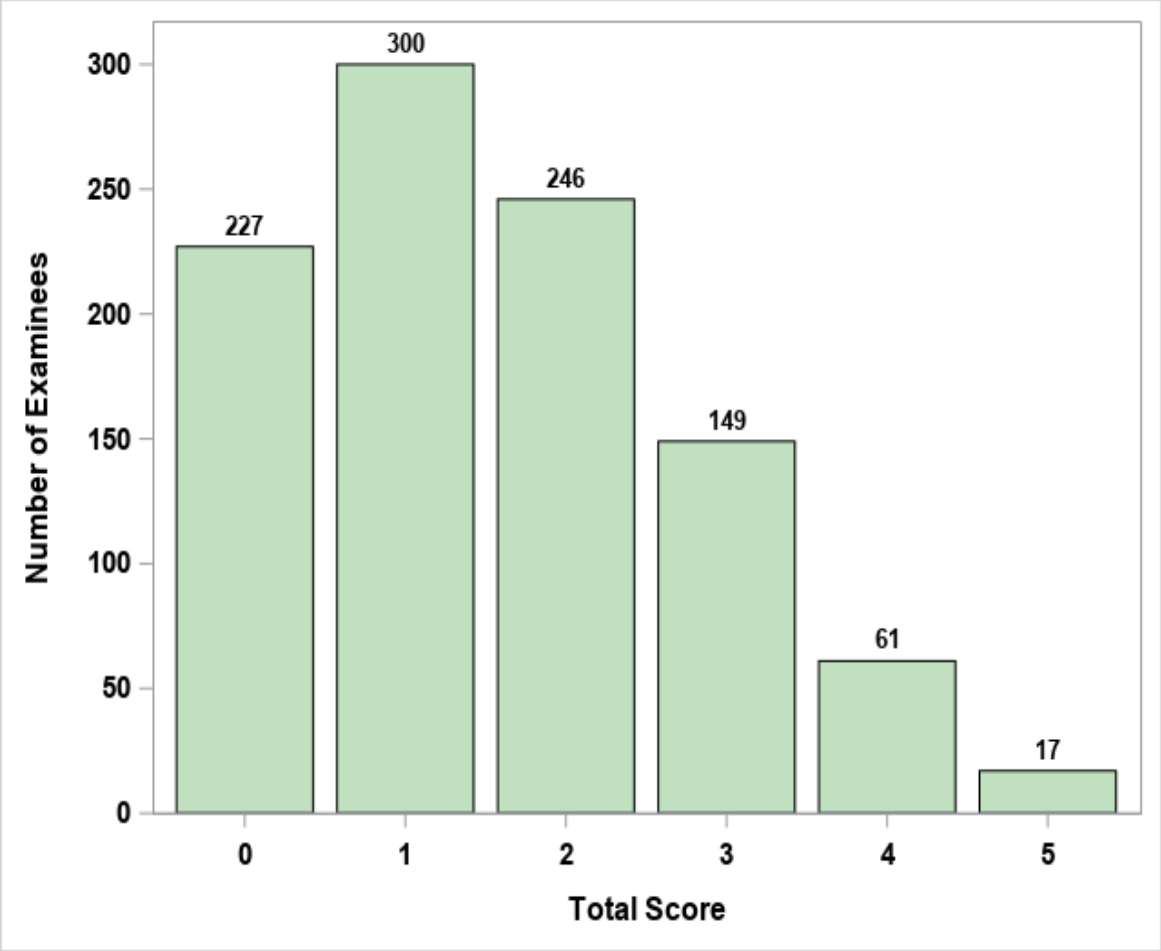

Figure 3: Distribution of the number of examinees obtaining each total score

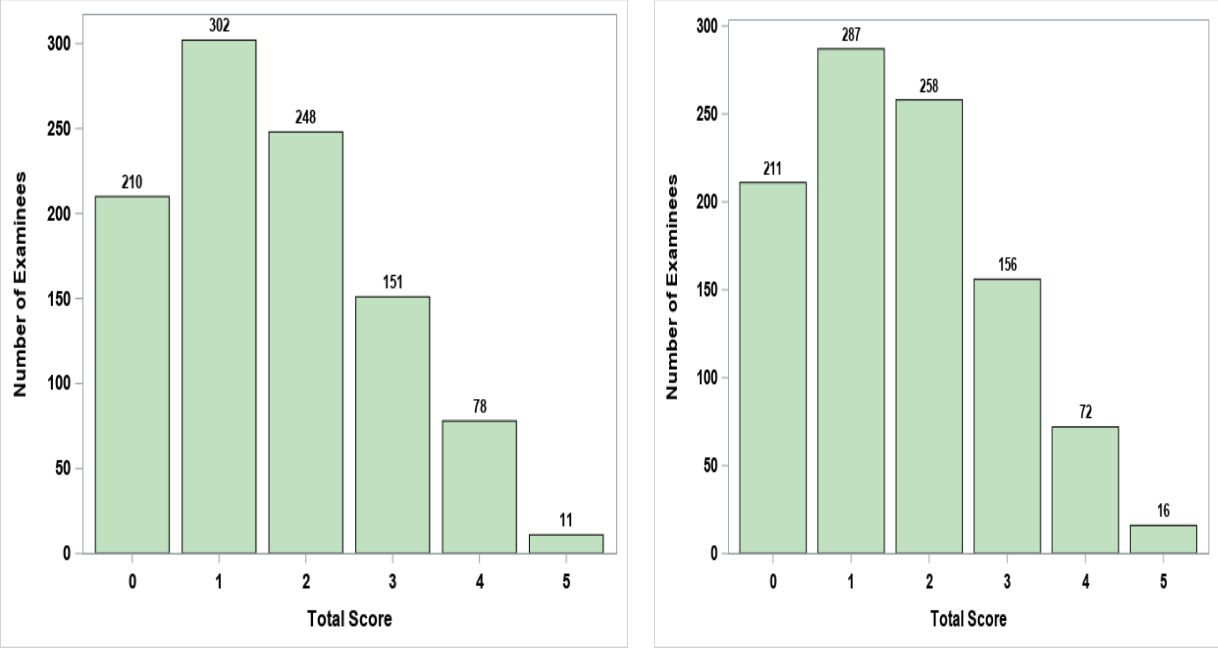

Figure 4: Distribution of total score of replicated data for the 1PL (left panel) and 2PL (right panel) models.

2.8.2 Discrepancy Measures

Discrepancy measures are statistics computed from the observed and replicated data. It could be simple summary statistics such as measures of center (e.g., mean, median, mode) and measures of spread (e.g., standard deviation, interquartile range, range). Other statistical measures such as correlation coefficients, item-total correlation, rater reliability, odd ratios, total test score and proportion of correct answers could also serve as discrepancy measures. There is no limit to the choice of discrepancy measure that could be employed but detecting a misfitting model hinge on the choice of the discrepancy measure used. Bayarri and Berger (2000) suggested that the choice of discrepancy measure allows the researcher to evaluate how compatible the fitted model is to the observed data. That is to say that for every given model, the choice of the discrepancy measure should be chosen to provide evidence of high power and low Type-I error rates.

The discrepancy measure of the replicated and observed data are typically denoted by $T(D_{rep})$ and $T(D_{obs})$, respectively. For the replicated datasets, S number of $T(D_{rep})$ is computed [i.e., $T(D_{rep})_1, T(D_{rep})_2, \ldots, T(D_{rep})_S$], where S is the number of replicated datasets. Typically, the 50th percentile of the discrepancy measure of the replicated datasets is compared to the discrepancy measure of the observed statistic. Discrepancy measures such as total score distribution, item-score correlation, Yen's Q_1, Q_3, G^2, $S-X^2$, $S-G^2$, and Agresti's (2002) global odds ratio have been employed in IRT to evaluate absolute model-data fit with PPMC techniques.

Total Score Distribution

The total score is the sum of the item scores (for each examinee). The total score distribution is the distribution of examinees' sums of item scores. The illustration presented in Section 2.8.1 is an example of total score distribution. This distribution may be useful in detecting misfit at the test level. As depicted in Section 2.8.1, the total score distribution using graphical checks alone may be inconclusive. However, pertinent features of the total score distribution can be described using summary statistics such as mean, standard deviation, and skewness. Summary statistics computed from the total score distribution can serve as discrepancy measures to examine different aspects of the distribution. For example, we see that the center of the total score distribution for the observed data (Figure 3) resembles the replicated 1PL data (Figure 4, left panel). The tails of the distribution do not appear to be adequately captured by the 1PL model. The center can be described with the mean of the total score distribution. In this case, the mean will be the discrepancy measure. The variability and shape of the data can be evaluated using standard deviation and skewness statistics, respectively. In addition, we can use the first and third quartiles to evaluate how well the values around the tails

of the distribution are captured.

Item-Total Correlation Coefficient

The item-total correlation is the correlation between a particular item's score and the total score, also referred to as the point-biserial correlation for dichotomous items and polyserial correlation for polytomous items. This coefficient measures the extent to which responses on a particular item are related to responses to other items on the test (Allen & Yen, 1979). The item-total correlation is regarded as a discrimination index. Essentially, item-total correlation indicates how well items discriminate between high-performing and low-performing examinees. Item-total correlation can be computed by including an item's scores in the total score or removing the item's score in the total score. To avoid spuriousness, it is recommended to remove a particular item's score from the total score (Crocker & Algina, 1986). Item-total correlation coefficient has been employed as a discrepancy measure to detect item misfit using PPMC in IRT studies (Li et al., 2017; Sinharay & Johnson, 2003).

Odds Ratio Statistic

The odds ratio is a pairwise measure statistic that measures the association between two items typically presented in two-way contingency tables. For example, the two-way contingency table for two dichotomous items (*j* and *j**) can be given as

Table 6. Two-way Contingency Table for Two Dichotomous Items

		Item *j*	
		0	1
Item *j**	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

where n_{00} is the observed number of examinees scoring 0 and 0 on items *j* and *j**, respectively. Similarly, n_{10}, n_{01}, and n_{11} are the observed number of examinees having pairwise scores of (1,

0), (0, 1), and (1, 1), respectively, on items *j* and *j**. These values provide the basis for computing the odds ratio. The odds ratio is computed as

$$OR = \frac{n_{11}n_{00}}{n_{10}n_{01}}. \tag{2.35}$$

Chen and Thissen (1997) used the odds ratio as a measure to evaluate local dependence for dichotomous IRT models. The scenario presented in Table 6 is suitable for items with binary responses (i.e., 0 and 1 responses). Previous studies (Li et al., 2017; Zhu & Stone, 2011, 2012) have applied the odds ratio statistic to polytomous items. These studies dichotomized polytomous items by categorizing them into two groups. For example, response score categories of 1, 2, and 3 could be recoded as 0 and score categories of 4 and 5 could be recoded as 1.

2.8.3 Posterior Predictive p-value

In PPMC, there are *S* replicated datasets. A thousand datasets are typically replicated in PPMC. Therefore, for a chosen discrepancy measure, *S* number of this discrepancy measure will be computed from the datasets [i.e., $T\left(D_{rep}\right)_1, T\left(D_{rep}\right)_2, \ldots, T\left(D_{rep}\right)_S$]. For example, if the item-total correlation is chosen as the discrepancy measure, then the item-total correlation will be computed for each replicated dataset. The distribution of the discrepancy measure is referred to as the reference distribution. This distribution provides the means to assess the extremeness of the observed discrepancy (Fox, 2010).

Frequentists construct the *p*-value under the assumption of a true null hypothesis. The *p*-value over the sampling distribution of the test statistic is mathematically given as

$$p_F = P\{T(X) \geq T(x)|H_o\}, \tag{2.36}$$

where p_F is the probability of observing the test statistic $T(X)$ or values more extreme as $T(x)$ given that the null hypothesis H_o is true. A small value of p_F is an indicative of an unlikely statistic under the sampling distribution. The Bayesian version of the *p*-value is the PPP (Meng,

1994). The frequentist *p*-value approach shares some features with the Bayesian *p*-value. The PPP-value is computed over the posterior predictive distribution and is mathematically given as

$$p_B = \Pr\{T(D_{rep}) \geq T(D_{obs}) | \theta\}, \quad (2.37)$$

where p_B is the probability that the replicated data $T(D_{rep})$ could be more extreme than the observed data $T(D_{obs})$ as measured by the discrepancy measure (Gelman et al., 2004). It is desired that $T(D_{obs})$ is located near the center of the distribution, therefore, a PPP-value close to 0.5 is indicative of adequate model-data fit and values less than 0.05 or greater than 0.95 is indicative of poor model-data fit (Sinharay, 2006).

2.8.4 Applications of PPMC in IRT

Bayesians regard the PPMC as a powerful diagnostic technique in assessing whether different aspects of the data are captured by the model. The frequentist methods of assessing IRT model-data fit are easily extended to a Bayesian framework using PPMC techniques. PPMC has been employed to evaluate model-data fit of dichotomous and polytomous IRT models using popular frequentist IRT fit indices.

Sinharay and Johnson (2003) evaluated the utility of PPMC in detecting misfit of dichotomous IRT models under different data generating and data analysis models. They generated data for 2,500 examinees responding to 30 items with the 1PL, 2PL, and 3PL models. Data generated with a more complex model were analyzed using the same model and models that are less complex. For example, data generated with the 2PL model were analyzed with the 1PL and 2PL models; but not the 3PL model. Data were also generated to have a *testlet* effect, two dimensions, and speededness, and were analyzed with the 2PL and 3PL models. The authors considered several discrepancy measures such as percentage-correct score for the items, observed score distributions, biserial correlation coefficient (including the mean, variance,

minimum, and maximum of the biserial correlation), proportion of examinees answering pairs of items correctly and odds ratio. They found that observed score distribution and odds were useful in detecting misfit of the 1PL model. Also, they found biserial correlation to be powerful in detecting lack of fit of the 1PL model when the generating models were 2PL and 3PL models. In detecting lack of unidimensionality, they found the odds ratio to be a powerful discrepancy measure. Sinharay et al. (2006) also found biserial correlation to be useful in detecting misfit of the 1PL model.

Sinharay (2006) further employed the PPMC to examine model-data fit of unidimensional dichotomous IRT models using item fit plots, and Orlando and Thissen's $S-X^2$ and $S-G^2$ statistics as the discrepancy measures. The $S-X^2$ and $S-G^2$ measures were useful in detecting misfit as their type I error rates did not exceed the nominal level. The $S-X^2$ also performed well for small test with 10 items and 500 examinees. Ames (2015, 2018) investigated the influence of the choice of prior distributions on the performance discrepancy measures using PPMC. Data for her study were generated using dichotomous IRT models. Percent correct, $S-X^2$, OUTFIT, and INFIT measures served as the discrepancy measures. Because the focus of Ames (2015, 2018) was the influence of the choice of prior distributions, four prior specifications were employed: noninformative, informative-accurate, informative-inaccurate, noninformative-inaccurate. The author found that the PPMC technique was influenced by prior specifications. This finding varied across the choice of discrepancy measure. For example, the effect of prior choice was small for the $S-X^2$ statistic but nonnegligible for the INFIT statistic.

Only a few simulation studies have applied PPMC to rigorously evaluate model-data fit of polytomous IRT models. Zhu and Stone (2011, 2012) extended PPMC to the GRM using discrepancy measures such as item score distribution, item-total correlation, Yen's Q_1 and Q_3,

and Agresti's (2002) global odds ratio. Consistent with previous studies, Zhu and Stone (2011, 2012) found PPMC to be effective in detecting model-data fit. Specifically, pairwise measures such as the global odds ratio and Yen's Q_3 were found to be powerful in detecting violations of unidimensionality. More recently, Li et al. (2017) investigated the performance of PPMC in detecting misfits of polytomous IRT models such as GPCM, PCM, rating scale model (RSM; Andrich, 1978), and the modified rating scale model (MRSM). Their study explored five discrepancy measures: (a) test score distribution; (b) item-total correlation; (c) $Q_1 - \chi^2$; (d) global odds ratio; and (e) Yen's Q_3. Li and colleagues generated their data using the GPCM. The generated data were analyzed using GPCM, PCM, RSM, and MRSM. Li et al. (2017) results showed that the discrepancy measures performed differently given the analysis model. They found that discrepancy measures such as item-total correlation, Yen's Q_1, Yen's Q_3, and the global odds ratio were dependent on sample size. Their results showed that larger sample sizes were associated with a larger percentage of flagged items or item pairs. However, test lengths were not shown to impact the effectiveness of the PPMC using these discrepancy measures.

As earlier noted, only Casabianca et al. (2017) and Nieto and Casabianca (2019) have applied the PPMC to the HRM. Casabianca et al. (2017) employed PPMC to the L-HRM using the total score distribution as the discrepancy measure. They defined the total score distribution as "the sum of the item score." The item score is an examinee's average rating based on multiple ratings on a particular item. Although Casabianca et al. (2017) did not provide the full results of their PPMC analyses, they conclude that the results allowed them to infer that the L-HRM with autoregressive time series was an adequate fit to their data. Nieto and Casabianca (2019) employed total score distribution and rater variability (standard deviation) to evaluate absolute model-data fit of the M-HRM. The results of the total score distribution suggest that the M-HRM

adequately captured features of the observed data. However, there were noticeable deviations between the observed rater variability and the replicated rater variability. These studies did not report the resulting PPP-values. In addition, Casabianca et al. (2017) and Nieto and Casabianca (2019) studies did not evaluate the utility of the PPMC when the IRT component of the HRM is misspecified. Hence, it is unclear as to whether the discrepancy measures they used can detect misfit when a correct or incorrect IRT model is specified.

2.9 Research Questions

The main purpose of the present study is to evaluate the absolute model-data fit of the HRM. This study asks the following research questions:

1. What is the Type I error rate and power of the test-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?
2. What is the Type I error rate and power of the item-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?
3. What is the Type I error rate and power of the rater-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?

2.10 Chapter Summary

It was important to understand the framework used in this study and previous work that has been done in this area. Chapter 2 outlined this framework and reviewed previous literature. This chapter first described traditional dichotomous and polytomous IRT models including when to apply these models. These traditional IRT models assume that an examinee's response to an item is independent of the examinee's response to any other item given the examinee's ability – an assumption that researchers have suggested to be lacking in performance assessments. This chapter also described appropriate models for performance assessments including the HRM,

which is the measurement model of interest in this study. In addition, this chapter documented different rater effects that have been studied in the context of the HRM. In addition, the implications of complete and incomplete rating designs were discussed. The essential component of this study is to evaluate the absolute model-data fit of the HRM. The chapter ended with an extensive discussion of the PPMC techniques including the application of PPMC in IRT and different discrepancy measures that have been previously applied in evaluating PPMC in IRT models.

CHAPTER 3:

METHODS

The focus of this dissertation is on the performance of PPMC in detecting misfits of the HRM under varying simulation conditions. The HRM models the hierarchy that exists in rating data by accounting for rater severity and variability. In the first stage, a signal-detection-like model describes the relationship between the ideal ratings and observed ratings, while an IRT model describes the relationship between ideal ratings and examinee latent ability in the second stage. The current study will investigate the empirical power and Type I error rates of different discrepancy measures at the test-level, item-level, and rater-level of the HRM. The implications of rater effects and rating designs on model-data fit will also be investigated. The study will also document the most effective discrepancy measures.

This chapter outlines the simulation design, discrepancy measures for the PPMC approach, data analyses, and evaluation methods. In addition, the methodology for a real data example is described in this chapter. The choice of the conditions selected for this dissertation is guided by previous studies involving performance assessments. The current study varied the IRT component model of the HRM, rating designs, number of examinees, and rater effects. Factors such as the number of item categories, test length, and distribution of examinee latent trait were fixed across all conditions.

3.1 Simulation Design Factors

3.1.1 Data Generation and Analysis Models

The second level of the HRM models the relationship between examinees' ideal ratings and the latent ability traits. The second level does not depend on the rater characteristics such as severity/leniency and consistency in the use of the scale. A polytomous IRT model is the

statistical model for this level of the hierarchy $\xi_{ij}|\theta_i$. Several simulation studies have applied the PCM or GPCM at the second level of the HRM. Nieto and Casabianca (2019) evaluated parameter recovery of the M-HRM. In generating data for the M-HRM, Nieto and Casabianca (2019) specified the GPCM as the polytomous IRT model. Similarly, Casabianca et al. (2017) specified the GPCM as the polytomous IRT model in their simulation studies which examined parameter recovery of the L-HRM. Patz et al. (2002) illustrated the utility of the HRM. The paper compared the HRM to the MFRM. Patz and colleagues employed the PCM as the IRT component of the HRM. None of these studies highlighted here varied the polytomous IRT models used. Thus, it is unclear of the effects of using the GPCM versus the PCM on model-data fit or parameter estimates. In the present study, the GPCM was the generating model and both PCM and GPCM serve as the analysis model. This study used HRM-GPCM to refer to HRM data generated or analyzed with the GPCM as the IRT component, whereas, HRM-PCM refer to HRM data generated using GPCM and analyzed with PCM as the IRT component.

3.1.2 Test Length

Unlike multiple-choice tests, performance assessments require examinees to perform fewer tasks because the tasks necessitate more time and resources for rating. For example, the Analytical Writing portion of the GRE, a high-stake assessment, consists of two analytical writing tasks. Similarly, the four components of the TOEFL comprise of two sections (writing and speaking) that require human raters to judge the quality of examinees' work. The Speaking portion consists of four tasks and the Writing portion consists of two writing tasks (one integrated and one independent). Simulation studies involving performance assessments have mostly utilized a small number of test items. Patz et al. (2002) and Casabianca et al. (2017) fixed the test lengths to five items. Both studies used a five-category scale per item for both observed

and ideal ratings. Casabianca and Wolfe (2017) considered test lengths of two, four, and eight items, while Nieto and Casabianca (2019) investigated parameter recovery of the M-HRM with test lengths of three and six items. Their results were consistent with Kim (2009) and DeCarlo (2008, 2010) that showed that bias in parameter recovery decreased with larger test lengths. For example, Casabianca and Wolfe (2017) showed that the correlation coefficients between observed and estimated examinee ability parameters were in the range of 0.796 and 0.802 for a 2-item test, between 0.875 and 0.881 for a 4-item test, and between 0.927 and 0.930 for an 8-item test, suggesting that a larger number of items improve the estimates of the person parameters. This study uses a test length of four items with five response categories, which is typical of the number of items in performance assessments.

3.1.3 Number of Examinees

Accurate estimation of IRT model parameters may largely depend on the number of examinees. Previous simulation studies have shown that larger sample sizes yield estimates that are close to the true parameter (e.g., Kieftenbeld & Natesan, 2012; Reise & Yu, 1990). Using the GRM with a test length of 25 items, Reise and Yu (1990) suggested that a minimum sample size of 500 examinees was needed to yield higher accuracy in parameter estimates. Reise and Yu's (1990) study employed MML as their estimation method. Comparing MML and Bayesian methods, Kieftenbeld and Natesan (2012) revealed that accuracy in estimates increased with an increase in sample size and test length. There is little guidance on the sample size required to accurately estimate item, rater, and ability parameters in performance assessments. Linacre (2020) suggests that a minimum sample size of 30 observations per item and at least 10 observations per rating-scale category are required to yield estimates that will have some degree of stability. In performance assessments, it is typical for each examinee to receive multiple

ratings. Hence, a design in which examinees receive multiple ratings on multiple items will result in large observations, which could augment a small sample size. Previous studies that involve the HRM have mostly used sample sizes between 250 and 1000 examinees (e.g., Casabianca et al., 2017; Nieto & Casabianca, 2019; Patz et al., 2002). Consistent with previous studies on performance assessments, sample sizes of 250 and 500 examinees were considered for the present study.

3.1.4 Number of Raters

The number of raters employed in performance assessments is relative to design considerations such as the rating design, number of items, number of examinees, and whether each examinee receives single or multiple ratings. Previous simulation studies have utilized varying number of raters. For example, Patz et al. (2002), Nieto and Casabianca (2019), and Casabianca and Wolfe (2017) all fixed the number of raters to 3, 25, and 100, respectively. On the other hand, Uto and Ueno (2020) had different number of raters (5, 10, and 30). This large variation in the number of raters employed in these studies is an indication that the choice of the number of raters in performance assessments is relative to the design considerations highlighted earlier. For example, the TOEFL Speaking section requires the contribution of a minimum of three raters to rate an individual examinee (ETS, 2016). Given the large number of examinees that take the TOEFL, an international assessment, it is easy to imagine that rating of assessments will require a large number of raters to rate all the test takers who take the TOEFL. Simulating data with 100 raters, Casabianca and Wolfe (2017) designed their study such that each examinee was rated by a subset of the raters. In their study, examinees received two, four, or eight ratings per item. Similarly, Nieto and Casabianca (2019) randomly assigned two ratings to each examinee, which resulted in an incomplete design. This type of design is common in large-scale

assessments (McClellan, 2010). To keep the design conditions manageable, this study fixed the number of raters to eight.

