Studies of Rater and Item Effects in Rater Models

Yihan Zhao

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Studies of Rater and Item Effects in Rater Models
Yihan Zhao

The goal underlying educational testing is to measure psychological constructs in a particular domain and to produce valid inferences about examinees' ability. To achieve this goal of getting a precise ability evaluation, test developers construct questions with different formats, such as multiple-choice (MC) items, and open-ended questions or constructed response (CR) test items, for example, essay items. In recent years, large-scale assessments have implemented CR items in addition to MC items as an essential component of the educational assessment landscape.

However, utilizing CR items in testing involves two main challenges, including rater effects and rater correlations. One challenge is the error added by human raters' subjective judgments, such as rater severity and rater central tendency. Rater severity effect refers to the effect that raters may tend to give consistently low or high ratings that cause biased ability evaluation (Leckie & Baird, 2011). Central tendency describes when raters tend to use middle categories in the scoring rubric and avoid using extreme criteria (Saal et al., 1980). The second challenge is that multiple raters usually grade an examinee's essay for quality control purposes; however, ratings based on the same item are correlated and need to be handled carefully by appropriate statistical procedures (Eckes, 2011; Kim, 2009).

To solve these problems, DeCarlo (2010) proposed an HRM-SDT model that extended the traditional signal detection theory (SDT) model used in the first level of HRM. The HRM-SDT model not only considers the hierarchical structure of rating data but also deals with various rater effects beyond rater severity. This research examined to what extent the HRM-SDT separates rater effects (i.e., rater severity and rater central tendency) from item effects (i.e., item

difficulty). Accordingly, one goal of this study was to simulate various rater effects and item effects to investigate the performance of the HRM-SDT model with respect to separating these effects. The other goal was to compare the fit of the HRM-SDT model with one commonly used model in language assessments, the Rasch model, in different simulation conditions and to examine the difference between these two models in terms of segregating rater and item effects.

To answer these questions, Simulation A and Simulation B were conducted. In Simulation A, seven sets of parameters were varied in the first set of simulations. Simulation B addressed some questions of particular interest using another four sets of parameters, where both the rater and item parameters were simultaneously varied. This study found the HRM-SDT accurately recovered parameters, and clearly detected and separated changes in rater severity, rater central tendency, and item difficulty in most conditions.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This dissertation could not have been completed without the support of many people. First, I want to express my deeply thanks to my advisor, Professor Lawrence T. DeCarlo. My long journey in measurement started with his lecture in Psychometric Measurement six years ago. His deep knowledge and passion for research always inspired me on my way to becoming a true scholar. He inspired me to start my research study, guided me throughout the whole process of the analysis, and provided me intensive edits on my writing. Without his constructive feedback and continuous encouragement, I would not have been able to complete my dissertation.

Second, I would like to thank my dissertation committee members, Professor Young-Sun Lee, Professor James Purpura, Professor George Gushue, and Dr. YoungKoung Kim. My dissertation is a product of every conversation with all the committee members. I am writing this in a very heavy mood. My defense date (March 24, 2020) was in the darkest period of New York City, USA and the whole world. Due to the outbreak of COVID-19, the whole university closed and everyone was in a horror about this pandemic. Even in this situation, all of the committee members accepted my defense request and agreed to hold it remotely. Their passion for research sets me an example of dedicated scholars and stimulated me to always work professionally in my career.

Third, I would like to say thanks to Dr. Alex J. Bowers who provided me with fellowship and countless fun insights on using big data to tell stories in educational leadership. I really enjoyed the time to work in “data cave” with his special design of lights, multiple monitors and everything. It is a small room but is always filled with big ideas. Dr. Bowers’ aloud laugh always brings me the best memory in Teachers College.

Fourth, I want to express my sincerest thanks to my mom, Bin Li, and my dad, Xifeng Zhao. They are always the most important people supporting me and respecting every decision I have made. To support me to study abroad and chase my dream, they sacrificed a lot of family time and always endured loneliness without my accompany. I understand they always put me as the number one priority in their life and do whatever they can to make me live better. They are the kindest and the most diligent people in the whole world. Without the inspiration, drive, and support that they have given me, I might not be the person I am today.

Fifth, I want to thank my parent-in-law, Yan Shi and Baozhong Tang. In the past several years, without their supports on taking care of my kids, I cannot fulfill my academics and start my dream career in data science. They could live a better life in your country but traveled to a place without friends to help me take care of the whole family.

Sixth, I am so grateful I have two cute little ones, Lucas Tang and Louis Tang, in these three years. I am so proud and so blessed that I have these two lovely boys. As a new mom, I have suffered sleep deprivation for a long time and have faced so many unprecedented challenges in my life; however, their coming always highlights my life and bring me surprise and joy every day.

A special thank you to my husband, Yiming Tang. I would like to thank you for being there for me right from the first day we met. We came to the U.S. in our early twenties. In this new world, I have nothing but you. You are the only person who knows all my difficulties and shares all my happiness in these years. I am extremely lucky to have you and your unflinching care and love remain in my heart forever.