3.1.5 Rating Design

There is little research into performance of HRM with rating designs that lack connectedness. None of the simulation studies that directly applied the HRM explored the issue of connectedness. However, as documented in Chapter 2, Patz et al. (2002) explored different rating designs using empirical data. Patz and colleagues referred to the rating designs as Modality One, Modality Two, and Modality Three. Modality One, in which raters scored the entire booklet of 11 items, resulted in the smallest absolute modality bias. Fully-crossed designs have been shown to perform better in terms of bias (Hombo et al., 2001) and in terms of classification of examinees (Wind et al., 2019). While fully-crossed design may be desirable, the costs associated with employing fully-crossed design are higher compared to an incomplete rating design. Incomplete rating designs have been shown to yield estimates comparable to the true parameters. For example, Casabianca and Wolfe (2017) employed an incomplete rating design in which they varied the number of ratings each examinee received. They found that the correlation coefficients between examinees' true and estimated parameters were considerably high. For example, for their 4-item condition with no biased raters, the correlation coefficient between true and estimated ability parameters was 0.877 when each examinee only received two ratings compared to 0.878 when each examinee received four or eight ratings.

To assess the performance of the PPMC in detecting misfits across rating designs, the fully-crossed and spiral rating designs were considered for this study. The two rating designs considered here fall under the connected rating design. The fully-crossed rating design provides a baseline for assessing how well the spiral rating design performs. In the fully-crossed design,

every rater rated every examinee on all items. This design is shown in Table 7. As seen in this table, each examinee received 8 ratings on each item. However, in the spiral rating design, each rater rated every examinee on a subset of the items. Table 8 shows the spiral rating design employed in this study. In this design, each examinee received 2 ratings on each item. As shown in Table 8, Raters 1 and 2 rated every examinee on Item 1 while Raters 3 and 4 rated every examinee on Item 2. In addition, Raters 5 and 6 scored every examinee on Item 3, and finally, Raters 7 and 8 scored every examinee on Item 4. Therefore, the spiral rating design employed in this study provided only 25% of the data of the fully-crossed rating design. Li et al. (2017) found evidence of higher effectiveness of PPMC in detecting misfit with larger samples. Although previous studies (Casabianca & Wolfe, 2017; Hombo et al., 2001) suggested that incomplete rating designs resulted in adequate parameter recovery, the performance of PPMC in detecting misfit of the HRM-PCM is expected to be higher under the fully-crossed rating design due to considerably larger number of ratings.

Table 7. Fully-Crossed Rating Design for Simulated Ratings

Rater	Item	Examinee				
		1	2	3	…	*n*
1	1 – 4	✓	✓	✓	…	✓
2	1 – 4	✓	✓	✓	…	✓
3	1 – 4	✓	✓	✓	…	✓
4	1 – 4	✓	✓	✓	…	✓
5	1 – 4	✓	✓	✓	…	✓
6	1 – 4	✓	✓	✓	…	✓
7	1 – 4	✓	✓	✓	…	✓
8	1 – 4	✓	✓	✓	…	✓

Table 8. Spiral Rating Design for Simulated Ratings

Rater	Item	Examinee				
		1	2	3	…	*n*
1 – 2	1	✓	✓	✓	…	✓
3 – 4	2	✓	✓	✓	…	✓
5 – 6	3	✓	✓	✓	…	✓
7 – 8	4	✓	✓	✓	…	✓

3.1.6 Proportion of Aberrant Raters

Raters have different characteristics that may influence the choice of rating categories that they apply. These characteristics are referred to as rater effects. To reduce rater effects, raters are typically trained, calibrated, and monitored during the rating process. Previous research has shown that rater effects still exist despite training and calibration of raters (e.g., Eckes, 2020; Ezike & Ames, 2021; Weigle, 1998). In the context of the HRM, severe raters are those with bias statistics that are less than -0.5 (i.e., $\phi < -0.5$) and raters who are deemed to be lenient are those whose bias statistics are greater than 0.5 (i.e., $\phi > 0.5$). In addition, smaller values of the variability statistic, ψ, is an indication of more consistent raters. Previous simulation studies have varied the proportion of aberrant raters. Casabianca and Wolfe (2017) considered three types of aberrant raters: normal, unreliable, and severe. Normal raters were classified as raters with small bias ($-0.5 \leq \phi \leq 0.5$) and small variability ($\psi \leq 0.75$). In the unreliable condition, 20% of raters were simulated to have small bias statistics but larger variability. Similarly, in the severe condition, 20% of the raters were simulated to exhibit large rater severity. These raters had rater bias statistics less than -0.5.

Two types of rater effects classification were employed in this study: no rater effects and rater effects. All raters in the *no rater effects* condition were simulated to have small bias and

small variability parameters. Similar to Casabianca and Wolfe (2017), 75% of the raters in the *rater effects* category have small bias and small variability parameters, while the remaining 25% have significantly large bias and large variability parameters. Hence, six out of the eight raters in the rater effects category have small bias and small variability parameters, while two out of the eight raters in the rater effects category have significant bias and are inconsistent raters. It is expected that conditions with rater effects with perform worse in comparison to the conditions without rater effects. The data generating parameters of the raters are presented in Section 3.2.2.

3.2 Data Generation Parameters

3.2.1 Item Parameters

The item parameters used in generating the IRT component of the HRM under the HRM-GPCM are provided in Table 9. The choice of the item parameters was guided by previous simulation studies. Consistent with Li et al. (2017), the item step parameters were drawn from $\mathcal{N}(\mu = -1.5, \sigma = 0.5), \mathcal{N}(\mu = -0.5, \sigma = 0.5), \mathcal{N}(\mu = 0.5, \sigma = 0.5)$ and $\mathcal{N}(\mu = 1.5, \sigma = 0.5)$. Only one set of step parameters shown in Table 9 were drawn from these distributions. These parameters were fixed across all replications.

There has been evidence to suggest that the performance of PPMC in detecting misfit of Rasch models is lower when the discrimination parameter is close to 1 (e.g., Li et al., 2017). The discrimination parameters for the present study were drawn from a lognormal distribution, $ln\mathcal{N}(0, 1)$. New discrimination parameters were drawn for every new replication to ensure generalizability of this study. Only the discrimination parameters were varied in this study since previous work has not suggested any impact of item difficulty or step parameters on the performance of PPMC in detecting misfits.

Table 9. Item Generating Parameters for the IRT Component of the HRM

Item	δ_{j1}	δ_{j2}	δ_{j3}	δ_{j4}
1	-2.788	-0.173	0.152	2.941
2	-1.498	-0.203	1.885	1.675
3	-1.993	-0.327	0.327	2.047
4	-1.468	-0.358	1.003	1.908

3.2.2 Rater Parameters

The rater parameters used in generating the data are provided in Table 10. Data were generated for raters without rater effects (i.e., raters with normal behaviors) and raters with rater effects (i.e., raters with aberrant behaviors). Using distributions suggested by Casabianca and Wolfe (2017), the bias and variability parameters, for raters without rater effects, were drawn from normal distribution $\mathcal{N}(\mu = 0, \sigma = 0.25)$ and a lognormal distribution $ln\mathcal{N}(\mu = -0.70, \sigma = 0.25)$, respectively. For raters with rater effects, the bias and variability parameters were drawn from $\mathcal{N}(\mu = -1.00, \sigma = 0.30)$, and $ln\mathcal{N}(\mu = 0, \sigma = 0.25)$, respectively.

All eight raters in the condition without rater effects have bias statistics within acceptable thresholds. The bias values for these raters are between -0.50 and 0.50. Also, the variability values are less than 0.75. Raters with bias parameters in the range of -0.50 and 0.50 and variability parameter less than 0.75 are more probable to award ratings that are close to the true ratings.

In the rater effects condition, Raters 6 and 8 were replaced with more biased and less consistent raters. These raters have large variability values ($\psi_6 = 1.487$ and $\psi_8 = 0.980$) and bias values outside the suggested acceptable thresholds ($\phi_6 = -0.931$ and $\phi_8 = -1.055$). Rater

6 is more likely to inconsistently award scores in some categories below the ideal scores and Rater 8 is more likely to award scores in some categories below the ideal scores in an inconsistent fashion. In sum, Rater 6 and Rater 8 are inconsistently severe raters. However, Rater 6 is more inconsistent compared to Rater 8. The rater generating parameters presented in Table 10 were fixed across all replications.

Table 10. Rater Generating Parameters

Rater	No rater effect		Rater effect**	
	ϕ_r	ψ_r	ϕ_r	ψ_r
1	-0.224	0.425		
2	0.046	0.270		
3	0.397	0.403		
4	-0.283	0.740		
5	-0.020	0.539		
6	0.033	0.604	-0.931	1.487
7	0.177	0.561		
8	-0.060	0.597	-1.055	0.980

**for Raters 6 and 8 have rater effects for the "rater effect" condition

3.2.3 Examinee Latent Ability Parameters

The ability distribution of examinees can take on any shape. In IRT, standard estimation approaches such as MML assume that examinee latent ability distribution is normally distributed. Estimates of the item, examinee, and rater parameters may be biased when θ is nonnormal. Bayesian estimation approach is one solution to this problem. Using MCMC, Conforti and Casabianca (2016) explored how well HRM parameters were recovered with nonnormal latent ability distribution. They observed nonignorable bias in examinee latent traits and item parameters, however, the rater parameters were robust to nonnormality in the examinee

latent traits. The implications of nonnormal examinee ability traits are not of interest in this study. In all conditions of this study, the latent traits were drawn from a normal distribution with a mean of zero and a standard deviation of 1, $\mathcal{N}(\mu = 0, \sigma = 1)$, in an effort to avoid any potentially biased estimates in the item and examinee parameters as accurate parameter estimations are crucial when employing PPMC.

3.3 Data Generation

Table 11 presents the summary of the fixed and manipulated simulation conditions. The number of response scale options, distribution of examinee latent trait, number of raters, generating model for the IRT component, and the test length were all fixed. However, the rating design, number of examinees, rater behavior, and analysis model for the IRT component were manipulated. The data for this study were generated in R (version 4.0.4; R Core Team, 2021). Fifty replicated datasets were generated for each simulation condition. The item and rater generating parameters were fixed across all 50 replications. Data generation for the fully-crossed complete rating design was relatively straightforward. However, the spiral design followed the layout shown in Table 9. The steps taken in generating the data for the fully-crossed and spiral rating designs are presented below:

Fully-crossed rating design:

Step 1: Generate ideal ratings using the person and item parameters. This was generated with GPCM.

Step 2: Generate observed ratings using the ideal ratings and rater bias and variability parameters.

Spiral rating design:

Step 1: Generate ideal ratings of each examinee using the examinee latent

trait and item parameters using the GPCM.

Step 2: Create four spirals. Each spiral represents each item.

Step 3:
- Assign Rater 1 and Rater 2 to Spiral 1.
- Assign Rater 3 and Rater 4 to Spiral 2.
- Assign Rater 5 and Rater 6 to Spiral 3.
- Assign Rater 7 and Rater 8 to Spiral 4.

Raters within each spiral are only raters who rated the item in that spiral.

Step 3: Generate observed ratings for all examinees within each spiral using the ideal ratings and rater bias and variability parameters.

The data generating syntax for the fully-crossed and spiral rating designs are documented in Appendix 2 and Appendix 3, respectively.

Table 11. Summary of Simulation Conditions

	Simulation Factor	Levels
Fixed	Rating scale response options	5 categories
	Distribution of examinee latent trait	Normal, $\mathcal{N}(\mu = 0, \sigma = 1)$
	Number of raters	8 raters
	Data generating model	HRM-GPCM
	Test length	4 items
Manipulated	Data analysis model	[1] HRM-PCM [2] HRM-GPCM
	Rating design	[1] Fully-crossed rating design [2] Spiral rating design
	Number of examinees	[1] 250 [2] 500
	Rater behavior	[1] 100% normal raters [2] 25% severe/inconsistent raters

3.4 Estimation of Model Parameters

As earlier noted, data were generated with the HRM-GPCM. All simulated data were analyzed using HRM-GPCM and HRM-PCM. Estimating the data with HRM-PCM introduces misfit because the discrimination parameter in the HRM-PCM is constrained to 1. Using Bayesian MCMC methods, model parameters were estimated by employing R2jags package (Su & Yajima, 2012) in R to interface with JAGS (Plummer, 2003).

Specification of prior distributions on all unknown parameters is one of the prerequisites of Bayesian estimation. Prior distributions can be informative, weakly informative, or noninformative, among others. Noninformative prior distributions were specified for all

estimated model parameters. Noninformative prior distributions reflect little knowledge about the model parameters. One of the benefits of noninformative prior distributions is that the results are similar to frequentist methods. The prior distributions chosen in the present study were similar to those employed by previous studies (e.g., Casabianca et al., 2017; Li et al., 2017; Patz et al., 2002). The prior distributions for the item discrimination follow a lognormal distribution ($\alpha_j \sim log\mathcal{N}(\mu = 0, \sigma^2 = 10)$) due to the non-negative requirement on the discrimination parameters. Lognormal priors or other priors like gamma priors allow α to be constrained to positive values. Normal priors were placed on the difficulty step parameters with mean of zero and variance of 10 ($\delta_{jv} \sim \mathcal{N}(\mu = 0, \sigma^2 = 10)$). Similarly, normal priors were placed on the rater bias parameter ($\phi_r \sim \mathcal{N}(\mu = 0, \sigma^2 = 10)$). The rater variability parameter of the HRM only takes on non-negative values. Hence, gamma priors were placed on the variability parameter, $\psi_r \sim Gamma(1,\ 1)$. To tackle potential identification issues, a normal prior with a mean of 0 and a variance of 1 was placed on the examinee latent traits ($\theta_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$).

For each dataset, two parallel chains, each with 40,000 iterations, were used. The first 15,000 iterations were discarded as burn-in. To reduce autocorrelation within the chains, the remaining 25,000 iterations were thinned. Every 10th iteration was retained. Thus, the resulting posterior distributions contained 5,000 iterations (2,500 iterations x 2 chains). The convergence of the chains was assessed using Gelman and Rubin's (1992) potential scale reduction factor, $\hat{R}$. For each parameter in the model, the $\hat{R}$ statistic was computed by comparing the *within-chain* variances to the *between-chain* variances. For example, if rater *r*'s bias (ϕ_r) parameter is of interest, then the $\hat{R}$ statistic associated with this parameter can be computed as:

$$\hat{R} = \sqrt{\frac{\hat{V}(\phi_r)}{W}}, \tag{3.1}$$

where W is the within-chain variance, and $\hat{V}(\phi_r)$ is weighted average of the between-chain (B) and within-chain variances. Assuming we have *M* chains with *N* number of iterations within each chain, then the computations for the between-chain and within-chain variances are provided in Equations 3.2 and 3.3, respectively.

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\bar{\phi}_{rm.} - \bar{\phi}_{r..})^2, \tag{3.2}$$

$$W = \frac{1}{M(N-1)}\sum_{m=1}^{M}\sum_{n=1}^{N}(\bar{\phi}_{rmn} - \bar{\phi}_{rm.})^2. \tag{3.3}$$

Subsequently, the weighted average of the between-chain and within-chain variances is computed as:

$$\hat{V}(\phi_r) = \left(\frac{N-1}{N}\right)W + \left(\frac{M+1}{MN}\right)B. \tag{3.4}$$

An $\hat{R}$ statistic close to 1 indicates that the chains have fully converged to the target posterior distributions. The number of iterations and burn-in samples provided here were chosen after preliminary evaluations and convergence assessments. In practice, $\hat{R}$ values smaller than 1.1 are considered acceptable fit but declare convergence prematurely (Gelman & Shirley, 2011). This study used an $\hat{R}$ value of 1.05 as the baseline value to assess the convergence of the chains. In addition to using the $\hat{R}$ statistic to assess convergence, trace plots of the estimated parameters were visually inspected to ensure adequate mixture of the chains.

3.5 Posterior Predictive Model Checking

The converged posterior distributions from the MCMC draws were used to perform the PPMC in R. For the PPMC, a thousand datasets were generated from the posterior predictive distributions. The following steps were taken:

Step 1: Randomly draw vector of parameter values from the joint posterior distribution.

Step 2: Use parameters drawn in Step 1 to simulate observed data based on the fitted model.

Step 3: Compute the discrepancy measures using the simulated data in Step 2.

Step 4: Repeat Steps 1–3 S times (S=1,000 times in this study).

Step 5: Compare the predictive and observed (or realized) discrepancy measures by calculating the proportion of times the predictive discrepancy measures exceeds the observed discrepancy measures.

3.5.1 Discrepancy Measures

In PPMC, there is no limit to the number of discrepancy measures that can be employed. This study considered discrepancy measures at three levels: test level, item level, and rater level. Some of the discrepancy measures employed have been previously used with traditional IRT models.

***Test-level measure*.** Casabianca et al. (2017) and Nieto and Casabianca (2019) both used the total score distribution for evaluating model-data fit of the L-HRM and M-HRM, respectively. The total score distribution is the distribution of the total score received by every examinee on the test. As a reminder, the total score is the sum of the item scores. An item score (denoted by $\bar{X}_{ij.}$) is the average rating given to an examinee by multiple raters on a particular

item. A snapshot of how the total score distribution is computed is shown in Table 12. The last column on the table is the total score. As shown in Table 12, Examinee 1 was scored by multiple raters on Item 1. This examinee received scores of 4, 3, and 3 on Item 1 from Raters 1, 2, and 3, respectively. This examinee's item score, computed as the average of the three ratings, was 3.33, and the item score on Item 2 was 4.00. Hence, Examinee 1 received a total score of 7.33 (3.33 + 4.00 = 7.33). The total score distribution is the distribution of the total score received by every examinee on the test.

Eight summary statistics were computed from the total score distribution: mean, standard deviation, first quartile, third quartile, minimum, maximum, skewness and kurtosis. These discrepancy measures capture the shape, spread, and center of the total score distribution.

Table 12. Hypothetical Total Score Distribution

Examinee	Item 1			Item 2			Item score		Total Score
	Rater			Rater			Item		
	R1	R2	R3	R1	R2	R3	1	2	
1	4	3	3	4	4	4	3.33	4.00	7.33
2	4	3	4	4	3	5	3.67	4.00	7.67
3	2	2	2	4	5	4	2.00	4.33	6.33
4	3	3	5	4	3	2	3.67	3.00	6.67
5	2	2	4	4	3	3	2.67	3.33	6.00

Note. R1 = Rater 1, R2 = Rater 2, R3 = Rater 3

***Item-level measure*.** Two discrepancy measures were employed at the item-level. The item-level discrepancy measures are specific to each item or pair of items. Hence, the performance of the discrepancy measures will be summarized for each item or item pairs. The first measure was the item-total correlation. The Item-total correlation is the correlation between

a particular item's score and the total score. Li et al. (2017) and Sinharay et al. (2006) found the item-total correlation to be effective in detecting misfit when the generating model included a discrimination parameter, but the analysis model did not. Going back to the illustration in Table 12, the item-total correlation for Item 1 is the correlation between the item score for Item 1 (i.e., Column 8) and the total score (i.e., Column 10). Crocker and Algina's (1986) recommendation to remove an item's score from the total was employed in the current study.

The odds ratio was chosen as the second discrepancy measure at the item level. Odds ratio is a pairwise measure that captures the association between two items. The odds ratio is computed by employing Equation 2.35 (described in Chapter 2). Using PPMC, the odds ratio has been shown to perform relatively well in detecting model misfit due to functional form misspecification (Li et al., 2017; Sinharay & Johnson, 2003) and due to multidimensionality (Levy et al., 2009). The former is of interest in this study since the current study generated data from a unidimensional model. To implement the odds ratio, polytomous or continuous data need to be dichotomized. Li et al. (2017) dichotomized their items with five response options by recoding scores of 3 and 4 to 1 and scores of 0, 1, and 2 were recoded as 0. As earlier defined, item score in this study is the average rating given to an examinee by multiple raters on a particular item. As seen in Table 12, the item scores can take on non-discrete values. Therefore, this study elected to dichotomize the scale using 2.50. The odds ratio was implemented by dichotomizing the item scores such that

$$\bar{X}^{*}_{ij.} = \begin{cases} 1 & \text{if} \quad \bar{X}_{ij.} \geq 2.50 \\ 0 & \text{if} \quad \bar{X}_{ij.} < 2.50, \end{cases} \tag{3.5}$$

where $\bar{X}_{ij.}$ is examinee *i*'s item score on item *j* and $\bar{X}^{*}_{ij.}$ is examinee *i*'s dichotomized item score on item *j*. For example, an item score 3.67 will receive a score of 1 and an item score of 2.00 will

receive a score of 0. The odds ratio was computed for all item pairs: (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), and (3, 4).

***Rater-level measure*.** Performance assessment models have not been rigorously evaluated using PPMC; hence, there is little guidance on the performance of different discrepancy measures at the rater-level. Three rater-level discrepancy measures were explored in this study: score-estimate correlation, rater-total correlation, and rater standard deviation.

The score-estimate correlation is the correlation between the ratings of a single rater (i.e., rater's scores) and the ability estimates of the examinees. A rater's score is defined as the average rating awarded by a particular rater to an examinee across all items. The formulation of the score-estimate correlation is presented in Section 2.7. Literature suggests that the score-estimate correlation is useful in detecting rater inconsistency (Wolfe, 2014; Wolfe & McVay, 2010). Misfit of the HRM was introduced in the functional form of the IRT component. It is expected that constraining the discrimination parameter to 1 for the HRM_{PCM} will impact the estimated examinee latent traits. Hence, the score-estimate correlation may prove useful in detecting rater-level misfit when the IRT functional form is misspecified.

The rater-total correlation is another rater-level discrepancy measure considered. The rater-total correlation is the correlation between a rater's score and the total score (i.e., the sum of the item scores). The rater-total correlation is analogous to the SR-ROR correlation that was described in Chapter 2. Essentially, rater-total correlation, like score-estimate correlation, reflects how consistent a single rater is compared to the rest of the raters. Finally, rater standard deviation was also employed as a rater-level discrepancy measure. The rater standard deviation is computed using the ratings of a single rater. The twelve discrepancy measures investigated in the current work are summarized in Table 13.

Table 13. Summary of Discrepancy Measures

Level	Discrepancy Measure	Description
Test level	*Total score distribution* • mean • standard deviation • first quartile • third quartile • minimum • maximum • skewness • kurtosis	The total score distribution is "the sum of the item score." The item score is the average rating given to an examinee by multiple raters on a particular item.
Item level	*Item-total correlation*	Item-total correlation is the correlation between a particular item's score (defined above) and the total score
	Odds ratio	Odds ratio is the measure of association between pairs of items
Rater level	*Score-estimate correlation*	The score estimate correlation is the correlation between a rater's score and the ability estimates of the examinees. The rater's score is the average rating awarded by a rater to an examinee across all items
	Rater-total correlation	The rater-total correlation is the correlation between a rater's score and the total score (i.e., the sum of the item score).
	Rater standard deviation	The rater standard deviation is the standard deviation of a rater's score to a group of examinees.

3.5.2 Outcome Measures

The main objective of this study was to explore the effectiveness of different discrepancy measures in detecting (mis)fit of the HRM when a model is correctly or incorrectly specified. To investigate this, the functional form of the IRT was misspecified for some conditions. The ability of the chosen discrepancy measures in detecting misfit was assessed by computing the PPP-values. The PPP-value is the proportion of replicated datasets for which the observed discrepancy measure exceeds the discrepancy measure of the replicated dataset. The PPP-value is computed over the posterior predictive distribution and is mathematically given as

$$p_B = \Pr\{T(D_{rep}) \geq T(D_{obs}) \big| \theta\}, \tag{3.7}$$

where p_B is the probability that the replicated data $T(D_{rep})$ could be more extreme than the observed data $T(D_{obs})$ as measured by the discrepancy measure. Essentially, the fraction of times that $T(D_{rep}) \geq T(D_{obs})$ is computed which implies that PPP-values take on values between 0 and 1. A PPP-value close to 0.5 indicates adequate model-data fit. Small values less than 0.05 or large values greater than 0.95 are indicative of poor model-data fit. A PPP-value of 0.05 indicates that 5% of $T(D_{rep})$ are less than $T(D_{obs})$, and 95% of $T(D_{rep})$ are greater than $T(D_{obs})$. Using the guidelines suggested by Sinharay (2006), a model was judged to have evidence of misfit when the PPP was less than 0.05 or greater than 0.95.

The PPP-values were used to compute the Type I error rates and power for each discrepancy measure. Type-I error was computed as the proportion of times a correctly specified model shows evidence of misfit. An alpha level of 5% was considered as an evidence of adequate model-data fit for this dissertation. It was expected that the Type I error rates of the discrepancy measures would not exceed 5% when the correct calibration model was specified. For each simulation condition, power was computed as the proportion of times an incorrectly

specified model shows evidence of misfit.

In addition, logistic regression was used to model the probability of detecting misfit of the HRM-PCM. The dependent variable for the logistic regression was whether or not the HRM-PCM was flagged for misfit. The independent variables were the simulation factors (i.e., rating design, rater effects, and sample size). Odds ratio coefficients from the logistic regression analyses were reported.

3.6 Chapter Summary

Chapter 3 outlined the methodology used in this study. Details of the simulation design, guided by previous work in performance assessments, were provided. The data generating parameters and steps for generating the data for the complete and incomplete rating designs were detailed. The estimation procedures of the model parameters were extensively discussed including the prior distributions, number of iterations and burn-in samples considered, and how convergence was assessed using PSRF. The selected discrepancy measures for this study were discussed. Finally, the outcome measures and guidelines for assessing the effectiveness of the discrepancy measures were provided.

CHAPTER 4:

RESULTS

The purpose of this study was to evaluate the performance of PPMC in detecting misfit of the HRM at the test-, item-, and rater-level. The number of examinees (N = 250 and 500), rating design (fully-crossed and spiral rating designs), rater effects (no rater effects and 25% of raters with rater effects), and analysis model (HRM-GPCM and HRM-PCM) were varied, leading to 16 fully-crossed conditions. This chapter summarizes the findings of the simulations. The power and Type I error rates of the different discrepancy measures employed are presented in this chapter. In addition, the effects of the design factors are outlined in this chapter.

4.1. Data Features

All data analyzed in this study were generated using HRM-GPCM. The descriptive statistics of two simulation conditions are presented to show the score distributions of the two simulation conditions and how they compare across raters. Table 14 presents the rater mean and standard deviation for a single simulation replication for two conditions: (a) 500 examinees with fully-crossed rating design condition without rater effects and (b) 500 examinees with fully-crossed rating design with rater effects. In Table 14, it can be seen that the rater mean was approximately 3 for raters with bias parameters close to zero (i.e., raters exhibiting no rater effects). For example, in the condition without rater effects, Raters 2, 5, 6, and 8 had mean scores of 3.024, 2.928, 3.012, and 2.882, respectively. These mean scores lie around the center of the 5-response scale point used in this study (i.e., score range of 1 to 5), indicating that the average ratings of the raters with bias parameters close to zero fall around the center of the score distribution. As expected, raters simulated to have rater severity effects had the smallest rater

means. Raters 6 and 8 were simulated to show significant negative bias (i.e., severity) and had rater means of 2.372 and 2.202, respectively.

The rater score distributions of the two conditions presented in Table 14 are further depicted in Figures 5 and 6. Figure 5 shows the score distribution of the eight raters in this study for the condition with 500 examinees, no rater effects, and a fully-crossed design. Figure 5 shows that raters with bias parameters close to zero have distributions that are symmetric compared to raters who have significant non-zero bias parameters (Figure 6). For example, in Figure 5, Raters 6 and 8, with bias parameters of 0.033 and -0.060, respectively, mostly asigned scores in the middle category. In the condition with rater effects, data were simulated using significant bias parameters for Raters 6 and 8. As shown in Figure 6, the score distributions of Raters 6 and 8 were positively skewed, indicating that these two raters are less likely to use higher score categories such as 4 and 5. It could also be seen that Raters 1 and 3 have score distributions that mimic the rater parameters used in generating the data. Rater 1 has a negative bias parameter indicating that this rater is more likely to use lower score categories such as 1, 2, and 3. Conversely, Rater 3 has a positive bias parameter, which indicates that this rater is more likely to use higher score categories. The score distributions of Raters 1 and 3 clearly depict what was expected based on the bias parameters of these raters.

Table 14. Rater Descriptive Statistics for a Single Simulation Replication

N	Rating design	Rater Effect	Rater	Rater Parameters		M	SD
				ϕ_r	ψ_r		
500	Full	No Rater Effect	1	-0.224	0.425	2.756	1.080
			2	0.046	0.270	3.024	1.023
			3	0.397	0.403	3.338	1.136
			4	-0.283	0.740	2.776	1.166
			5	-0.020	0.539	2.928	1.123
			6	0.033	0.604	3.012	1.111
			7	0.177	0.561	3.134	1.178
			8	-0.060	0.597	2.882	1.134
500	Full	Rater Effect	1	-0.224	0.425	2.748	1.118
			2	0.046	0.270	2.978	1.083
			3	0.397	0.403	3.330	1.122
			4	-0.283	0.740	2.758	1.139
			5	-0.020	0.539	2.926	1.135
			6	-0.931	1.487	2.372	1.131
			7	0.177	0.561	3.236	1.152
			8	-1.055	0.980	2.202	1.056

Note. M = mean; SD = standard deviation; N = number of examinees; ϕ_r = rater *r*'s bias parameter; ψ_r = rater *r*'s variability parameter

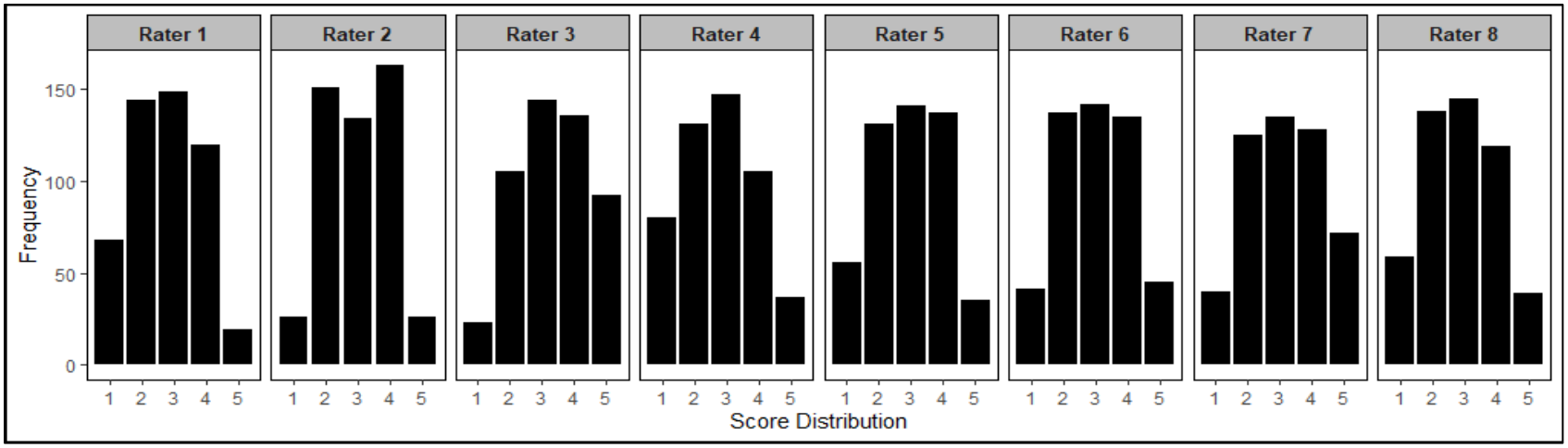

Figure 5. Rater score distribution for a single item generated with 500 examinees using a fully-crossed rating design with raters with no rater effects.

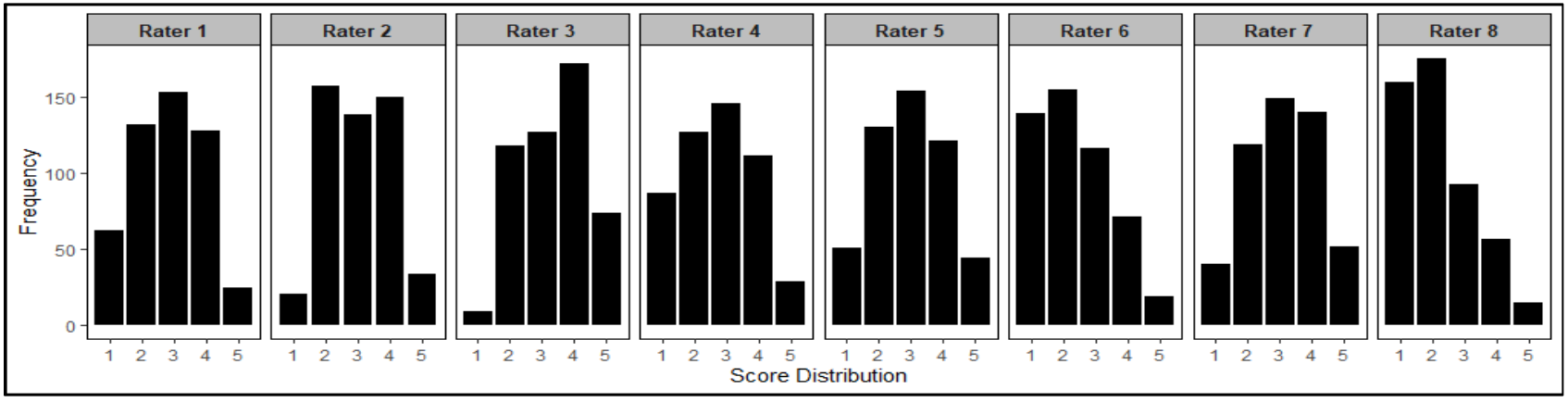

Figure 6. Rater score distribution for a single item generated with 500 examinees using a fully-crossed rating design with some raters exhibiting rater effects.

Research Question 1: *What is the Type I error rate and power of the test-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?*

4.2 Test-Level Discrepancy Measures

Total Score Distribution

The total score, computed at the test level, is the sum of the item scores. An item score is the average rating given to an examinee by multiple raters on a particular item. Figure 7 illustrates an example of a total score distribution of a single replication for the condition with 500 examinees, no rater effects, and with a fully-crossed rating design. The distribution of the total score shown in Figure 7 was roughly symmetric, with an average total score of 11.43 (*SD* = 2.91) and a skewness statistic of -0.03. The score distribution ranged from 5.00 to 18.50.

The summary statistics of the total score distribution shown in Figure 8 resembles statistics from Figure 7. However, the mean of the total score distribution for Figure 8 was 10.77 (*SD* = 2.71), which is slightly smaller than the mean from Figure 7. In addition, the total scores of the lowest and highest scoring examinees were 4.75 and 18.13, respectively. Data for Figure 8 were generated under the condition with rater effects (i.e., N=500, rater effects, and fully-crossed design).

Eight summary statistics were computed from the total score distribution for the PPMC procedures: 1) mean, 2) standard deviation, 3) first quartile, 4) third quartile, 5) minimum, 6) maximum, 7) skewness, and 8) kurtosis. These discrepancy measures capture the shape, variability, and center of the total score distribution. For each measure, this study separates results into subsets for which the generating model (GM) equals analysis model (AM), in which the generating model (HRM-GPCM) and analysis model match (HRM-GPCM), denoted by GM=AM. In contrast, GM≠AM is the other subset, in which the generating model (HRM-

GPCM) and analysis model do not match (HRM-PCM). The hypothesis is that when the GM=AM, discrepancy statistics will show model-data fit resulting to a low Type I error rate and when GM≠AM, discrepancy statistics may detect misfit resulting in high power.

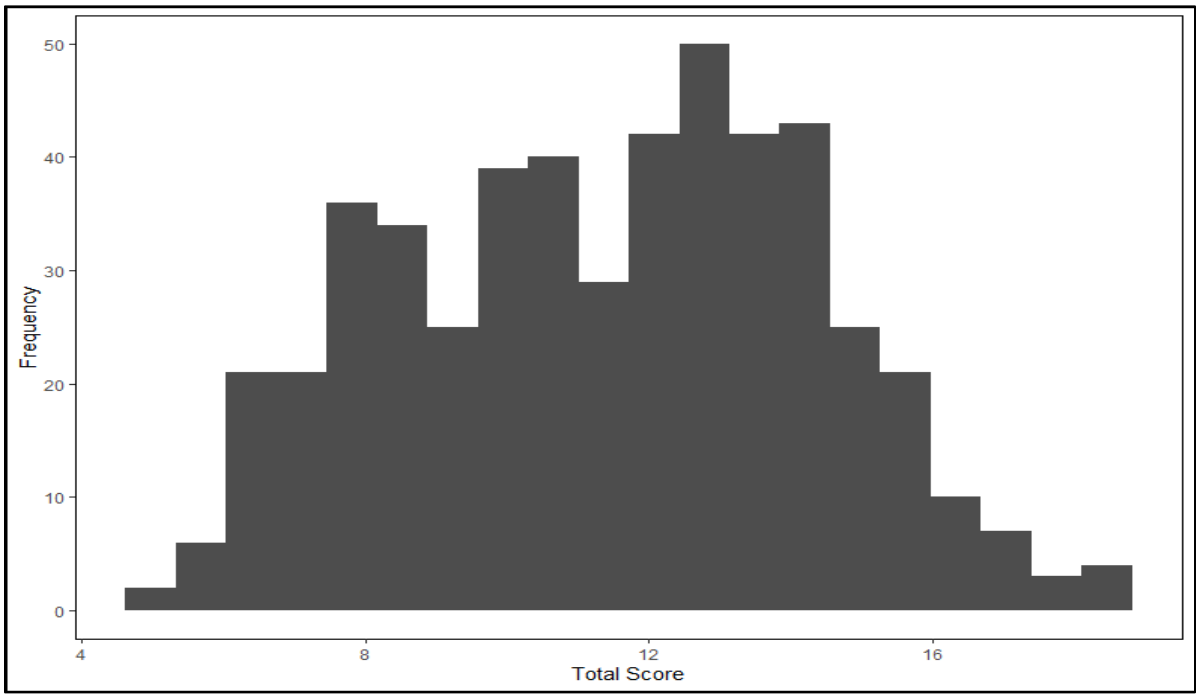

Figure 7. Total score distribution for data generated with 500 examinees using a fully-crossed rating design with raters with no rater effects.

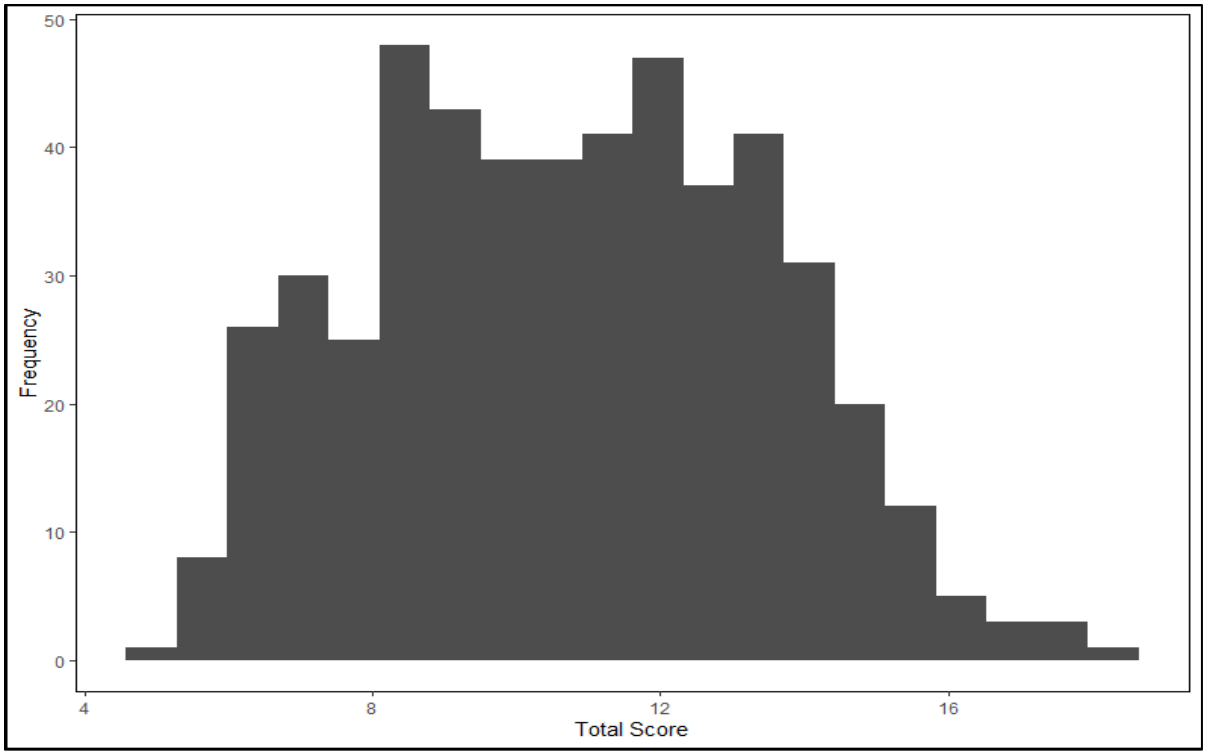

Figure 8. Rater score item distribution for data generated with 500 examinees using a fully-crossed rating design with some raters exhibiting rater effects.

4.2.1 Mean Discrepancy Measure

4.2.1.1 Summary of Observed and Replicated Mean of Total Score Distribution

The observed mean estimates of the total score distribution for the 50 replications used in this study are presented in Figure 9. Each boxplot in Figure 9 contains 50 values – each value is the observed mean of the total score distribution of a set of simulation conditions for the i^{th} replication. In general, conditions with raters with no rater effects have higher means compared to conditions with rater effects. The shape of the distributions, across all simulation conditions, appears to be roughly symmetric.

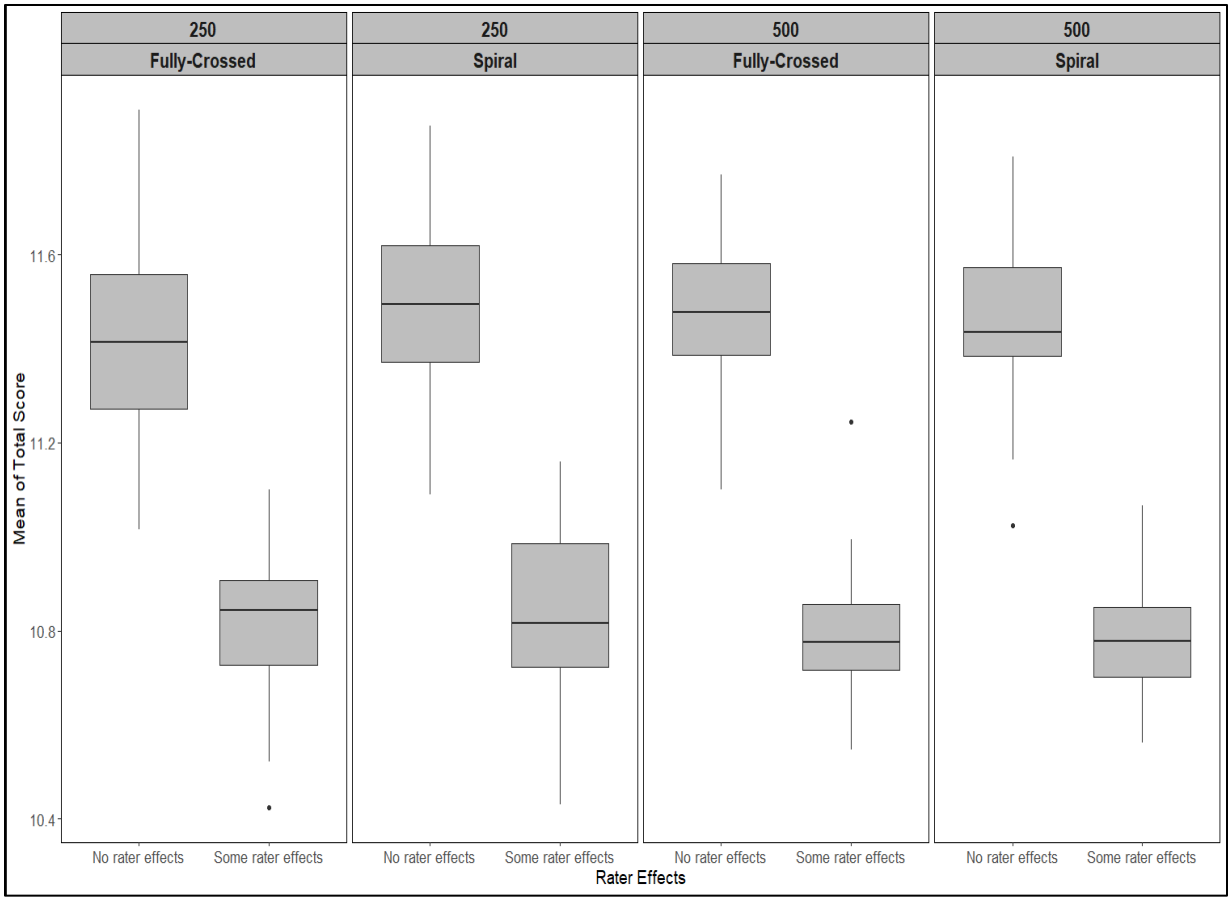

Figure 9. Observed mean of total score distribution across simulation conditions

GM=AM. For each i^{th} observed mean of the total score distribution, there are 1000 PPMC replicated means of the total score distribution based on the posterior predictive samples. The PPMC replicated means at the 5th, 50th, and 95th percentiles were compared to the observed mean. This comparison shown in Figure 10 illustrates that the observed means closely resemble the PPMC replicated means at the three percentiles. For example, the smallest observed mean of the total score was 11.01 for the condition with 250 examinees, fully-crossed rating design, and no rater effects (Panel 1 in the top row of Figure 10). When the data were fitted with the correct model (i.e., HRM-GPCM), the mean of the total score of the replicated dataset at the 5th, 50th, and 95th percentiles were 10.84, 11.02, and 11.20, respectively.

GM≠AM. Fitting the data with HRM-PCM (i.e., incorrectly specified model) resulted in means of 10.79, 11.01, and 11.22 at the 5th, 50th, and 95th percentiles, respectively. The replicated datasets based on the posterior predictive samples when the HRM was correctly and incorrectly specified resulted in similar mean estimates of the total score distribution. This suggest that the mean of the total score distribution is not an effective discrepancy measure for detecting misfit of the HRM at the test level.

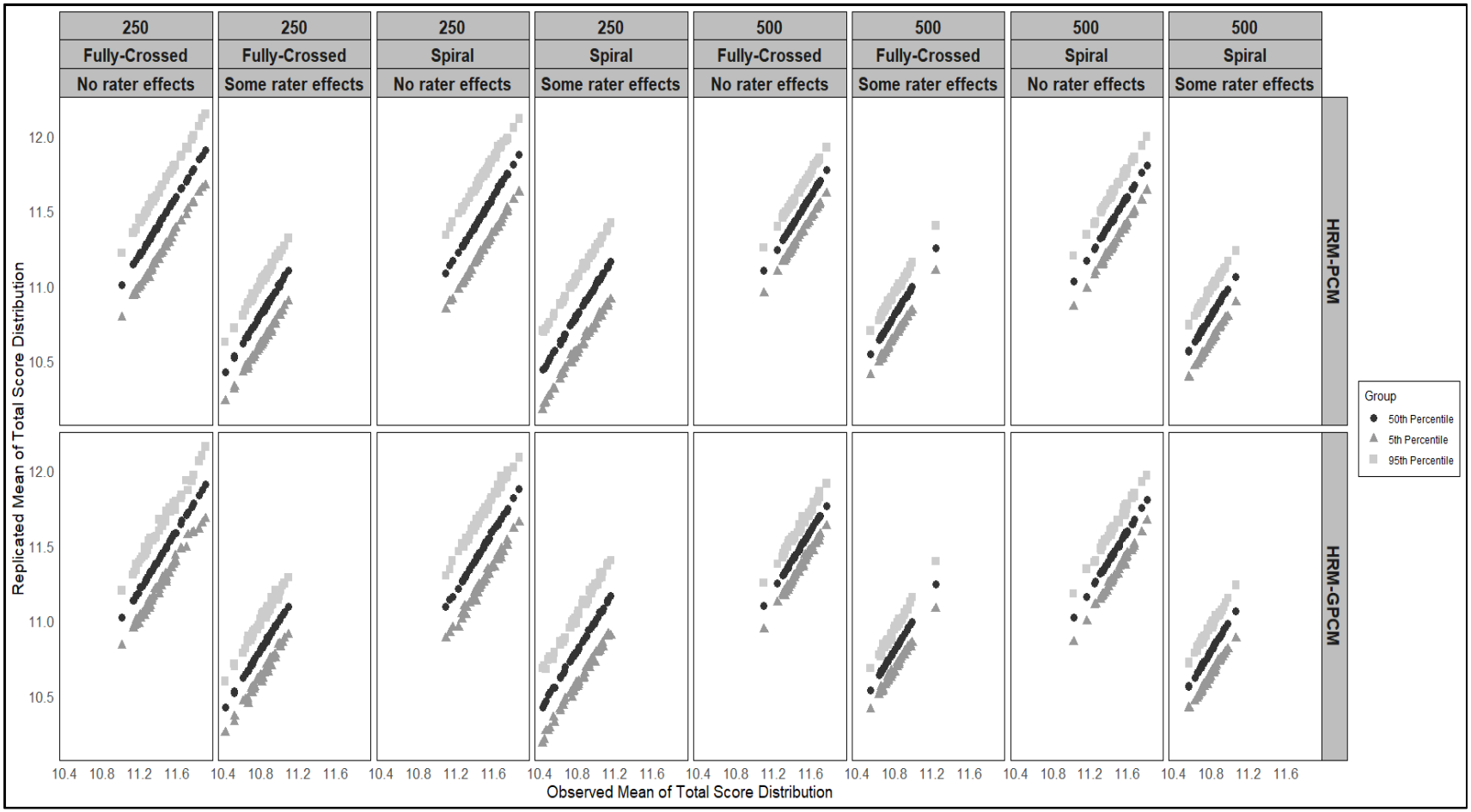

Figure 10. Scatterplot of observed mean of total score distribution and replicated median, 5th, and 95th percentile of the total score distributions based on the posterior predictive samples across simulation conditions. The top row represents GM≠AM and bottom row GM=AM.

4.2.1.2 Type I Error and Power of Mean Discrepancy Measure

The PPP-values resulting from comparing the observed and replicated means across the simulation conditions are presented in Figure 11. Figure 11 has 16 boxplots. Each boxplot contains 50 PPP-values corresponding to 50 replications for each set of simulation conditions used in this study. There are two boxplots within each panel - the first boxplot contains PPP-values for data analyzed using the misfitted model (i.e., HRM-PCM) and one second boxplot contains PPP-values for data analyzed with the correct model (i.e., HRM-GPCM). Overall, the distributions of the PPP-values center around 0.50. Across all sets of conditions, the PPP-values were between 0.45 and 0.57. PPP-values around 0.50 suggest good model-data fit (Gelman et al., 1996).

GM=AM. Previous studies (e.g., Sinharay, 2006) have used PPP-values less than 0.05 or greater than 0.95 to suggest evidence of a misfitting model. The proportions of extreme PPP-values are presented in Table 15. The Type I error rate is the proportion of extreme PPP-values when the generating model is the same as the analysis model. Essentially, this is the proportion of times the HRM-GPCM does not fit the data. As shown in Table 15, all Type I error rates when GM=AM are 0% across the simulation conditions. A Type I error rate of 0% implies that, out of the 50 replications used in this study, the HRM-GPCM was never flagged for misfit using the mean discrepancy measure.

GM≠AM. The power rates reported in Table 15 were all 0% in all simulation conditions employed in this study. Power was used to show evidence of misfit. Power is the proportion of extreme PPP-values when the generating model is different from the analysis model (i.e., data are generated with HRM-GPCM but analyzed with HRM-PCM). A power rate of 0% implies that, when the mean discrepancy measure was employed, the HRM-PCM was not flagged for

misfit in all 50 replications in the current study.

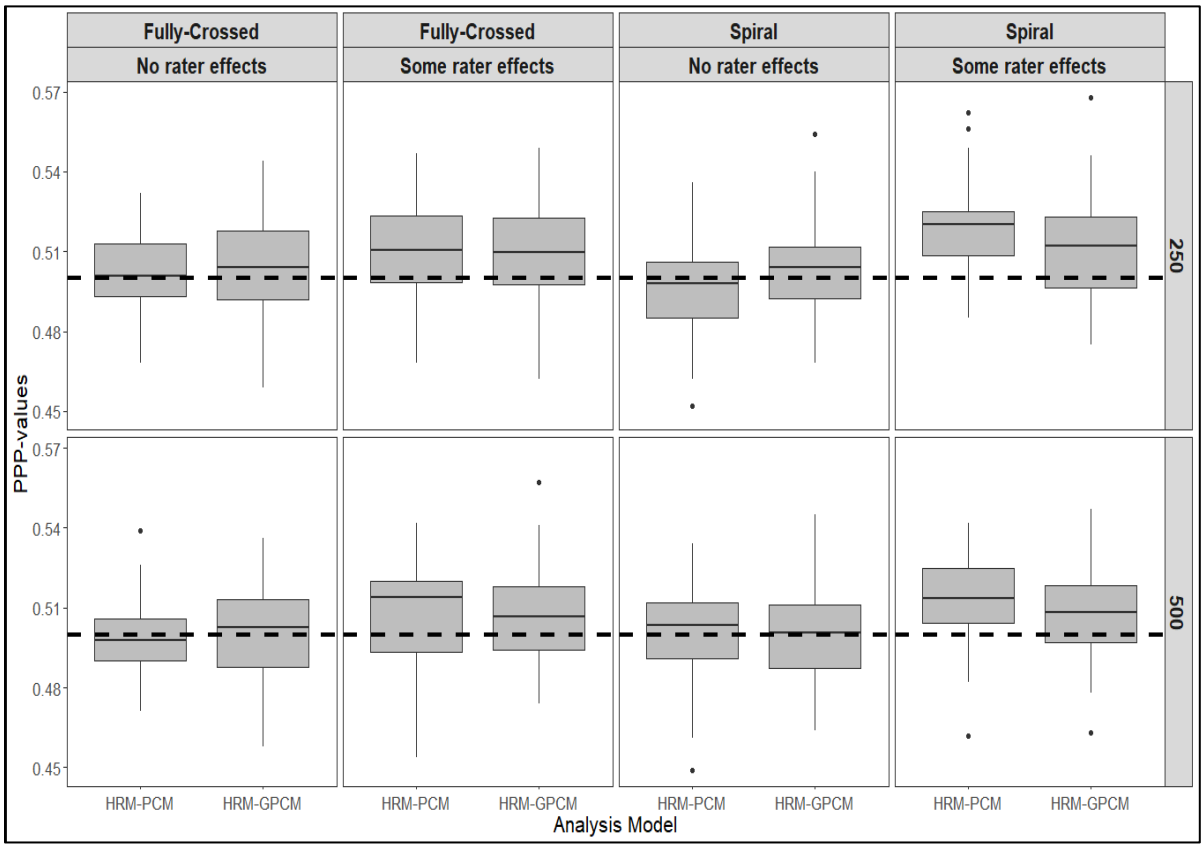

Figure 11. Distribution of PPP-values for the mean of total score distribution discrepancy measure across simulation conditions. The dashed horizontal line represents where the PPP-value is 0.50

Table 15. Type I Error Rates and Power for the Mean of the Total Score Discrepancy Measure

Examinee	Rater Effect	Rating Design	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)	$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)
250	No Rater Effect	Full	0.00	0.00
		Spiral	0.00	0.00
	Rater Effect	Full	0.00	0.00
		Spiral	0.00	0.00
500	No Rater Effect	Full	0.00	0.00
		Spiral	0.00	0.00
	Rater Effect	Full	0.00	0.00
		Spiral	0.00	0.00

Note. GPCM = generalized partial credit model; PCM = partial credit model.

4.2.2 Standard Deviation Discrepancy Measure

4.2.2.1 Type I Error and Power of Standard Deviation Discrepancy Measure

Standard deviation was used to measure the spread of the total score distribution. The distributions of the resulting PPP-values across the simulation conditions are presented in Figure 12. As shown in Figure 12, the boxplots of the PPP-values when the analysis model was HRM-GPCM are narrower and have values that ranged from 0.13 to 0.74. However, there were noticeably more variability and large extreme values in the boxplot of the PPP-values when HRM-PCM was the analysis model.

GM=AM. Table 16 displays the proportions of extreme PPP-values of the standard deviation discrepancy measure. The Type I error rates, when HRM-GPCM was used in analyzing the data, were all zero percent across the simulation conditions. This suggests that the replicated standard deviations of the total scores based on the posterior predictive samples were similar to the observed standard deviations of the total scores.

GM≠AM. The power rates for the standard deviation discrepancy measure are also presented in Table 16. The power rates ranged from 24% to 76% when HRM-PCM was used in analyzing the data, indicating that the standard deviation of the total score performed moderately well in detecting misspecification of HRM-PCM at the test level. The performance of the standard deviation discrepancy measure when the model was misspecified (i.e., HRM-GPCM → HRM-PCM) varied across simulation conditions. Conditions with 500 examinees performed better than conditions with 250 examinees. For instance, the power was 76% for the set of conditions with 500 examinees, no rater effects, and with fully-crossed design in comparison to 62% for the condition with 250 examinees, no rater effects, and with fully crossed design. Within conditions with the same number of examinees, the conditions with no rater effects outperformed

the conditions with rater effects. For example, when the number of examinees was 250, the power was 48% for the condition without rater effects generated with spiral design compared to 24% for the condition with rater effects generated with spiral design. In addition, the power rates suggest that the standard deviation of the total score is more useful in detecting misspecification of the HRM-PCM under a fully-crossed rating design compared to a spiral rating design. For example, in conditions with 500 examinees, the power rates under the fully-crossed rating design were 76% (without rater effects) and 66% (with rater effects) compared to power rates of 58% under the spiral rating design.

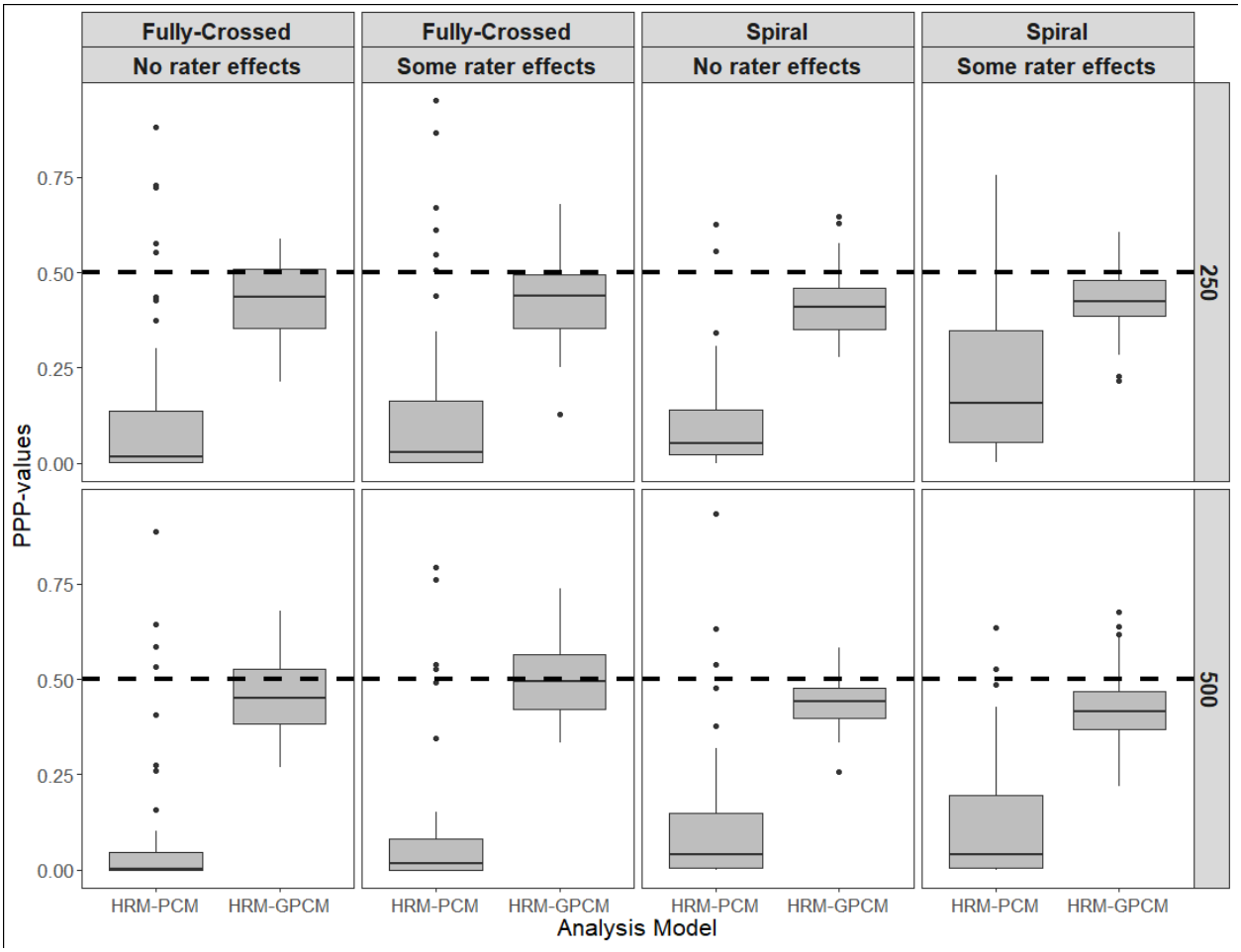

Figure 12. Distribution of PPP-values for the standard deviation of total score distribution discrepancy measure across simulation conditions. The dashed horizontal line represents where the PPP-value is 0.50

Table 16. Type I Error Rates and Power for the Standard Deviation of the Total Score Discrepancy Measure

Examinee	Rater Effect	Rating Design	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)	$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)
250	No Rater Effect	Full	0.00	0.62
		Spiral	0.00	0.48
	Rater Effect	Full	0.00	0.58
		Spiral	0.00	0.24
500	No Rater Effect	Full	0.00	**0.76**
		Spiral	0.00	0.58
	Rater Effect	Full	0.00	0.66
		Spiral	0.00	0.58

Note. GPCM = generalized partial credit model; PCM = partial credit model. Power above 70% is bolded

4.2.3 First and Third Quartiles Discrepancy Measures

4.2.3.1 Type I Error and Power of First and Third Quartiles Discrepancy Measures

GM=AM. Table 17 presents the Type I error and power rates of the first and third quartiles of the total score discrepancy measures. The 25% and 75% percentiles allow for the assessment of how well values at the extreme ends of the distribution were adequately captured by the fitted models. When the generating model is the same as the analysis model (i.e., HRM-GPCM → HRM-GPCM), the results show that the proportions of extreme PPP-values (i.e., Type I error rates) for the first quartile discrepancy measure ranged from 0.00 to 0.38. Interestingly, all the conditions under the fully-crossed rating design had Type I error rates below 5%. However, the Type I error rates of the first quartile discrepancy measure were inflated under the spiral rating design. Similar patterns in the Type I error rates were observed for the third quartile discrepancy measure. For the correctly specified model with a fully-crossed design, all Type I error rates for the third quartile discrepancy measure were less than 5%, while the Type I error

rates were inflated under the spiral design. The patterns observed suggest that the tails of the total score distribution was better captured under a fully-crossed rating design.

GM≠AM. The proportions of extreme PPP-values of the first and third quartiles of the total score discrepancy measures for the misspecified model (i.e., power) were small. Most conditions resulted in power less than 50%. Only one condition with 500 examinees, under spiral design, and no rater effect resulted in power above 70%. These low power rates suggest that the first and third quartiles may not be useful measures of detecting misfit on the HRM-PCM.

Table 17. Type I Error Rates and Power for the First and Third Quartiles of the Total Score Discrepancy Measures

Examinee	Rater Effect	Rating Design	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)		$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)	
			Q1	Q3	Q1	Q3
250	No Rater Effect	Full	0.00	0.02	0.18	0.24
		Spiral	0.10	0.28	0.00	0.54
	Rater Effect	Full	0.00	0.00	0.24	0.26
		Spiral	0.22	0.16	0.06	0.34
500	No Rater Effect	Full	0.02	0.00	0.38	0.50
		Spiral	0.38	0.42	0.14	**0.80**
	Rater Effect	Full	0.04	0.04	0.32	0.28
		Spiral	0.44	0.44	0.16	0.68

Note. Q1 = first quartile; Q3 = third quartile; GPCM = generalized partial credit model; PCM = partial credit model. Power above 70% is bolded

4.2.4 Minimum and Maximum Discrepancy Measures

4.2.4.1 Type I Error and Power of Minimum and Maximum Discrepancy Measures

The minimum and maximum values of the total score distribution also served as discrepancy measures. The Type I error rates and power of these two discrepancy measures are presented in Table 18.

GM=AM. For both the minimum and maximum discrepancy measures, the Type I error rates were relatively large. The Type I error rates were within the range of 2% and 22% suggesting that the model performed poorly in predicting the tails of the total score distribution.

GM≠AM. The power for the minimum discrepancy measure ranged from 6% to 18% and ranged from 8% to 26% for the maximum discrepancy measure. These results suggest that the minimum and maximum discrepancy measures underperformed in detecting misfit of the HRM-PCM.

Table 18. Type I Error Rates and Power for the for the Minimum and Maximum values of the Total Score Discrepancy Measures

Examinee	Rater Effect	Rating Design	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)		$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)	
			Minimum	Maximum	Minimum	Maximum
250	No Rater Effect	Full	0.10	0.16	0.18	0.16
		Spiral	0.22	0.12	0.10	0.22
	Rater Effect	Full	0.04	0.02	0.06	0.08
		Spiral	0.20	0.10	0.14	0.10
500	No Rater Effect	Full	0.12	0.08	0.12	0.26
		Spiral	0.10	0.16	0.08	0.26
	Rater Effect	Full	0.04	0.06	0.08	0.10
		Spiral	0.22	0.08	0.18	0.18

Note. GPCM = generalized partial credit model; PCM = partial credit model.

4.2.5 Skewness and Kurtosis Discrepancy Measures

4.2.5.1 Type I Error and Power of Skewness and Kurtosis Discrepancy Measures

GM=AM. The performance of the fitted models in predicting the shape of the total score distribution was assessed using skewness and kurtosis statistics as the discrepancy measures. The Type I error rate and power for these measures are presented in Table 19. In all conditions, the Type I error rates of the skewness discrepancy measure were all less than 5%. For the kurtosis discrepancy measure, only one set of simulation conditions (N=500, no rater effect, and spiral design) resulted in Type I error rate above 5%. This indicates that HRM-GPCM performed adequately well in modeling the shape of the total score distribution.

GM≠AM. When the analysis model was different from the generating model, the proportions of extreme PPP-values for the skewness discrepancy measure ranged from 0.00 (0%) to 0.02 (2%). This suggests that replicated datasets based on data fitted with HRM-PCM result in similar skewness statistics as the observed data (i.e., data generated using HRM-GPCM). This provides evidence that the skewness discrepancy measure may not be useful in detecting misfit of the HRM-PCM. The performance of the kurtosis discrepancy measure in detecting misfit of the HRM-PCM was also low. The power ranged from 4% to 26%. The largest power was observed for conditions without rater effects with large sample size (i.e., N = 500).

Table 19. Type I Error Rates and Power for the Skewness and Kurtosis Statistics of the Total Score Discrepancy Measures

Examinee	Rater Effect	Rating Design	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)		$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)	
			Skewness	Kurtosis	Skewness	Kurtosis
250	No Rater Effect	Full	0.00	0.04	0.00	0.18
		Spiral	0.00	0.00	0.00	0.14
	Rater Effect	Full	0.00	0.02	0.00	0.06
		Spiral	0.00	0.00	0.00	0.04
500	No Rater Effect	Full	0.00	0.00	0.00	0.22
		Spiral	0.00	0.08	0.00	0.26
	Rater Effect	Full	0.00	0.04	0.00	0.16
		Spiral	0.04	0.00	0.02	0.08

Note. GPCM = generalized partial credit model; PCM = partial credit model.

Summary of Test-Level Discrepancy Measures

The choice of discrepancy measure is important to implementing the PPMC technique in order to make accurate conclusion of aspects of the data captured by the model. The findings of the current study reveal that the mean, standard deviation, skewness, and kurtosis of the total score distribution were well captured by the HRM-GPCM. Overall, the Type I error rates of the mean, standard deviation, skewness, and kurtosis discrepancy measures were less than 5%. However, for some simulation conditions, the tails of the total score distribution were not adequately captured by the HRM-GPCM. The minimum, maximum, first quartile, and third quartile discrepancy measures of these simulation conditions resulted in Type I error rates significantly greater than 5%.

Furthermore, the results of the eight discrepancy measures assessed at the test-level suggest that only the standard deviation of the total score distribution was useful in detecting misfit of the HRM-PCM. However, only one simulation condition resulted in power greater than 70%. Examining the results more closely, Table 20 presents factors that are associated with the misfit of the HRM-PCM for the standard deviation discrepancy measure. A logistic regression was performed to ascertain the effects of sample size, rater effects, and rating design on whether or not the HRM-PCM was detected for misfit. The results indicate that the main effects of number of examinees, rating design, and rater effects were all significant predictors of misfit of the HRM-PCM using the standard deviation discrepancy measure. The number of examinees had the largest odds ratio of 4.382 ($\beta_{Sample\ Size} = 1.478, p = 0.0008, OR = 4.382$). This statistic implies that, using the standard deviation as the discrepancy measure, the odds of detecting misfit of the HRM-PCM for a sample size of 500 was 4.382 higher in comparison to a sample size of 250, holding other variables constant.

Rating design had the second largest odds ratio of 4.373 ($\beta_{Rating\ Design} = 1.476, p =$

$0.0008, OR = 4.373$). These findings indicate that the odds of detecting misfit of the HRM-PCM under the fully-crossed rating design compared to the spiral design was 4.373 higher, holding all other variables constant. Additionally, holding all other factors constant, the odds of detecting misfit of the HRM-PCM using the standard deviation was higher when the raters had no rater effects compared to when raters exhibited some rater effects ($\beta_{Rater\ Effects} = 1.073, p = 0.0138, OR = 2.923$).

Table 20. Results of Logistic Regression to Predict Misfit of HRM-PCM of the Standard Deviation Discrepancy Measure

Effect	Estimate	*z*-value	p-*value*	OR
Intercept	-1.153	-3.481	0.0005	0.316
Sample size	1.478	3.370	0.0008	4.382
Rating design	1.476	3.369	0.0008	4.373
Rater effects	1.073	2.462	0.0138	2.923
Sample size x rating design	-1.135	-1.884	0.0596	0.321
Sample size x rater effects	-1.073	-1.803	0.0714	0.342
Rater effects x rating design	-0.906	-1.517	0.1294	0.404
Sample size x rater effects x rating design	1.395	1.645	0.1001	4.036

Note. OR = odds ratio. Reference group are 250 (sample size), spiral rating design, some rater effects.

Research Question 2: *What is the Type I error rate and power of the item-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?*

4.3 Item-Level Discrepancy Measures

Item-total correlation and odds ratio are two item-level discrepancy measures presented in this section. The item-total correlation is illustrated for each item, while the odds ratio is demonstrated for item pairs.

4.3.1 Item-Total Correlation Discrepancy Measure

4.3.1.1 Summary of Observed and Replicated Item-Total Correlation

Item-total correlation is the correlation between a particular item's score and the total score without that item. As defined in Section 3, the item score is the average rating given to an examinee by multiple raters. The total score is the sum of the item scores. Item-total correlation indicates how well items discriminate between high- and low-achieving examinees. The average observed and replicated item-total correlation estimates across simulation conditions are presented in Figure 13. The replicated item-total correlation estimates in the top row of Figure 13 are based on HRM-PCM, while the bottom row consists of item-total estimates replicated based on HRM-GPCM. As shown in Figure 13, the line graphs of the average observed and replicated item-total correlations overlapped when the generating model was the same as the analysis model (i.e., HRM-GPCM → HRM-GPCM), suggesting that the observed item-total correlations were similar to the replicated item-total correlations when the correct model was specified. It can be seen from Figure 13 that, on average, the item-total correlations for posterior predictive samples based on HRM-PCM were underestimated.

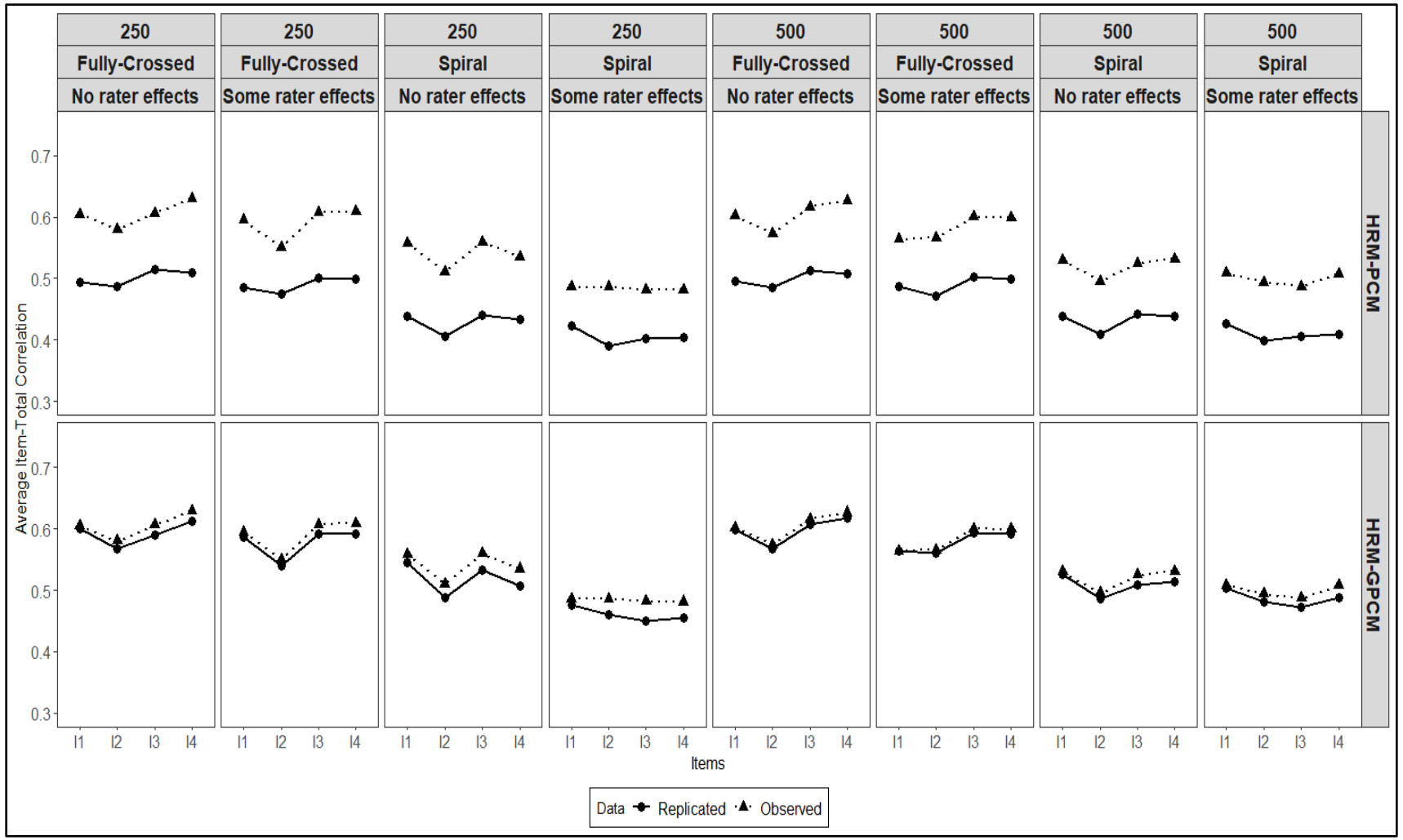

Figure 13. Average observed and replicated item-total correlation of all four items across simulation conditions

4.3.1.2 Type I Error and Power of Item-Total Correlation Discrepancy Measure

The distributions of the PPP-values are presented in Figure 14. As seen in the distributions, there were only a few extreme PPP-values when HRM-GPCM was the analysis model. This further highlight that the replicated item-total correlations under HRM-GPCM were similar to the observed item-total correlations. However, there were noticeable large extreme PPP-values when the data were analyzed with HRM-PCM. The spread of the distributions of the PPP-values when the model was misspecified were larger compared to when the correct model was specified.

GM=AM. The Type I error rates and power for the item-total correlation discrepancy measure are presented in Table 21. The Type I error rates are presented for each item. The results demonstrate that the proportions of extreme PPP-values when the correct model was specified (i.e., Type I error) ranged from 0.00 to 0.02. An item's Type I error rate of 0% suggests that, out of the 50 replications, the item was never flagged to misfit the HRM-GPCM, while the item with a Type I error rate of 2% suggests that the item was only flagged to misfit the HRM-GPCM only once. These findings suggest adequate model-data fit of the HRM-GPCM.

GM≠AM. The percentage of times data analyzed with HRM-PCM showed evidence of misfit ranged from 44% to 86%. Simulation conditions with 500 examinees outperformed conditions with 250 examinees. All simulation conditions with 500 examinees resulted in power above 70%. Compared to conditions with spiral rating design, the fully-crossed rating design yielded higher power. There were no noticeable differences in the power for conditions with and without rater effects.

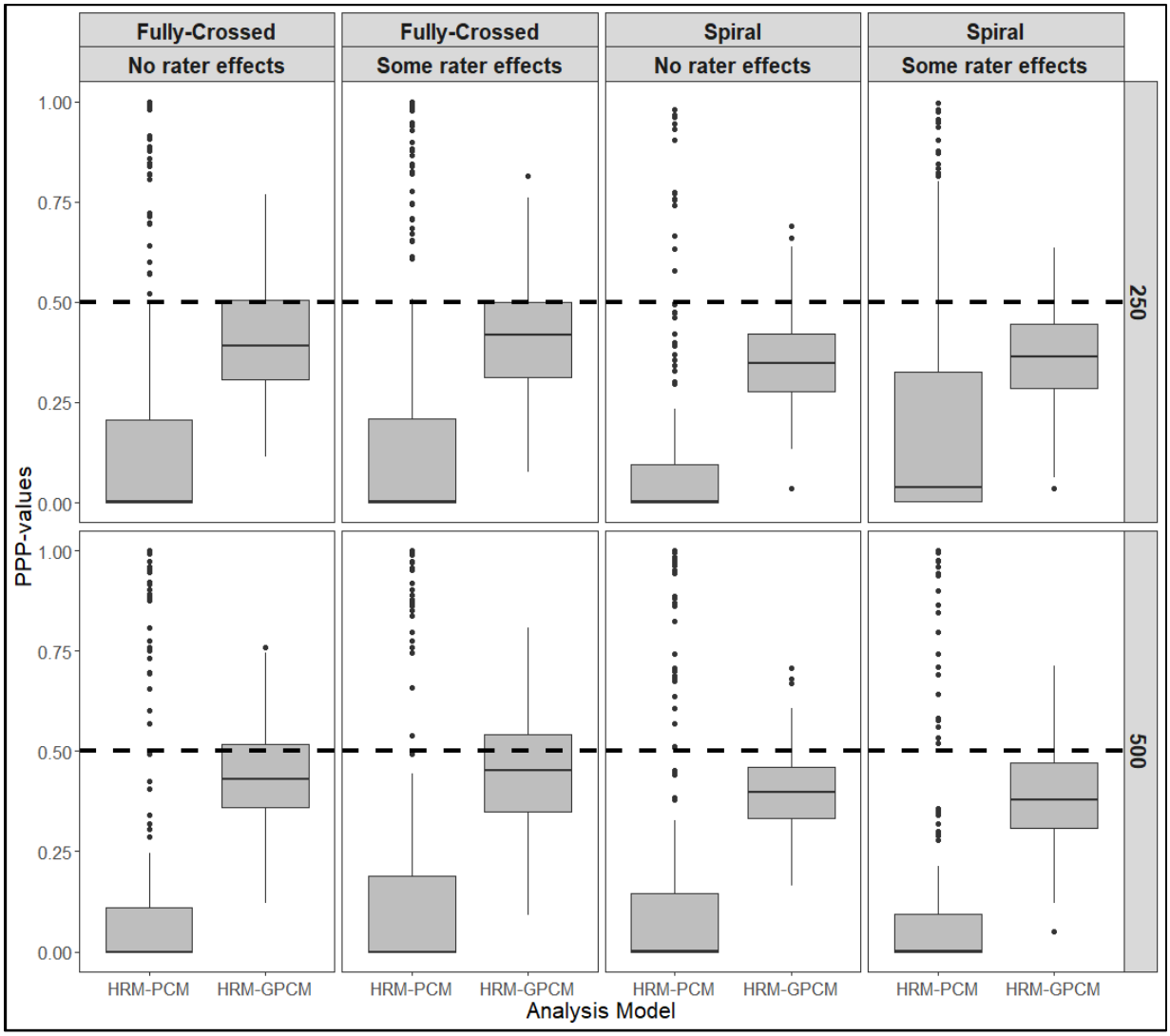

Figure 14. Boxplot of PPP-values for the item-total correlation discrepancy measure, combined for all items, across simulation conditions. The dashed horizontal line represents where the PPP-value is 0.50

Table 21. Type I Error Rates and Power for the Item-Total Correlation Discrepancy Measure

Examinee	Rater Effect	Rating Design	Item ID	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)	$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)
250	No Rater Effect	Full	1	0.00	0.64
			2	0.00	**0.72**
			3	0.00	**0.70**
			4	0.00	**0.74**
		Spiral	1	0.00	**0.72**
			2	0.02	0.66
			3	0.00	0.66
			4	0.00	0.68
	Rater Effect	Full	1	0.00	0.66
			2	0.00	0.64
			3	0.00	**0.74**
			4	0.00	**0.78**
		Spiral	1	0.00	0.56
			2	0.00	0.60
			3	0.02	0.64
			4	0.00	0.44
500	No Rater Effect	Full	1	0.00	**0.74**
			2	0.00	**0.70**
			3	0.00	**0.84**
			4	0.00	**0.82**
		Spiral	1	0.00	**0.80**
			2	0.00	**0.74**
			3	0.00	**0.72**
			4	0.00	**0.72**
	Rater Effect	Full	1	0.00	**0.86**
			2	0.00	**0.68**
			3	0.00	**0.78**
			4	0.00	**0.80**
		Spiral	1	0.00	**0.80**
			2	0.00	**0.76**
			3	0.00	**0.72**
			4	0.00	**0.78**

Note. GPCM = generalized partial credit model; PCM = partial credit model. Power rates above 70% are bolded

4.3.1.3 Power of Item-Total Correlation Discrepancy Measure by Discrimination Parameter Classification

Previous research (e.g., Li et al., 2017) suggests that the item-total correlation discrepancy measure is sensitive to the discrimination parameters of the item. In essence, Li and colleagues implied that items with discrimination parameters close to 1 were less likely to be flagged for misfit by the item-total correlation discrepancy measure. This study further explored the power of item-total correlation discrepancy measure when the discrimination parameters were considered.

This study randomly generated item discrimination parameters. That is, new discrimination parameters were generated for every new replication. Baker (2001) classified item discrimination parameters using seven categories: none, very low, low, moderate, high, very high, and perfect. The range of values for each category is presented in Table 22. Using the range of values suggested by Baker (2001), two categories of discrimination values were created: Category 1 contained items with low and high discrimination parameters (i.e., discrimination parameters less than 0.65 and greater than 1.34: $\alpha_j < 0.65$ and $\alpha_j > 1.34$) and Categories 2 contained items with moderate discrimination parameters (i.e., $0.65 \leq \alpha_j \leq 1.34$).

Table 22. Labels for Item Discrimination Parameter Values

Verbal label	Range of values
None	0
Very low	0.01 – 0.34
Low	0.35 – 0.64
Moderate	0.65 – 1.34
High	1.35 – 1.69
Very high	> 1.70
Perfect	+ infinity

Source. Baker (2001)

Figure 15 presents the power of detecting misfit of the HRM-PCM across the two categories of the discrimination parameter. As expected, items with high and low discrimination parameters were mostly flagged for misfit than items with moderate discrimination parameters. The results indicate that conditions with high or low item discrimination parameters yielded significantly high power, close to 100%. Only one simulation condition resulted in power less than 70%. Interestingly, simulation conditions with fully-crossed rating designs with 500 examinees performed the best. It can also be seen that for smaller sample size (N = 250 examinees), the fully-crossed design outperformed the spiral design.

Furthermore, the results suggest that the power of detecting misfit of the HRM-PCM when the items have moderate discrimination parameters was low. Majority of the simulation conditions resulted in power less than 40% when items have moderate discrimination parameters. These findings parallel Li et al. (2017) that suggests that the item-total correlation discrepancy measure performed poorly in detecting misfit when the item discrimination parameters were close to 1.

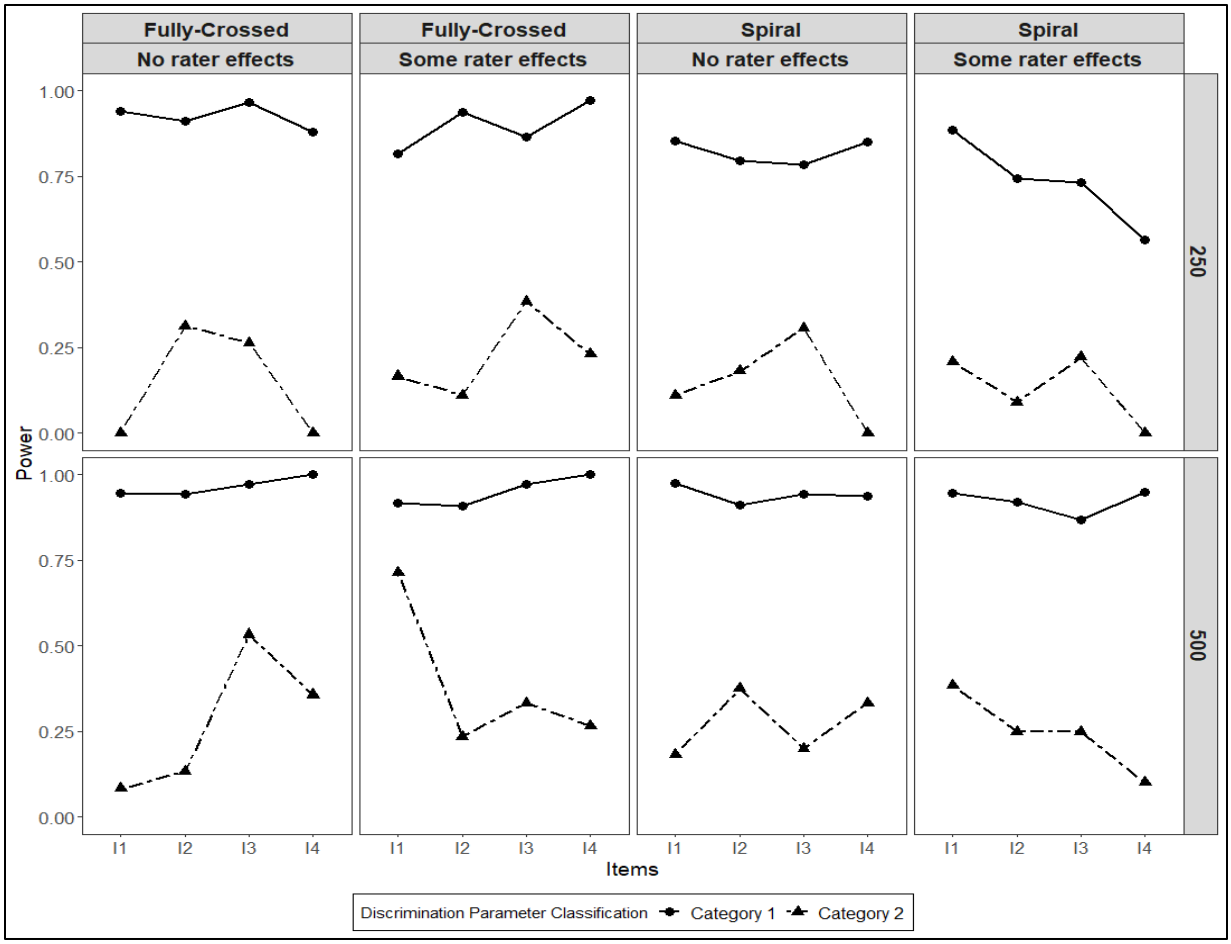

Figure 15. Power of detecting misfit for the item-total discrepancy measure by discrimination parameter classification. *Note.* Combination of levels of discrimination parameter → Category 1 (Low and High), and Category 2 (Moderate).

4.3.2 Odds Ratio Discrepancy Measure

4.3.2.1 Type I Error and Power of Odds Ratio Discrepancy Measure

GM=AM. Table 23 summarizes the Type I error rates and power of the odds ratio discrepancy measure for the different item pairs. The results demonstrate that the Type I error rates were relatively low across all conditions. However, there were a few item pairs with Type I error rates above 5%. Overall, the results suggest that the fitted model adequately captured the association among test items.

GM≠AM. When the HRM-PCM was fitted to the data, the results showed that the odds ratio performed moderately well under the fully-crossed rating design with 500 examinees. For example, the power ranged from 58% to 68% for the condition without rater effects, with a fully-crossed rating design, and 500 examinees. However, the power ranged from 44% to 58% under similar conditions but with a smaller sample size of 250 examinees. This clearly suggests that the odds ratio performed better with larger sample size and fully-crossed rating design.

Table 23. Type I Error Rates and Power for the Odds Ratio Discrepancy Measure

Rater Effect	Rating Design	Item Pair	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)		$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)	
			N=250	N=500	N=250	N=500
No Rater Effect	Full	1,2	0.04	0.00	0.58	0.64
		1,3	0.00	0.07	0.52	0.62
		1,4	0.08	0.00	0.50	0.62
		2,3	0.02	0.00	0.44	0.68
		2,4	0.00	0.00	0.46	0.58
		3,4	0.02	0.07	0.52	0.66
	Spiral	1,2	0.02	0.00	0.38	0.46
		1,3	0.00	0.00	0.38	0.50
		1,4	0.00	0.04	0.36	0.56
		2,3	0.02	0.02	0.44	0.48
		2,4	0.08	0.06	0.36	0.60
		3,4	0.04	0.00	0.30	0.34
Rater Effect	Full	1,2	0.04	0.04	0.46	0.62
		1,3	0.00	0.04	0.50	0.54
		1,4	0.02	0.00	0.52	0.56
		2,3	0.00	0.00	0.48	0.56
		2,4	0.00	0.07	0.48	0.60
		3,4	0.04	0.00	0.58	0.58
	Spiral	1,2	0.04	0.00	0.42	0.54
		1,3	0.00	0.00	0.26	0.44
		1,4	0.04	0.04	0.28	0.44
		2,3	0.02	0.00	0.34	0.44
		2,4	0.04	0.02	0.36	0.46
		3,4	0.00	0.06	0.16	0.48

Note. N = number of examinees; GPCM = generalized partial credit model; PCM = partial credit model.

4.3.2.2 Power of Odds Ratio Discrepancy Measure by Discrimination Parameter Classification

Item pairs can take on six combinations of the item discrimination parameters: (1) high/high, (2) low/low, (3) moderate/moderate, (4) high/low, (5) high/moderate, (6) moderate/low. For example, a pair of items with a high/high combination both have item discrimination parameters above 1.34, whereas a pair of items with a moderate/low combination implies that one of the items has a discrimination parameter less than 0.65 and the second item has a discrimination parameter between 0.65 and 1.34. The six combinations were further grouped into three categories: (1) Category 1 contains pair of items with high/high, low/low, and high/low combinations, (2) Category 2 contains moderate/moderate item pairs, and (3) Category 3 contains item pairs with moderate/high and moderate/low combinations.

Figure 16 illustrates the power of the item pairs for the three categories. The results suggest that item pairs that fall under Category 1 were easily detected for misfit of the HRM-PCM. Item pairs with moderate discrimination parameters (i.e., Category 2) resulted in low power, again suggesting that the odds ratio discrepancy measure may fail to capture the association between pairs of items with moderate discrimination parameters. All pairs of items in Category 3 resulted in power less than 50%. This suggests that the odds ratio may not be useful in detecting misfit of the HRM-PCM when item combinations include discrimination parameters that are moderately classified (i.e., items with discrimination parameters that are between 0.65 and 1.34).

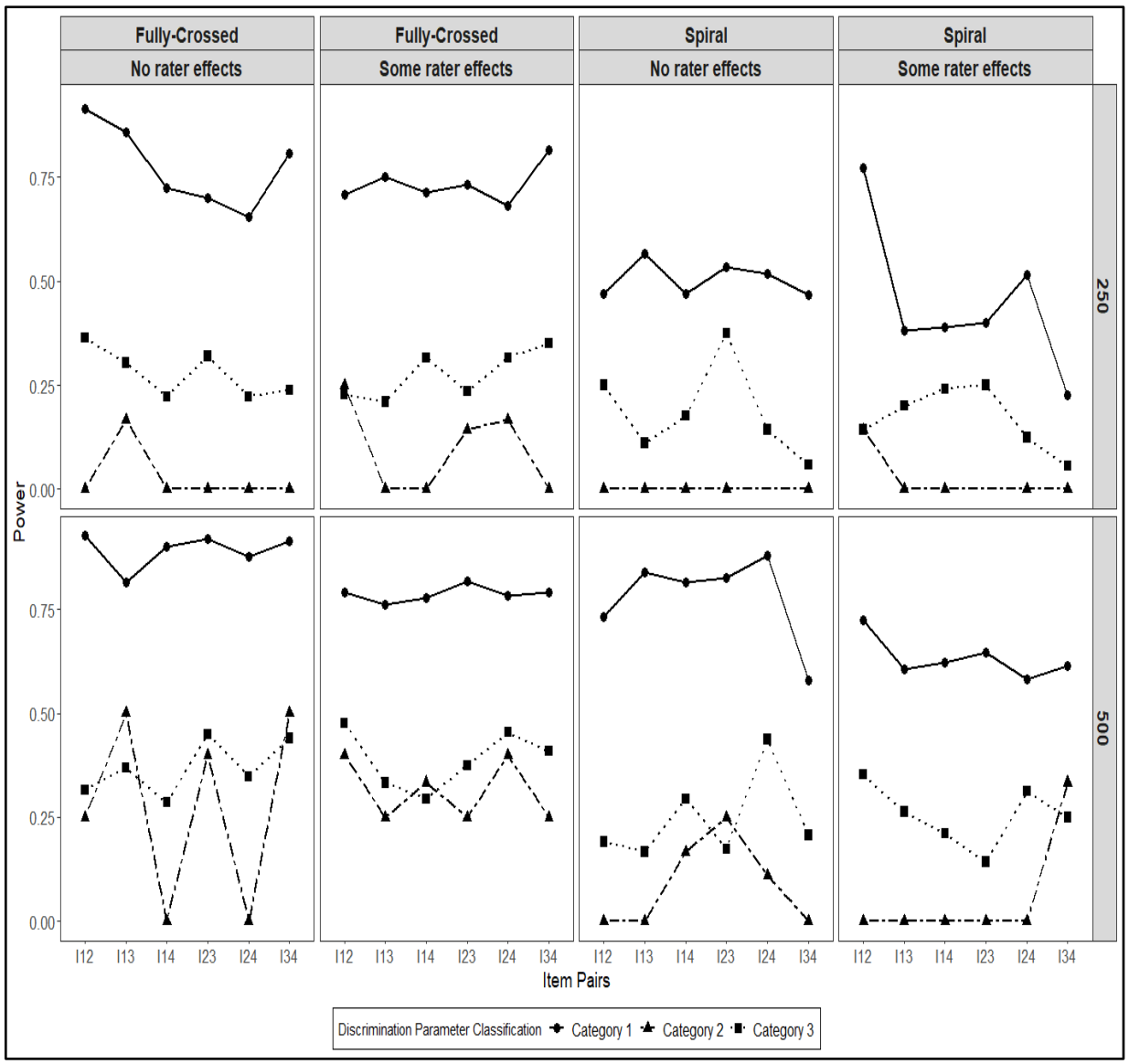

Figure 16. Power of detecting misfit for the odds ratio discrepancy measure by discrimination parameter classification. *Note.* Combination of levels of discrimination parameter → Category 1 (Low/Low, Low/High, and High/High), Category 2 (Moderate/Moderate), and Category 3 (Moderate/High, and Moderate/Low). I12 = Items 1 and 2; I13 = Items 1 and 3; I14 = Items 1 and 4; I23 = Items 2 and 3; I24 = Items 2 and 4; I34 = Items 3 and 4.

Summary of Item-Level Discrepancy Measures

The power of the item-total correlation and odds ratio discrepancy measures were evaluated at the item-level of the test. The results suggest that item-total correlation was most powerful in detecting the misfit of the HRM-PCM. There was also evidence to suggest that item discrimination parameters influenced the performance of these discrepancy measures, especially with the odds ratio. These findings parallel Li et al. (2017).

The results of the logistic regression analysis presented in Table 24 further provide evidence of the impact of the design factors in detecting inadequacy of the HRM-PCM. For the item-total correlation, the findings show that the level of the item discrimination parameter is significantly associated with detecting misfit of the HRM-PCM ($\beta_{Discrim} = 3.541, p < 0.0001, OR = 34.498$). This suggests that, holding all other conditions constant, the odds of detecting misfit of the HRM-PCM using item-total correlation for items with high or low discrimination parameters (Category 1) is 34.498 times when compared to items with moderate discrimination parameters (Category 2). The evidence also shows that sample size ($\beta_{Sample\ Size} = 1.405, p < 0.0001, OR = 4.075$), rating design ($\beta_{Rating\ Design} = 1.127, p < 0.0001, OR = 3.086$), and rater effects ($\beta_{Rater\ Effects} = 0.537, p = 0.0412, OR = 1.711$) were all significantly associated with detecting misfit of the HRM-PCM. For example, the odds ratio associated with rating design suggests that the fully-crossed rating design had an odds in detecting misfit of the HRM-PCM that is 3.086 higher in comparison to the spiral rating design.

Similarly, the level of the discrimination parameter was an important factor in detecting the inadequacy of the HRM-PCM when the odds ratio discrepancy measure was employed. The odds ratios associated with the discrimination parameter category were 18.675 and 2.699. This implies that the odds of detecting misfit of the HRM-PCM for Category 1 compared to Category

2 was 18.675. Furthermore, when compared to Category 2, the odds of items that fall under Category 3 in detecting misfit of HRM-PCM was 2.699, holding all other variables constant. The rating design was the second most important factor in detecting misfit of the HRM-PCM using the odds ratio discrepancy measure ($\beta_{Rating\ Design} = 1.042, p < 0.0001, OR = 2.834$). The odds of detecting misfit of the HRM-PCM in conditions with rater effects compared to conditions without rater effects were not statistically significant ($\beta_{Rater\ Effect} = 0.148, p = 0.4370, OR = 1.159$).

Table 24. Results of Logistic Regression to Predict Misfit of HRM-PCM of the Item-Total Correlation and Odds Ratio Discrepancy Measures

Effect	Item-Total Correlation Discrepancy Measure				Odds Ratio Discrepancy Measure			
	Estimate	*z*-value	p-*value*	OR	Estimate	*z*-value	p-*value*	OR
Intercept	-2.479	-10.839	<0.0001	0.084	-2.978	-11.239	<0.0001	0.051
Sample size	1.405	4.831	<0.0001	4.075	0.678	3.580	0.0003	1.970
Rating design	1.127	4.011	<0.0001	3.086	1.042	5.438	<0.0001	2.834
Rater effects	0.537	2.041	0.0412	1.711	0.148	0.777	0.4370	1.159
Discrimination parameter category								
Category 1	3.541	21.570	<0.0001	34.498	2.927	12.502	<0.0001	18.675
Category 3	-	-	-	-	0.993	4.228	<0.0001	2.699
Sample size x rating design	-0.587	-1.399	0.1618	0.556	-0.185	-0.692	0.4890	0.832
Sample size x rater effects	-0.317	-0.774	0.4391	0.729	0.246	0.923	0.3562	1.278
Rater effects x rating design	-0.489	-1.224	0.2208	0.613	-0.072	-0.271	0.7862	0.930
Sample size x rater effects x rating design	0.120	0.201	0.841	1.127	-0.111	-0.296	0.7674	0.895

Note. OR = odds ratio. Reference group are 250 (sample size), spiral rating design, and some rater effects. For the discrimination parameter category, Category 2 is the reference category for both the item-total correlation and odds ratio discrepancy measures. Item-total correlation has two discrimination parameter categories, while odds ratio has three discrimination parameter categories.

Research Question 3: What is the Type I error rate and power of the rater-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?

4.4 Rater-Level Discrepancy Measures

This section summarizes the findings of three rater-level discrepancy measures: score-estimate correlation, rater-total correlation, and rater standard deviation. First, the summary of the observed and replicated samples for each discrepancy measure is presented. Next, the Type I error and power are discussed. The impact of rater and item characteristics on the discrepancy measures are further evaluated.

4.4.1 Score-Estimate Correlation Discrepancy Measure

4.4.1.1 Summary of Observed and Replicated Score-Estimate Correlation

Literature suggests that score-estimate correlation is a measure of rater accuracy in the use of the rating scale (Wolfe, 2014; Wolfe & McVay, 2010). The observed and replicated score-estimate correlations when the model was correctly and incorrectly specified are presented in Figure 17. First, it can be seen that Rater 2 has the highest average score-estimate correlation. This is unsurprising because Rater 2 had the smallest bias and variability parameters ($\phi_2 = 0.046, \psi_2 = 0.270$). In conditions with rater effects, Rater 6 has the smallest average score-estimate correlation. Rater 6's variability parameter was the largest among all raters ($\psi_6 = 1.487$). In HRM, a rater's variability parameter is a measure of the rater's consistency (i.e., accuracy) in the use of the rating scale. The patterns in Figure 17 suggest that the score-estimate correlations of the most accurate raters was larger than the score-estimate correlation of less accurate raters.

Figure 17 further shows that, when the correct model was specified, the score-estimate correlations from the replicated datasets based on the posterior predictive samples resembled the

observed score-estimate correlation. However, it can be seen that, on average, the observed score-estimate correlations were higher than the score-estimate correlations from the replicated datasets under a misspecified model. This suggests that the HRM-PCM underestimated the score-estimate correlation when the item discrimination parameters are not 1.

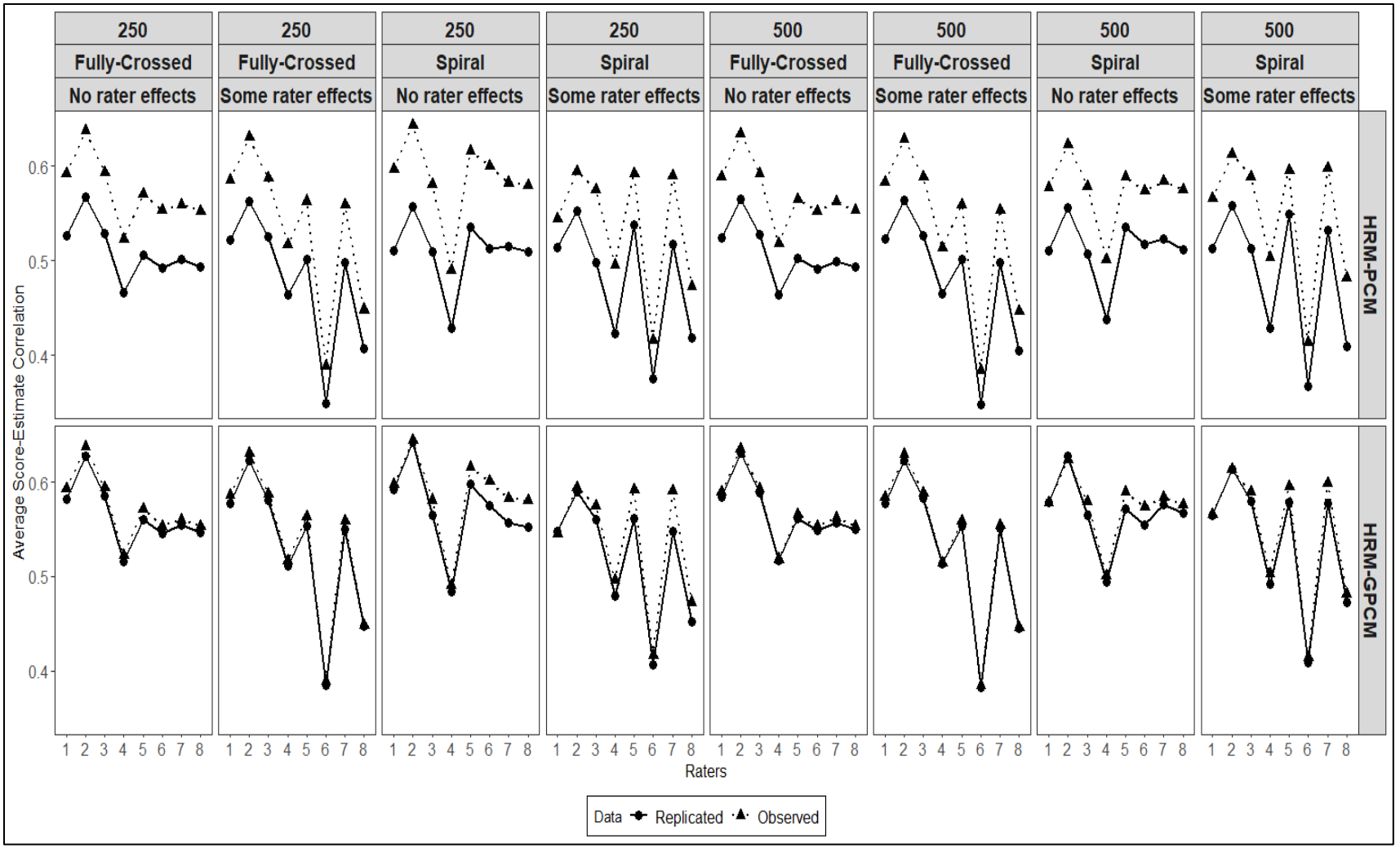

Figure 17. Average observed and replicated score-estimate correlation across simulation conditions

4.4.1.2 Type I Error Rates and Power of Score-Estimate Correlation Discrepancy Measures

GM=AM. The Type I error rates and power of the score-estimate correlation discrepancy measure resulting from the correct specification and misspecification of the HRM are presented in Table 25. The results are presented for each rater since raters have different bias and variability parameters. In general, the Type I error rates were less than 5%. There were only two raters with Type I error rates greater than 5%: (1) Rater 1 with Type I error rate of 10% when N=250, without rater effects, and fully-crossed rating design, and (2) Rater 7 with Type I error rate of 8% when N=500, without rater effects, and fully-crossed rating design. Overall, the findings suggest that, when the correct model was specified, the score-estimate correlation adequately captured the relationship between the average scores awarded by a single rater and examinees' true ability.

GM≠AM. Using latent measurement models (i.e., GPCM, PCM, RSM), Wolfe (2014) demonstrated that the discrimination parameter was a measure of rater accuracy. Extending this to the HRM, the results of this study showed that the rater score-estimate correlation was effective in detecting misfit of the HRM-PCM when the discrimination parameters were constrained to 1, with power ranging from 30% to 86%. The results vary by simulation conditions. Conditions with 500 examinees outperformed conditions with 250 examinees. The average power for 500 examinees was 74% compared to 62% when N=250. Similarly, conditions with fully-crossed rating design performed better than conditions with spiral design.

The effects of raters exhibiting rater effects on score-estimate correlation were further examined. For conditions with rater effects, Rater 6 and Rater 8 had the lowest power, on average. These two raters were simulated to exhibit large variability and significant bias parameters. For example, when N=250 under a fully-rating design, Rater 6 and 8 had power of

42% and 54%, respectively, compared to power between 70% and 80% for the rest of the raters in the same condition. We can infer that detecting misfit of the HRM-PCM using score-estimate correlation is adequate, but the performance is lower when raters have large variability parameters.

Table 25. Type I Error Rates and Power for the Score-Estimate Correlation Discrepancy Measure

Rater Effect	Rating Design	Rater	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)		$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)	
			N=250	N=500	N=250	N=500
No Rater Effect	Full	1	0.10	0.02	**0.78**	**0.82**
		2	0.00	0.00	**0.80**	**0.86**
		3	0.02	0.02	**0.70**	**0.84**
		4	0.02	0.02	0.66	**0.80**
		5	0.02	0.00	**0.74**	**0.84**
		6	0.02	0.00	**0.72**	**0.82**
		7	0.02	0.08	**0.70**	**0.86**
		8	0.02	0.00	**0.74**	**0.84**
	Spiral	1	0.04	0.00	**0.74**	**0.76**
		2	0.00	0.00	**0.74**	**0.84**
		3	0.00	0.02	0.66	**0.74**
		4	0.00	0.00	0.44	0.64
		5	0.02	0.02	0.58	**0.82**
		6	0.00	0.00	0.56	**0.72**
		7	0.02	0.00	0.44	**0.66**
		8	0.04	0.02	0.46	**0.70**
Rater Effect	Full	1	0.04	0.02	**0.78**	**0.82**
		2	0.02	0.00	**0.80**	**0.86**
		3	0.02	0.00	**0.74**	**0.82**
		4	0.04	0.00	**0.74**	**0.74**
		5	0.00	0.00	**0.80**	**0.78**
		6*	0.04	0.02	0.42	0.52
		7	0.04	0.02	**0.70**	**0.80**
		8*	0.00	0.00	0.54	**0.70**
	Spiral	1	0.02	0.02	0.62	**0.80**
		2	0.00	0.04	0.56	**0.78**
		3	0.00	0.00	0.60	**0.64**
		4	0.02	0.04	0.58	**0.64**
		5	0.00	0.02	0.48	0.50
		6*	0.00	0.00	0.32	0.44
		7	0.00	0.00	0.50	**0.62**
		8*	0.00	0.00	0.30	**0.64**

Note. N = number of examinees; GPCM = generalized partial credit model; PCM = partial credit model. Power rates above 70% are bolded. *Raters simulated to exhibit rater effects

4.4.1.3 Power of Score-Estimate Correlation Discrepancy Measures by Discrimination Parameter Classification

In performance assessments, examinees typically take all items on the test, which implies that the characteristics of the items on the test, including rater characteristics, may indicate how well examinees' ability parameters are estimated. For example, an item with a discrimination parameter near zero could be problematic even if even raters are experienced and well-trained. The performance of the score-estimate correlation when the item discrimination parameters were considered was further evaluated. Three discrimination categories were created based on the discrimination parameters. Category 1 contained replications with at most two items with high discrimination parameters (i.e., $\alpha_j > 1.34$). Category 2 contained replications with exactly three items with high discrimination parameters. Category 3 contained replications that all four items have high discrimination parameters.

The effect of the item and rater characteristics on the power of score-estimate correlation is illustrated in Figure 18. The results suggest that score-estimate correlation is powerful in detecting misfit of the HRM-PCM when all items have high discrimination parameters (i.e., Category 3). The performance of Category 3 was highest under fully-crossed designs with 500 examinees. For example, the power was 100% for every rater in the condition with fully-crossed and 500 examinees regardless of rater effects indicating that constraining the slopes of all the items to 1 clearly have an impact on the fit of the raters. Category 1 with only two highly discriminating items performed the least. This suggests that score-estimate correlation was less effective in detecting misfit when only half of the items have high discrimination parameters.

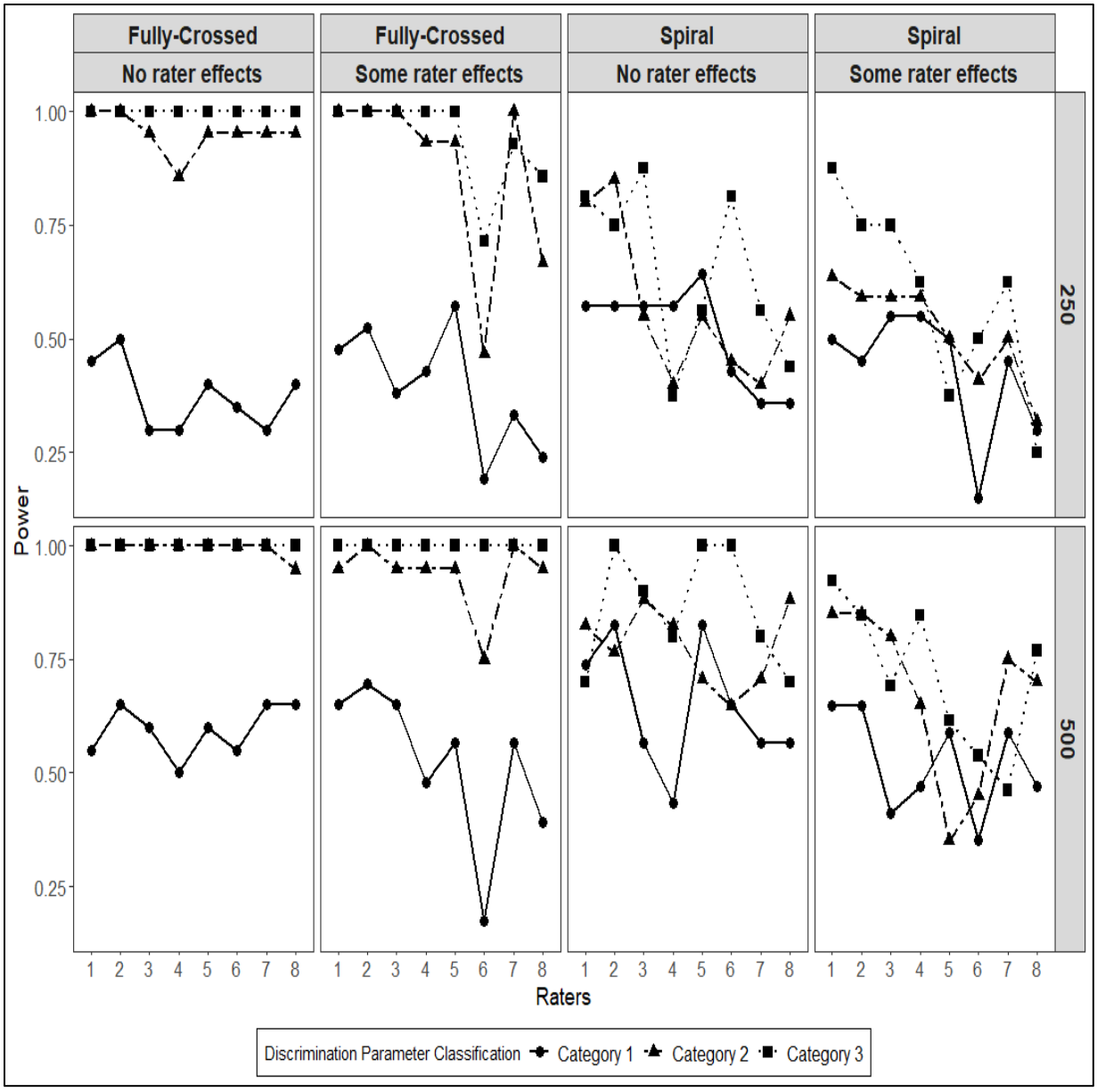

Figure 18. Power of detecting misfit for the score-estimate correlation discrepancy measure by discrimination parameter classification. *Note.* Combination of levels of discrimination parameter → Category 1 (at most two highly discriminating items), Category 2 (exactly three highly discriminating items), Category 3 (four highly discriminating items).

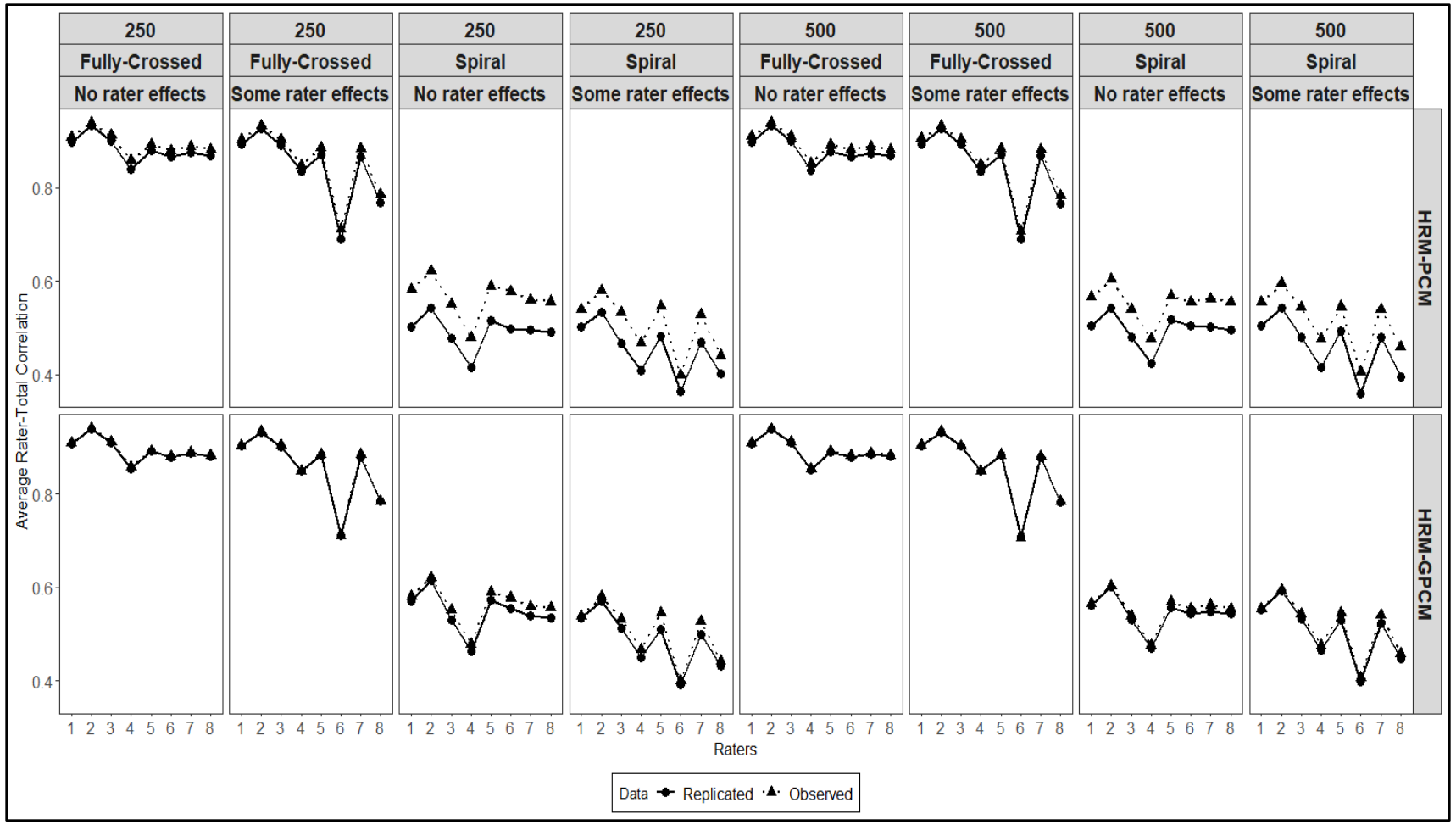

Figure 19. Average observed and replicated rater-total correlation across simulation conditions

4.4.2 Rater-Total Correlation Discrepancy Measure

4.4.2.1 Summary of Observed and Replicated Rater-Total Correlation Coefficient Discrepancy Measure

The average observed and replicated rater-total correlation coefficients are presented in Figure 19. As with the score-estimate correlation, Rater 2, with the smallest variability parameter, had the largest average rater-total correlation. However, for the conditions with rater effects, Rater 6, with the largest variability parameter, had the smallest average rater-total correlation. This figure also shows that the observed and replicated rater-total correlation coefficients were similar when the data were fitted with the correct model. However, misspecification of the model led to noticeable differences in the average observed and replicated rater-total correlation, especially under the spiral rating design.

4.4.2.2 Type I Error and Power of Rater-Total Correlation Discrepancy Measure

GM=AM. The proportions of extreme PPP-values when the correct model was specified is summarized in Table 26. Overall, the Type I error rates did not exceed the 5% in all simulation conditions. There were only two conditions with raters that had Type I error rates above 5%. The Type I error rates for the spiral design were mostly 0%. Across all conditions, the Type I error rates for the spiral design ranged from 0% to 2% compared to 0% to 6% under the full-crossed design. For example, all eight raters, including raters with rater effects, had Type I error rates of 0% for the condition with 500 examinees under the spiral design. These findings suggest that it is more beneficial to have total scores (i.e., the sum of the item scores) from fewer raters (e.g., two raters) than from all eight raters.

GM≠AM. Table 26 also presents the proportions of extreme PPP-values when the model was misspecified. Across all simulation conditions, the power ranged from 10% to 72%. It is

clear from Table 26 that power increased as the number of examinees increased from 250 to 500. For example, for the condition with fully-crossed rating design with no rater effects, the power rates were between 16% and 26% for a sample size of 250 examinees compared power rates between 26% and 58% when a sample size of 500 examinees was used. It is worth mentioning that percentage of times the HRM-PCM was flagged for misfit under the spiral design was larger than the fully-crossed rating design. For example, for the condition with rater effects and 500 examinees, the power ranged from 14% to 36% under the fully-crossed rating design compared to a superior 38% to 68% under the spiral rating design. Again, the superiority of the spiral rating design with rater-total correlation suggests that total scores from fewer raters is preferred.

Raters 6 and 8 simulated to be less accurate (i.e., large variability parameters) have the lowest power rates. For example, in the condition with 500 examinees, spiral design, and with rater effects, Rater 6 and Rater 8 had power rates of 38% and 46%, respectively, compared to power between 50% and 68% for the rest of the raters. This suggests that the performance of rater-total correlation was weaker for raters with rater effects, implying that the ability to detect misfit is reduced when a dataset includes raters with rater effects.

Table 26. Type I Error Rates and Power for the Rater-Total Discrepancy Measure

Rater Effect	Rating Design	Rater	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)		$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)	
			N=250	N=500	N=250	N=500
No Rater Effect	Full	1	0.00	0.02	0.18	0.58
		2	0.00	0.00	0.16	0.34
		3	0.02	0.02	0.26	0.44
		4	0.02	0.00	0.26	0.26
		5	0.02	0.02	0.24	0.48
		6	0.04	0.04	0.16	0.44
		7	0.04	0.00	0.28	0.34
		8	0.06	0.04	0.26	0.40
	Spiral	1	0.02	0.00	0.64	**0.74**
		2	0.00	0.00	0.64	**0.70**
		3	0.00	0.00	0.50	0.56
		4	0.00	0.00	0.30	0.48
		5	0.00	0.00	0.48	0.52
		6	0.00	0.00	0.50	0.52
		7	0.00	0.00	0.30	0.58
		8	0.00	0.00	0.38	0.58
Rater Effect	Full	1	0.02	0.04	0.22	0.36
		2	0.00	0.02	0.14	0.30
		3	0.06	0.00	0.22	0.30
		4	0.04	0.04	0.22	0.30
		5	0.02	0.04	0.18	0.20
		6*	0.02	0.00	0.12	0.14
		7	0.02	0.02	0.30	0.34
		8*	0.00	0.02	0.14	0.22
	Spiral	1	0.00	0.00	0.36	0.64
		2	0.00	0.00	0.42	0.68
		3	0.00	0.00	0.36	0.52
		4	0.00	0.00	0.32	0.50
		5	0.00	0.00	0.32	0.54
		6*	0.00	0.00	0.12	0.38
		7	0.00	0.00	0.28	0.54
		8*	0.00	0.00	0.10	0.46

Note. N = number of examinees; GPCM = generalized partial credit model; PCM = partial credit model. Power rates above 70% are bolded. *Raters simulated to exhibit rater effects

4.4.2.3 Power of Rater-Total Correlation Discrepancy Measures by Discrimination Parameter Classification

Figure 20 illustrates the effect of the rater and item characteristics on the proportions of extreme PPP-values. The results indicate that the performance of the rater-total correlation in detecting misfit of the HRM-PCM increased with increase in the number of highly discriminating items on the test. For the condition with 500 examinees, fully-crossed design, and with rater effects, the power for Category 3 ranged from 55% to 100% compared to 20% to 50% for Category 2 and 0% to 20% for Category 1. Clearly, this suggests that the rater-total correlation benefits from having test items that have high discrimination parameters.

For replications with highly discriminating items, the performance of rater-total correlation increased with increase in sample size. In addition, it can be observed that when the number of examinees was 500, the performance of Category 3 for the fully-crossed rating design was similar to the spiral rating design. However, increased sample size benefitted Category 1 and Category 2 items under the spiral rating design than the fully-crossed design. It is also worth noting that conditions without rater effects performed in a similar fashion as conditions with rater effects.

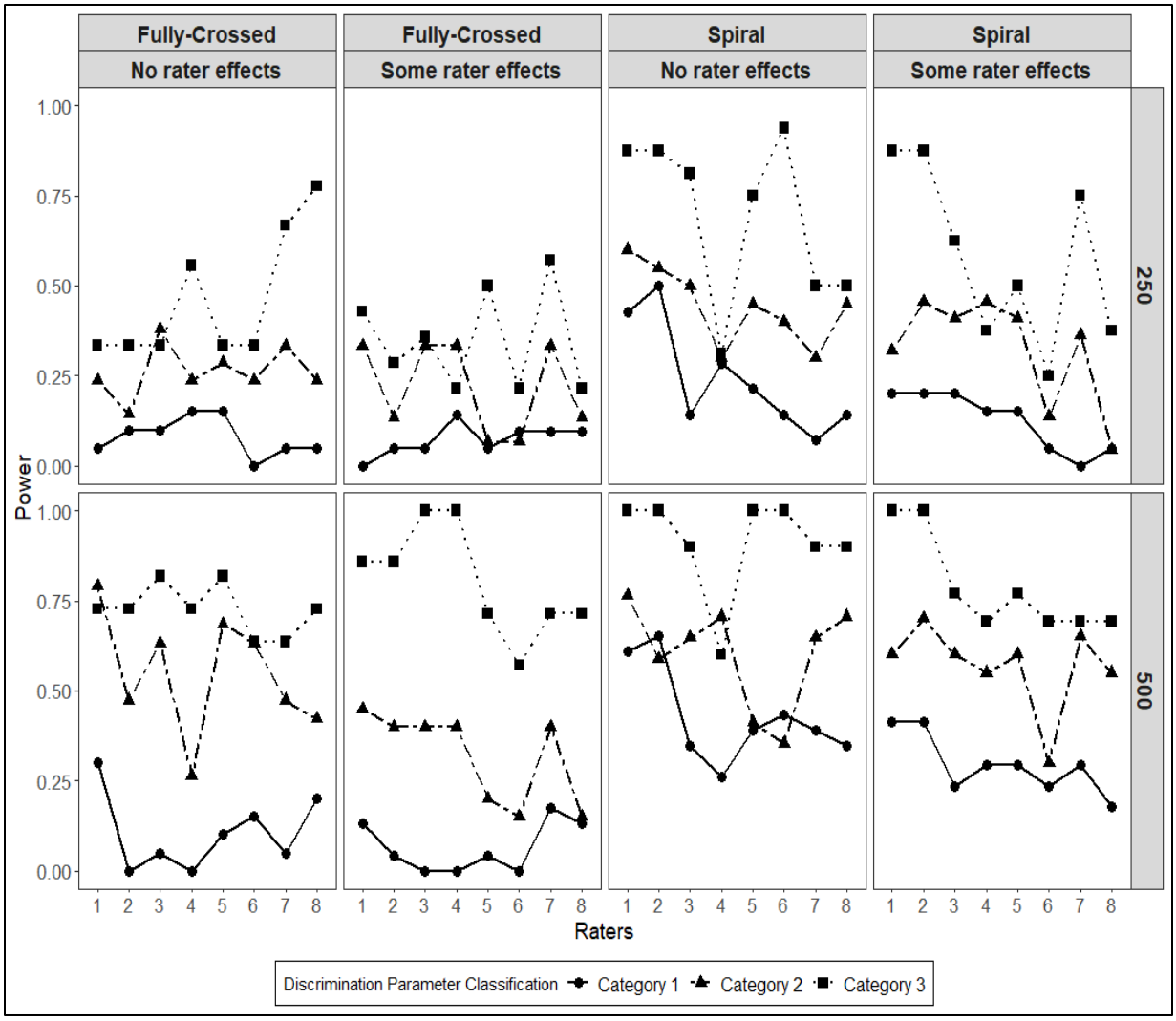

Figure 20. Power of detecting misfit for the rater-total correlation discrepancy measure by discrimination parameter classification. *Note.* Combination of levels of discrimination parameter → Category 1 (at most two highly discriminating items), Category 2 (exactly three highly discriminating items), Category 3 (four highly discriminating items).

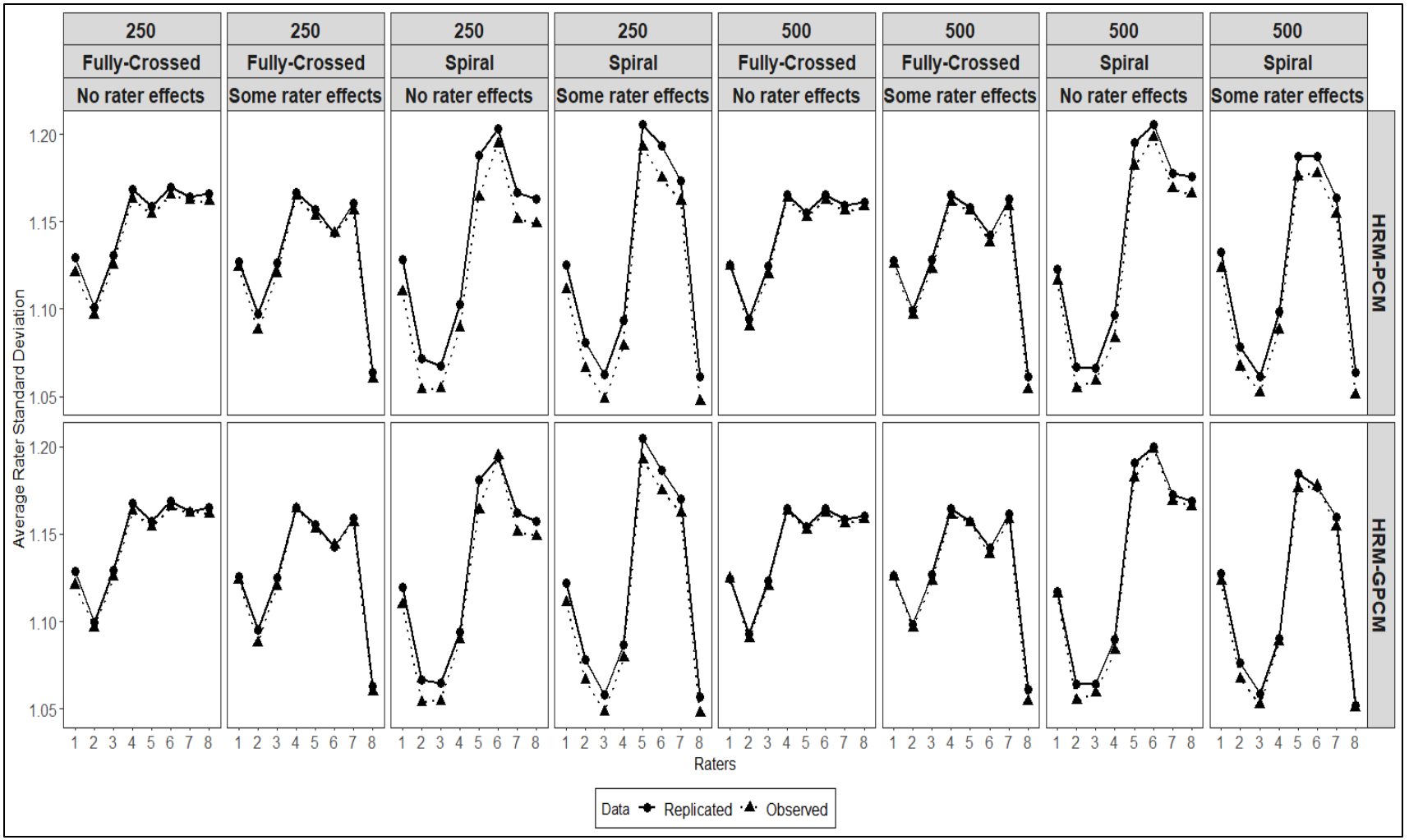

Figure 21. Average observed and replicated rater standard deviation across simulation conditions

4.4.3 Rater Standard Deviation Discrepancy Measure

4.4.3.1 Summary of Observed and Replicated Rater Standard Deviation

The average observed and replicated standard deviation of the raters are presented in Figure 21. As seen from this figure, on average, rater standard deviations were between 1.05 and 1.20. Rater 2 had the smallest average standard deviation among raters without rater effects. In conditions without rater effects, Rater 2 has the smallest bias and variability parameters. Hence, this rater can be said to consistently assign scores that reflect examinees' true abilities. However, as shown in Figure 21, Rater 8 had the smallest average standard deviation in conditions with rater effects. One explanation of this is that Rater 8 has the most extreme bias parameter ($\phi_r = -1.055$) suggesting that Rater 8 award scores around the lower end of the score distribution. As earlier depicted in Figure 6, Rater 8 most frequently used Score 1 and Score 2 of the score distribution.

When the observed and replicated rater standard deviations were compared, the patterns illustrated in Figure 21 reveal that, on average, the rater standard deviation based on the posterior predictive samples of the HRM-GPCM resemble the observed rater standard deviation. Comparably, the rater standard deviation based on the posterior predictive samples of the HRM-PCM were equally similar to the observed rater standard deviation, on average.

4.4.3.2 Type I Error and Power for the Rater Standard Deviation Discrepancy Measure

GM=AM. The Type I error rates and power for the rater standard deviation discrepancy measure are presented in Table 27. The Type I error rates are presented for each rater. The results demonstrate the Type I error rates ranged from 0% to 10%. For example, a rater's Type I error rate of 0% implies that of the 50 replications in a set of conditions, this rater's observed standard deviation was similar to the rater's replicated standard deviation based on the posterior

predictive samples of the HRM-GPCM in all the replications. These findings suggest adequate model-data fit of the HRM-GPCM.

GM≠AM. The performance of the rater standard deviation discrepancy measure when the model was misspecified (i.e., HRM-GPCM → HRM-PCM) is also presented in Table 27. The power ranged from 0% to 10%. The low power observed across the simulation conditions for this discrepancy measure indicates that the rater standard deviation is not useful in detecting misfit of the HRM-PCM.

Table 27. Type I Error Rates and Power for Rater Standard Deviation Discrepancy Measure

Rater Effect	Rating Design	Rater	$HRM_{GPCM} \rightarrow HRM_{GPCM}$ (Type I error)		$HRM_{GPCM} \rightarrow HRM_{PCM}$ (Power)	
			N=250	N=500	N=250	N=500
No Rater Effect	Full	1	0.06	0.02	0.06	0.02
		2	0.00	0.00	0.00	0.00
		3	0.00	0.02	0.00	0.00
		4	0.08	0.00	0.08	0.00
		5	0.00	0.02	0.00	0.02
		6	0.02	0.00	0.02	0.00
		7	0.02	0.02	0.02	0.00
		8	0.00	0.00	0.00	0.00
	Spiral	1	0.04	0.02	0.06	0.00
		2	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.02
		6	0.00	0.00	0.02	0.02
		7	0.00	0.00	0.00	0.00
		8	0.00	0.00	0.00	0.00
Rater Effect	Full	1	0.04	0.02	0.00	0.00
		2	0.00	0.00	0.00	0.00
		3	0.06	0.00	0.08	0.00
		4	0.00	0.04	0.00	0.04
		5	0.00	0.04	0.00	0.02
		6	0.06	0.02	0.06	0.02
		7	0.10	0.02	0.10	0.04
		8	0.00	0.02	0.00	0.02
	Spiral	1	0.00	0.00	0.02	0.02
		2	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.04
		7	0.00	0.00	0.00	0.00
		8	0.00	0.00	0.02	0.00

Note. N = number of examinees; GPCM = generalized partial credit model; PCM = partial credit model.

Summary of Rater-Level Discrepancy Measures

The results presented at the rater-level show that the score-estimate correlation was powerful in detecting misfit of the inadequacy of the HRM-PCM. The score-estimate correlation yielded high power across all simulation conditions. The results of the logistic regression analyses in Table 28 indicate that sample size, rating design, and discrimination parameter category were all significant predictors of misfit of the HRM-PCM. Although the rater effects factor was not significant, the interaction between sample size and rater effects was significant in detecting misfit of the HRM-PCM. It can be observed from Table 28 that the odds ratios associated with the categories of the discrimination parameters were largest. When compared to Category 1, the odds of detecting misfit of the HRM-PCM when the test has four highly discriminating items (Category 3) was 6.155 times higher. Similarly, the odds of detecting misfit of the HRM-PCM was 4.289 higher for Category 2 in comparison to Category 1.

The rater-total correlation also performed moderately well in detecting the misfit of HRM-PCM. It was evident, as reported in Section 4.4.2.3, that the power of detecting misfit of the inadequacy of the HRM-PCM increased when the test contained highly discriminating items. This evidence is further shown in the logistic regression analyses reported in Table 28. As seen in this table, the odds of detecting misfit of the HRM-PCM was 11.222 higher for Category 3 compared to Category 1. This proves that the rater-total correlation performance is highest for highly discriminating test items. The findings also show that for the rater-total correlation discrepancy measure, the spiral rating design performed better than the fully-crossed rating design ($\beta_{Rating\ Design} = -4.192, p < 0.0001, OR = 0.464$). This study also found that the conditions with larger sample size and the conditions without rater effect yielded higher odds of detecting misfit of the HRM-PCM.

Table 28. Results of Logistic Regression to Predict Misfit of HRM-PCM of the Score-Estimate Correlation and Rater-Total Correlation Discrepancy Measures

Effect	Score-Estimate Correlation Discrepancy Measure				Rater-Total Correlation Discrepancy Measure			
	Estimate	*z*-value	p-*value*	OR	Estimate	*z*-value	p-*value*	OR
Intercept	-0.967	-8.026	<0.0001	0.380	-2.037	-14.598	<0.0001	0.130
Sample size	0.517	3.338	0.0008	1.677	1.040	6.393	<0.0001	2.829
Rating design	0.945	5.929	<0.0001	2.572	-0.767	-4.192	<0.0001	0.464
Rater effects	0.131	0.855	0.3923	1.140	0.567	3.492	0.0005	1.763
Discrimination parameter category								
Category 2	1.456	15.344	<0.0001	4.289	1.314	13.083	<0.0001	3.720
Category 3	1.817	14.881	<0.0001	6.155	2.418	20.411	<0.0001	11.222
Sample size x rating design	-0.051	-0.221	0.8250	0.950	-0.255	-1.037	0.2998	0.775
Sample size x rater effects	0.632	2.794	0.0052	1.882	-0.064	-0.281	0.7789	0.938
Rater effects x rating design	0.083	0.364	0.7156	1.086	-0.197	-0.789	0.4303	0.822
Sample size x rater effects x rating design	-0.413	-1.221	0.2222	0.662	0.249	0.736	0.4620	1.283

Note. OR = odds ratio. Reference group are 250 (sample size), spiral rating design, and some rater effects. For the discrimination parameter category, Category 1 is the reference category for both discrepancy measures

CHAPTER 5:

DISCUSSION

In educational assessments, traditional polytomous IRT models are employed to analyze data where an examinee can obtain partial or full credits. Examinees and item parameters, which are typically of interest in traditional IRT models, can be estimated with the PCM and GPCM. Assessments that require human raters introduces additional considerations such as rater effects. Consequently, measurement models that compensate for these effects are needed in performance assessment. The HRM is one of the IRT models that accounts for rater severity/leniency and accuracy. The formulation of the HRM allows for the estimation of rater bias and consistency in the use of the rating scale. The second stage of the HRM allows for the use of a polytomous IRT model to describe the relationship between an examinee's ideal ratings and latent ability. The PCM and GPCM can be used in the first stage of the HRM. This study documents the consequences of using the wrong polytomous IRT model in the HRM.

Researchers and practitioners risk making invalid inferences if a posited model does not fit the data. Using PPMC techniques, this study investigated the performance of different discrepancy measures in detecting misfit of the HRM. This study varied the sample size (N = 250 and 500), rating design (fully-crossed and spiral rating designs), rater effects (no rater effects and 25% of raters with rater effects), and analysis model (HRM-GPCM and HRM-PCM), leading to 16 fully-crossed simulation conditions. One thousand posterior predictive samples were used for the PPMC. Furthermore, the PPP-values were computed as the proportions of times for which the observed discrepancy measure exceeds the discrepancy measure based on the posterior predictive samples. Models with extreme PPP-values were flagged for misfit. Using guidelines from previous studies (e.g., Sinharay, 2006; Sinharay & Johnson, 2003), extreme PPP-values

were defined as less than 0.05 or greater than 0.95. The performance of the discrepancy measures employed was assessed at the test-, item-, and rater-level. This study addressed three research questions.

Research Question 1: *What is the Type I error rate and power of the test-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?*

At the test level, the results reveal that the proportions of extreme PPP-values were zero when the generating model was the same as the analysis model (i.e., Type I error rates) for the mean discrepancy measure, across all the simulation conditions, suggesting that the HRM-GPCM adequately captures the center of the total score distribution. In addition, the proportions of extreme PPP-values, for the mean discrepancy measure, when the generating model was different from the analysis model (i.e., power) were zero in all the simulation conditions. This implies that the mean discrepancy measure did not flag the HRM-PCM for misfit in all 50 simulation replications across the simulation conditions. In PPMC, the discrepancy measure used is crucial to the inferences made regarding model fit (Levy & Mislevy, 2016). For example, the mean of the total score leads us to believe that the HRM-PCM adequately fits the data even though that HRM-PCM was the incorrect model. Although the center of the total score distribution was captured by the HRM-PCM, other features of the data may not be adequately captured by HRM-PCM. Therefore, using only the mean of the total score distribution will lead to incorrect inferences about the fitted model if other features of the data are of interest.

The standard deviation of the total score performed moderately well in detecting misfit of the HRM-PCM in certain conditions. Overall, the HRM-PCM underestimated the standard deviation of the total score distribution. The Type I error rates were 0% across all simulation conditions, suggesting that the HRM-GPCM adequately captured the spread of the total score of

the data. The performance of the standard deviation discrepancy measure in detecting misfit of the HRM-PCM varied across simulation conditions. The percent of times the HRM-PCM was flagged for misfit (i.e., power) was between 24% and 76%. Sample size and rating design played the most significant role in detecting misfit of the HRM-PCM. Larger sample size (N = 500) and fully-crossed rating design were associated with higher odds of detecting misfit of the HRM-PCM. In addition, it was found that detecting misfit was higher in conditions without rater effects compared to conditions with rater effects.

The first and third quartiles were used to assess how well the lower and upper ends of the distribution were adequately captured by the models. The Type I error rates suggest that the HRM-GPCM adequately captured the lower and upper 25% of the total score distribution for tests with fully-crossed rating design. The Type I error rates under the spiral rating design were relatively large (between 10% and 44%). The Type I error rates were noticeably larger for tests with 500 examinees compared to tests with 250 examinees. The power for the first quartile ranged from 0% to 38% and from 24% to 80% for the third quartile. The high Type I error rates and low power suggest that the quartiles may not be useful measures to assess model-data fit of the HRM. Under the fully-crossed design, these measures suggest that the HRM-GPCM was an adequate fit to the data; however, the fully-crossed design is not commonly used in practice (Eckes, 2011). Also, the low power observed for the fully-crossed design further implies that these measures are not useful in detecting misfit of the HRM-PCM.

It was worthwhile to investigate how well the extreme tails of the total score distribution were captured by the fitted models. The results indicate that the Type I error rates were as high as 22% for the minimum discrepancy measure and 16% for the maximum discrepancy measure. Surprisingly, fully-crossed design conditions with rater effects had the lowest Type I error rates,

ranging from 2% to 6%. The power rates were low. The maximum power was 14% for the minimum discrepancy measure and 26% for the maximum discrepancy measure suggesting that these measures were not useful in detecting the inadequacy of the HRM-PCM. In practice, it may be uncommon for researchers to be interested in the extremes of the distributions. However, these discrepancy measures indicate that extreme values in the total score distribution are not adequately captured by the model.

Distributional shapes were assessed using the skewness and kurtosis discrepancy measures. For both discrepancy measures, the proportions of extreme values when the correct model was used to analyze the data were zero across all simulation conditions. This suggests that the HRM-GPCM adequately captures the shape of the data. Model misspecification with HRM-PCM yielded power that ranged from 0 to 2% for the skewness discrepancy measure and from 4% to 26% for the kurtosis discrepancy measure. These low power rates suggest that skewness and kurtosis discrepancy measures should not be used to assess misfit of the HRM-PCM.

Overall, it can be seen that shape of the total score distribution under the HRM-PCM and HRM-GPCM tends to be similar, as suggested by the skewness statistics. However, the spread of the distribution was narrower when data generated with HRM-GPCM were fitted with HRM-PCM leading to detecting misfit using the standard deviation discrepancy measure. The standard deviation discrepancy measure outperformed all the discrepancy measures employed at the test level. This measure is recommended for detecting deviations in the observed data and replicated data at the total score distribution level.

Research Question 2: *What is the Type I error rate and power of the item-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?*

The key distinction between the HRM-GPCM and HRM-PCM is that the item

discrimination parameter is fixed to 1 under the HRM-PCM but allowed to vary under the HRM-GPCM. The item-total correlation and odds ratio were employed to detect the misfit of the HRM-PCM at the item level. Allen and Yen (1979) described item-total correlation as a discrimination index indicating how well items distinguish between high- and low-performing examinees. The results of the present study suggest that the true item-total correlations were adequately captured when HRM-GPCM was fitted to the data but underestimated when HRM-PCM was the analysis model. The Type I error rates ranged from 0% to 2% across all simulation conditions indicating that the HRM-GPCM adequately captures how well items discriminate between high- and low-achieving examinees. The item-total correlation also yielded moderate to high power especially with conditions with 500 examinees. The power in conditions with 500 examinees ranged from 68% to 86%. This suggests that the item-total correlation is powerful with larger (N = 500) sample size conditions in detecting misfit when the item discrimination parameters were constrained to 1. These findings parallel previous studies that found item-total correlation to be useful in detecting misfit of Rasch models (e.g., Li et al., 2017; Sinharay & Johnson, 2003). The discrimination parameters for this study were generated from a lognormal distribution. New data generating parameters were drawn for each new simulation replication. There was also evidence that items with discrimination parameters close to 1 were less likely to be detected for misfit of the HRM-PCM compared to items with discrimination parameters farther from 1.

Using global odds ratio, Li et al. (2017) found that when the GPCM was the data generating model and analysis model, the Type I error rates were 0% across all their simulations conditions. In the present study, the Type I error rates for the odds ratio ranged from 0 to 8%. A considerably large number of this study's simulation conditions resulted in Type I error rates less

than 5% (see Table 23). It is important to note that the test length and number of examinees in Li et al. (2017) are considerably larger than the test length and number of examinees used in this study. Furthermore, Li and colleagues found that the power of the odds ratio detecting misfit of PCM ranged from 25% to 54%. Similar rates were found in the current study. The power of detecting misfit of the HRM-PCM ranged from 16% to 68%. The present study further explored the impact of the item discrimination parameters in detecting misfit of HRM-PCM using odds ratio. There was significant evidence to conclude that the odds of detecting misfit of HCM-PCM was almost 19 times when the item pairs consisted of items with discrimination parameters farther from 1 compared to item pairs with discrimination parameters close to 1.

Research Question 3: *What is the Type I error rate and power of the rater-level discrepancy measures in detecting model-data misfit of the HRM using PPMC?*

Wolfe (2014) and Wolfe and McVay (2012) extensively discussed score-estimate correlation and rater-total correlation as indexes for detecting the accuracy of ratings awarded by raters. The score-estimate correlation is analogous to the point-measure correlation. Using latent measurement models such as GPCM, Wolfe (2014) indicated that the score-estimate correlation and rater-total correlation increase with rater accuracy and decrease with rater inaccuracy. Literature that utilized score-estimate correlation or rater-total correlation to diagnose misfit of the functional form of performance assessment models is scarce. The present study extends these two statistics to evaluate misfit of the functional form of the HRM.

The score-estimate correlation was computed as the correlation between a rater's average ratings the latent ability of the examinees. Therefore, employing the wrong model would have an impact on accurately estimating the true abilities of the examinees. Feuerstahler (2018) found that the trait estimates were biased when data generated from a more complex model was fit to a

less complex model. In the present study, it was found that HRM-PCM underestimated the score-estimate correlations. The Type I error rates were less than 5% in all but one simulation condition. The power of detecting misfit of the HRM-PCM ranged from 30% to 86%. The results of this study further demonstrate that the odds of detecting misfit for the fully-crossed design was more than twice higher compared to the spiral rating design. Most significantly, the odds of detecting misfit when the test contained more highly discriminating items were significantly higher than when there were fewer items with high discrimination parameters.

The rater-total correlation performed adequately well in detecting misfit of the HRM-PCM. The proportion of times the HRM-GPCM was flagged for misfit was mostly less than 0.05, demonstrating that the HRM-GPCM adequately captured how well raters' scores and the total scores are correlated. When the model was misspecified, the results indicate that rater-total correlation performed better in conditions with spiral rating design. This would imply that fewer ratings in the total score are preferred when using rater-total correlation. Each examinee only received two ratings on each item under the spiral rating design compared to eight ratings per item under the fully-crossed rating design. More ratings could lead to more discrepancies in the scores, especially if there are raters exhibiting rater effects. In fact, the results showed that the odds of detecting misfit were almost twice higher in conditions without rater effects compared to conditions with rater effects. The rater-total correlation performed well in detecting misfit when the test contained highly discriminating items compared to tests with less discriminating items.

Implications for Researchers and Practitioners

One of the benefits of employing Rasch models such as the PCM is that an examinee's raw score is a sufficient statistic for θ. In Rasch models, examinees are considered to have equivalent latent ability traits if the examinees have the same raw scores on a set of items. Rasch

models assume that all the items are equally related to the latent ability trait. This assumption may not be attainable in practice (Embretson & Reise, 2000). The GPCM allows the items to have different slopes. Studies in traditional IRT have shown that model misspecification may lead to biased item and examinee latent ability parameters (e.g., Feuerstahler, 2018).

The present study found overwhelming evidence of the role of the magnitude of the discrimination parameter in detecting misfit of the HRM-PCM. The power rates were considerably higher in conditions with slope parameters farther from 1. An implication of this finding is that parameters of all items should be carefully assessed when using the discrepancy measures employed in this study. Sinharay and Haberman (2014) discussed the practical significance of misfit of IRT models. One of their recommendations is omitting misfitting items when misfit is practically significant. The present study showed that the score-estimate and rater-total correlations performed well in detecting misfit of the HRM-PCM when the test contains exactly three (out of four) highly discriminating items. However, when only two out of the four items are highly discriminating, the results revealed that the score-estimate and rater-total correlations underperformed in detecting misfit of the HRM-PCM. This implies that tests with only 50% misfitting items may not be flagged for misfit using the rater-level discrepancy measures. Thus, it is recommended to assess the item-level misfits alongside rater- and test-level misfits. Assessing item-level misfits would give the researchers and practitioners more information to make decisions about whether it is necessary to omit the misfitting items. Most importantly, the impact of omitting misfitting items on the proficiency levels of the examinees should be examined.

Rater training, calibration, and monitoring are some of the measures taken to improve rating quality in performance assessments (Johnson et al., 2008). McClellan (2010) discussed

several approaches for controlling rating quality including backscoring, validity papers, double scoring, and trend scoring. Research has extensively discussed the impacts of rater effects in performance assessments (e.g., Myford & Wolfe, 2003; Wind, 2019; Wolfe, 2014; Wolfe & McVay, 2012). The current study revealed that that the discrepancy measures useful in detecting misfits of the HRM-PCM slightly underperformed in conditions with rater effects. These findings further stress the importance of rater training and monitoring to ensure that raters have a shared understanding of the rubrics, thus potentially decreasing the possibilities of having severe/lenient and inconsistent raters.

Rating design is an integral component of performance assessment. This study used the fully-crossed and spiral rating designs. The fully-crossed rating design has been employed in practice (e.g., Ezike & Ames, 2021), however, this design may not be realistic in large operational testing with large examinee-rater-item combinations due to the time constraints and cost implications. Overall, the results of the current study showed that the fully-crossed rating design outperformed the spiral rating design in nearly all conditions. The spiral design performed relatively well in conditions with large sample size (N = 500). Noteworthy is that the spiral rating design performed better than the fully-crossed rating design for the rater-total discrepancy measure. Researchers and practitioners that use incomplete rating designs should consider using rater-total correlation alongside the score-estimate correlation in evaluating model-data fit.

Limitations and Future Research

There are a few limitations in the current study. First, the current study only used 50 replications due to computing time. In addition, categorizing the discrimination parameter imply that the number of replications within each category was less than 50. Future studies should consider using larger number of replications.

Second, the HRM has been extended to account for longitudinal designs (i.e., L-HRM) and multidimensional tests (i.e., M-HRM). The current study did not consider these newer extensions of the HRM. The current study provides a foundation to further explore the L-HRM and M-HRM, and other models used for calibrating data from performance assessments.

Third, the distribution of latent ability traits was fixed in this study. The distribution of latent ability traits is not limited to a normal distribution. Previous IRT studies have employed distributions such as uniform and skewed distributions. Conforti and Casabianca (2016) investigated parameter recovery of the HRM with nonnormal data. They found that rater parameters were robust to nonnormality in the latent ability traits. It would be interesting to investigate the performance of the discrepancy measures employed in the present study in the presence of nonnormal latent ability distributions.

Forth, two rating designs were used in the current study. There are other incomplete rating designs that should be considered in future research. For example, raters can be randomly assigned which papers they score. This type of rating design is commonly used in large testing programs. Future studies should consider employing this type of rating design.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51.

Agresti, A. (2002). Categorical Data Analysis.,(Wiley: Hoboken, NJ.).

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory, 1973* (pp. 268-281). Publishing House of the Hungarian Academy of Sciences.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Ames, A. J. (2015). *Bayesian model criticism: Prior sensitivity of the posterior predictive checks Method* (Doctoral dissertation).

Ames, A. J. (2018). Prior Sensitivity of the Posterior Predictive Checks Method for Item Response Theory Models. *Measurement: Interdisciplinary Research and Perspectives*, *16*(4), 239-255.

Ames, A.J., Ezike, N.C., & Myers, A.J. (n.d.). Model-data fit evaluation: Posterior checks and Bayesian model selection.

Ames, A. J., Leventhal, B. C., & Ezike, N. C. (2020). Monte Carlo simulation in item response theory applications using SAS. *Measurement: Interdisciplinary Research and Perspectives*, *18*(2), 55-74.

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for Item Response Theory models. *Educational Measurement: Issues and Practice*, *34*(3), 39-48.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Bayarri, M. J., & Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, *95*(452), 1127-1142.

Birnbaum, A. (1958). On the estimation of mental ability. *Series Rep*, *15*, 7755-7723.

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

Bock, R. D. (1960). Methods and applications of optimal scaling. *The University of North Carolina Psychometric Laboratory Research Memorandum*, *25*.

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

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied psychological measurement*, *26*(4), 364-375.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, *35*(2), 179-197.

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Cham, Switzerland: Springer.

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.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Conforti, P., & Casabianca, J. (2016, April). The Hierarchical Rater Model with Nonnormal Populations. In *Annual Meeting of the National Council on Measurement in Education, Washington, D.C.*

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

DeCarlo, L. T. (2008). Studies of a latent class signal detection model for constructed response Scoring. *ETS Research Report Series*, *2008*(2), i-55.

DeCarlo, L. T. (2010). Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs. *ETS Research Report Series*, *2010*(1), i-65.

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.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt am Main, Peter Lang GmbH.

Eckes, T. (2020). Rater-Mediated Listening Assessment: A Facets Modeling Approach to the Analysis of Raters' Severity and Accuracy When Scoring Responses to Short-Answer Questions. *Psychological Test and Assessment Modeling*, *65*(4), 449-471

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

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 Jr, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*(1), 19-33.

Ezike, N. C., & Ames, A. J. (2021). The Diagnostic Rating System: Rater Behavior for an Alternative Performance Assessment Rating Method. *Psychological Test and Assessment Modeling*, *63*(3), 273-304.

Feuerstahler, L. M. (2018). Sources of error in IRT trait estimation. *Applied psychological measurement*, *42*(5), 359-375.

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education*, *32*(2), 77-96.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, *33*(3), 291-314.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian Data Analysis Chapman & Hall. *CRC Texts in Statistical Science*.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733-760.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457-472.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of markov chain monte carlo*, *6*, 163-174.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1995). *Markov chain Monte Carlo in practice*. CRC press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

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

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press.

Hombo, C., & Donoghue, J. R. (2001, April). Applying the hierarchical raters model to NAEP. In *Annual Meeting of the National Council on Measurement in Education, Seattle Washington*.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). A simulation study of the effect of rater designs on ability estimation. *ETS Research Report Series*, *2001*(1), i-41.

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.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*(2), 121-138.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331-358.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391-406.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*(5), 399-419.

Kim, Y. (2009). *Combining constructed response items and multiple choice items using a*

hierarchical rater model. Columbia University.

Lane, S. (2010). *Performance assessment: The state of the art.* (SCOPE Student Performance Assessment Series). Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Lane, S., & Stone, C. A. (2006). Performance assessments. In B. Brennan, (Ed.), *Educational measurement* (4th ed., pp. 387-432).

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks* (Doctoral dissertation).

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519-537.

Li, T., Xie, C., & Jiao, H. (2017). Assessing fit of alternative unidimensional polytomous IRT models using posterior predictive model checking. *Psychological methods*, *22*(2), 397.

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

Linacre, J. M. (2003). The hierarchical rater model from a Rasch perspective. *Rasch Measurement Transactions (Transactions of the Rasch Measurement SIG American Educational Research Association)*, *17*(2), 928.

Linacre, J. M. (2020). A User's Guide to FACETS Rasch-Model Computer Programs Program Manual. Retrieved from https://www.winsteps.com/manuals.htm

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*.

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments* (Doctoral dissertation).

Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics*, *32*(3), 287-314.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

McClellan, C. A. (2010). Constructed-response scoring—Doing it right. *R&D Connections*, *13*, 1-7.

McDonald, R. P. (1994). Testing for approximate dimensionality. In: Levault, D.Zumbo, BD.Gessaroli, ME., Boss, MW., *Modern theories of measurement: Problems and issues*. Ottawa, Canada: University of Ottowa; 1994. p. 63-86.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics.

Applied Psychological Measurement, 9(1), 49-57.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*(3), 1142-1160.

Mislevy, J. L., Rupp, A. A., & Harring, J. R. (2012). Detecting local item dependence in polytomous adaptive data. *Journal of Educational Measurement, 49*(2), 127-147.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i-30.

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

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227

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.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied psychological measurement, 24*(1), 50-64.

Park, C. (1997). *Accuracy of parameter estimation on polytomous IRT models*. Doctoral dissertation, University of Massachusetts Amherst.

Patz, R. J. (1996). *Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress*. Doctoral dissertation, Carnegie Mellon University

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.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice, 33*(1), 36-48.

Perron, B. E., & Gillespie, D. F. (2015). *Key concepts in measurement*. Oxford University Press, USA.

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, No. 125.10, pp. 1-10).

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Ascd.

Priestley, M. (1982). *Performance assessment in education and training: Alternative techniques*. Educational Technology.

Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 176.

R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. URL https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of educational Measurement*, *27*(2), 133-144.

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.

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

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin*, *88*(2), 413.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*(2), 461-464.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). sage.

Sheng, Y. (2017). Fitting psychometric models: issues and new developments. *Frontiers in psychology*, *8*, 856.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British journal of mathematical and statistical psychology*, *59*(2), 429-449.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant?. *Educational Measurement: Issues and Practice*, *33*(1), 23-35.

Sinharay, S., & Johnson, M. S. (2003). Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models. *ETS Research Report Series*, *2003*(2), i-55.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298-321.

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. van der, Linde A., 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*(4), 583-639.

Song, Y. A. (2019). *A comparative study of IRT models for rater effects and double scoring* (Doctoral dissertation, The University of Iowa).

Stone, C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement*, *60*(6), 974-991.

Stone, C.A. & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. Cary, NC: SAS Publications.

Su, Y. S., & Yajima, M. (2012). Package 'R2jags'. A Package for Running jags from R.

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-theory and Methods*, *7*(1), 13-26.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, *55*(2), 371-390.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics*, *13*(2), 117-130.

Uto, M., & Ueno, M. (2020). A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, *47*(2), 469-496.

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In *Essays on item response theory* (pp. 89-108). Springer, New York, NY.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103-118.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational measurement*, *37*(2), 141-162.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language testing*, *15*(2), 263-287.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*(3), 335-353.

Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147-170.

Wiggins, G. (1990). The case for authentic assessment. *Practical assessment, research, and evaluation*, *2*(1), 2.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve*

Student Performance. Jossey-Bass Publishers, 350 Sansome Street, San Francisco, CA 94104.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. *Objective measurement: Theory into practice, 5*, 113-134.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283-306.

Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, *43*(2), 159-171.

Wind, S. A., & Stager, C. G. (2019). The impacts of characteristics of disconnected subsets on group anchoring in incomplete rater-mediated assessment networks. *Psychological Test and Assessment Modeling*, *61*(1), 13-36.

Wind, S. A., Ooi, P. S., & Engelhard Jr, G. (2019). Exploring decision consistency and decision accuracy across rating designs in rater-mediated music performance assessments. *Musicae Scientiae*, *23*(4), 465-485.

Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes* (White Paper). Iowa City, IA: Pearson Education.

Wolfe, E. W., & McVay, A. (2010). Rater effects as a function of rater training context. *Retrieved from Pearson at http://images.pearsonassessments.com/images/tmrs/tmrs_rg/RaterEffects_101510. pdf*.

Woods, C. M. (2014). Estimating the latent density in unidimensional IRT to permit non-normality. In *Handbook of item response theory modeling* (pp. 78-102). Routledge.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological measurement*, *29*(1), 23-48.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245-262.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, *30*(3), 187-213.

Yen, W.M., & Fitzpatrick, A. R. (2006). Item response theory. In B. Brennan, (Ed.), *Educational measurement* (4th ed., pp. 111-153).

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement*, *48*(1), 81-97.

Zhu, X., & Stone, C. A. (2012). Bayesian comparison of alternative graded response models for performance assessment applications. *Educational and Psychological Measurement*, *72*(5), 774-799.

Appendices

Appendix 1: Item Generating Parameters for PPMC Illustration

Item	α	β
1	1.000	0.882
2	1.000	-0.107
3	1.000	1.316
4	1.000	1.497
5	1.000	1.060

Appendix 2: Syntax for Data Generation for Fully-Crossed Rating Design

```
#################################################################
# Simulate Data Using HRM for Fully-Crossed Design
#################################################################

simulate_data_GPCM <- function(theta, a, b, phi, psi )
{
 RR <- ncol(phi)  ##Raters
 I <- nrow(b)     ##Items
 ####################
 ## Ideal Ratings
 ####################
 K <- ncol(b)      #Location parameters
 N <- length(theta)  ##test length
 KM <- outer(rep(1,N), 0:K)+1

 simulate_response <- function(alpha, theta, beta) {
  unsummed <- c(alpha*theta, alpha*(theta - beta))
  numerators <- exp(cumsum(unsummed))
  denominator <- sum(numerators)
  response_probs <- numerators/denominator
  simulated_y <- sample(1:length(response_probs), size = 1,
                        prob = response_probs)
  return(simulated_y)
 }

 ii <- rep(1:I, times = N)
 jj <- rep(1:N, each = I)

 total=N*I

 gpcm <- numeric(total)

  for(n in 1:total) {
    gpcm[n] <- simulate_response(a[ii[n]], theta[jj[n]], b[ii[n], ])
 }

 gpcm_y <- matrix(gpcm, byrow=TRUE, ncol=I)

 ##Simulate data for all raters
 items <- paste0("I", 1:I)
 dat <- NULL
 for (rr in 1:RR){
  dat.rr <- matrix( NA, nrow=N, ncol=I)
```

```
    colnames(dat.rr) <- items
    for (ii in 1:I){
      # ii <- 1
      probs <- exp( - ( KM - ( gpcm_y[,ii] + phi[ii,rr] ) )^2 / psi[ii,rr] / 2 )
      probs <- probs / rowSums(probs )
      probs <- sirt::rowCumsums.sirt(probs)
      vals <- sirt::rowIntervalIndex.sirt(matr=probs, rn=stats::runif(N))
      dat.rr[,ii] <- vals
    }
    dat <- rbind( dat, dat.rr )
  }

  dat1 <- data.frame( "pid"=rep(1:N, RR), "rater"=rep(1:RR, each=N), dat )
  dat2 <- dat1[order( dat1$pid ), ]
  rownames(dat2) <- NULL
  return(dat2)

}
```

Appendix 3: Syntax for Data Generation for Spiral Rating Design

```
#################################################################
# Simulate Data Using HRM for Spiral-Rating Design
#################################################################

simulate_spiral_data_GPCM <- function(theta, a, b, phi, psi ){
  RR <- ncol(phi)  ##Raters
  I <- nrow(b)     ##Items
  ####################
  ## Ideal Ratings
  ####################
  K <- ncol(b)      #Location parameters
  N <- length(theta)  ##test length
  KM <- outer(rep(1,N), 0:K)+1

  simulate_response <- function(alpha, theta, beta) {
    unsummed <- c(alpha*theta, alpha*(theta - beta))
    numerators <- exp(cumsum(unsummed))
    denominator <- sum(numerators)
    response_probs <- numerators/denominator
    simulated_y <- sample(1:length(response_probs), size = 1,
                          prob = response_probs)
    return(simulated_y)
  }

  ii <- rep(1:I, times = N)
  jj <- rep(1:N, each = I)

  total=N*I

  gpcm <- numeric(total)

  for(n in 1:total) {
    gpcm[n] <- simulate_response(a[ii[n]], theta[jj[n]], b[ii[n], ])
  }

  gpcm_y <- matrix(gpcm, byrow=TRUE, ncol=I)


  ###Simulate data for all raters
  Total_Rater = seq(1,RR)


  RR1 <- split(Total_Rater, sort(1:length(Total_Rater) %% 4))[["0"]]
```

```
  RR2 <- split(Total_Rater, sort(1:length(Total_Rater) %% 4))[["1"]]
  RR3 <- split(Total_Rater, sort(1:length(Total_Rater) %% 4))[["2"]]
  RR4 <- split(Total_Rater, sort(1:length(Total_Rater) %% 4))[["3"]]

  Total_Item = seq(1,I)
  I1 = split(Total_Item, sort(1:length(Total_Item) %% 4))[["0"]]
  I2 = split(Total_Item, sort(1:length(Total_Item) %% 4))[["1"]]
  I3 = split(Total_Item, sort(1:length(Total_Item) %% 4))[["2"]]
  I4 = split(Total_Item, sort(1:length(Total_Item) %% 4))[["3"]]

    ##Spiral One
  items <- paste0("I", 1:I)

  dat1 <- NULL
  for (rr in min(RR1):max(RR1)){
    dat.rr <- matrix( NA, nrow=N, ncol=I)
    colnames(dat.rr) <- items
    for (ii in min(I1):max(I1)){
      # ii <- 1
      probs <- exp( - ( KM - ( gpcm_y[,ii] + phi[ii,rr] ) )^2 / psi[ii,rr] / 2 )
      probs <- probs / rowSums(probs )
      probs <- sirt::rowCumsums.sirt(probs)
      vals <- sirt::rowIntervalIndex.sirt(matr=probs, rn=stats::runif(N))
      dat.rr[,ii] <- vals
    }
    dat1 <- rbind(dat1, dat.rr )
  }

  datS1 <- data.frame( "pid"=rep(1:N, length(RR1)), "rater"=rep(min(RR1):max(RR1), each=N),
dat1 )
  datS1 <- datS1[order( datS1$pid ), ]

  ##Spiral Two
  dat2 <- NULL
  for (rr in min(RR2):max(RR2)){
    dat.rr <- matrix( NA, nrow=N, ncol=I)
    colnames(dat.rr) <- items
    for (ii in min(I2):max(I2)){
      # ii <- 1
      probs <- exp( - ( KM - ( gpcm_y[,ii] + phi[ii,rr] ) )^2 / psi[ii,rr] / 2 )
      probs <- probs / rowSums(probs )
      probs <- sirt::rowCumsums.sirt(probs)
      vals <- sirt::rowIntervalIndex.sirt(matr=probs, rn=stats::runif(N))
      dat.rr[,ii] <- vals
    }
    dat2 <- rbind( dat2, dat.rr )
```

```
  }

  datS2 <- data.frame( "pid"=rep(1:N, length(RR2)), "rater"=rep(min(RR2):max(RR2), each=N),
dat2 )
  datS2 <- datS2[order( datS2$pid ), ]

  ##Spiral Three
  dat3 <- NULL
  for (rr in min(RR3):max(RR3)){
    dat.rr <- matrix( NA, nrow=N, ncol=I)
    colnames(dat.rr) <- items
    for (ii in min(I3):max(I3)){
      # ii <- 1
      probs <- exp( - ( KM - ( gpcm_y[,ii] + phi[ii,rr] ) )^2 / psi[ii,rr] / 2 )
      probs <- probs / rowSums(probs )
      probs <- sirt::rowCumsums.sirt(probs)
      vals <- sirt::rowIntervalIndex.sirt(matr=probs, rn=stats::runif(N))
      dat.rr[,ii] <- vals
    }
    dat3 <- rbind( dat3, dat.rr )
  }

  datS3 <- data.frame( "pid"=rep(1:N, length(RR3)), "rater"=rep(min(RR3):max(RR3), each=N),
dat3 )
  datS3 <- datS3[order( datS3$pid ), ]

  ##Spiral Four
  dat4 <- NULL
  for (rr in min(RR4):max(RR4)){
    dat.rr <- matrix( NA, nrow=N, ncol=I)
    colnames(dat.rr) <- items
    for (ii in min(I4):max(I4)){
      # ii <- 1
      probs <- exp( - ( KM - ( gpcm_y[,ii] + phi[ii,rr] ) )^2 / psi[ii,rr] / 2 )
      probs <- probs / rowSums(probs )
      probs <- sirt::rowCumsums.sirt(probs)
      vals <- sirt::rowIntervalIndex.sirt(matr=probs, rn=stats::runif(N))
      dat.rr[,ii] <- vals
    }
    dat4 <- rbind( dat4, dat.rr )
  }

  datS4 <- data.frame( "pid"=rep(1:N, length(RR4)), "rater"=rep(min(RR4):max(RR4), each=N),
dat4)
  datS4 <- datS4[order( datS4$pid ), ]
```

```
  #Combine
  spiral_data <- rbind(datS1, datS2, datS3, datS4)
  rownames(spiral_data) <- NULL
  return(spiral_data)
}
```