DEDICATION

To my parents, Bin Li and Xifeng Zhao,
my husband, Yiming Tang, who always love and support me

Chapter I

INTRODUCTION

The goal underlying educational testing is to measure psychological constructs in a particular domain and to produce valid inferences about examinees' ability. To achieve this goal of getting a precise ability evaluation, test developers construct questions with different formats. Two well-known formats are selected response questions, for example, multiple-choice (MC) items, and open-ended questions or constructed response (CR) test items, for example, essay items. In recent years, large-scale assessments, such as the Graduate Record Examinations (GRE) and the Scholastic Aptitude Test (SAT), have implemented CR items in addition to MC items as an essential component of the educational assessment landscape. One reason for the popularity of CR items is that they raise the overall measurement accuracy by reducing the test-wiseness strategies commonly used in MC items, such as guessing (Haladyna et al., 2002). Second, CR items offer more in-depth information on students who have insufficient skills or expertise in testing; otherwise, this information would be absent with MC items only (Ercikan et al., 1998). Third, CR items more closely resemble real-world tasks and are more authentic than MC items associated with the measured construct (Kim & Moses, 2013). Accordingly, many research studies have found that CR items function well in assessing more complex problem-solving skills and provide more diagnostic information than MC items (Kuo et al., 2016).

Along with their benefits, however, utilizing CR items in testing involves two main challenges, including rater effects and rater correlations, when multiple raters are assigned to one item. One challenge is the error added by human raters' subjective judgments in the measurement process since there are no correct preassigned responses in CR items scoring (Kim

& Moses, 2013). There are three main types of rater effects: severity, central tendency, and halo effects (Kingsbury, 1922; Leckie & Baird, 2011). First, the rater severity effect refers to the effect that raters may tend to give consistently low or high ratings that cause biased ability evaluation (Leckie & Baird, 2011). Second, central tendency describes when raters tend to use middle categories in the scoring rubric and avoid using extreme criteria (Saal et al., 1980). Third, a halo effect may occur in analytic scoring when raters are required to give several scores for different domains of an essay (Lai et al., 2012). Raters may fail to distinguish between different levels of performance among different domains, such as the effectiveness of topical control and language control. Researchers have found even experienced and well-trained raters seem unable to avoid all these rater effect, thereby impairing the precision of student ability evaluation and item parameter estimation (Cumming, 2007; DeCarlo, 2008; Myford & Wolfe, 2003).

The second challenge is that multiple raters usually grade an examinee's essay for quality control purposes; however, ratings based on the same item are correlated and need to be handled carefully by appropriate statistical procedures (Eckes, 2011; Kim, 2009). The use of multiple raters to grade one essay is similar to implementing repeated measurements on one experimental design, which also requires researchers to recognize the correlated data structure. Many studies have pointed out that ignoring the dependency among multiple ratings can result in underestimation of the standard errors for examinee ability estimation (DeCarlo, Kim, & Johnson, 2011; Donoghue & Hombo, 2000; Patz, Junker, Johnson, & Mariano, 2002).

To deal with the above two main problems in human ratings, researchers have proposed statistical models such as generalizability theory (G-theory; Brennan, 1992; Koretz, Stecher, Klein, & McCaffrey, 1994) and item response theory (IRT) (Linacre, 1989). Among all rater studies in the journal "Language Testing" and "Language Assessment Quarterly" between 2007

and 2017, 19.1% of the studies used the G-theory (Cronbach, Rajaratnam, & Gleser, 1963), 51.5% applied the many-facet Rasch measurement (MFRM) model under IRT framework (Linacre & Wright, 1993), and 29.4% implemented other methods (Heine et al., 2018). G-theory examines how to increase measurement precision by increasing the item length or/and adding more raters per item. To answer this question in the G-theory context, researchers usually first conduct a G-study to examine the source of error components, and then implement a D-study to examine the possibility of making error components small. G-theory thus estimates the sources of rater variance; however, it does not provide a solution to bias due to rater effects.

Accordingly, researchers have also proposed rater models based on the IRT framework that makes some improvements by estimating the rater bias to create a more precise measurement of examinee's proficiency. In the context of IRT, one commonly used model is the many-facet Rasch measurement (MFRM) model that partitions all the effects for examinees ability, item difficulty, and rater severity on a logit scale (Engelhard, 1992, 1994; Linacre, 1989). MFRM extends the polytomous IRT models, such as the generalized partial credit model used in the National Assessment of Educational Progress (NAEP) or the Rasch model used in PISA (Programme for International Student Assessment). The drawback of the MFRM model is that the measurement of examinee's ability approaches infinite precision when the number of raters increases (Patz, 1996; Patz, Junker, & Johnson, 2000). Furthermore, the MFRM model uses joint maximum likelihood estimation, which would result in statistical inconsistency and bias with a small sample size (Jong & Linacre, 1993). This problem forces researchers to realize that raters do not directly measure an examinee's proficiency but measure the quality of a CR item written by the examinee.

Patz (1996) recognized the hierarchical structure of the rating data and developed the hierarchical rater model (HRM). The HRM uses a signal detection model in the first level to estimate the relationship between ratings and the quality of the essay and an IRT model in the second level to model the relationship between the essay quality and the examinee's proficiency. There are two limitations to the model used in the first level of HRM. First, Patz et al. (2002) suggested that problems likely arise when estimating the rater severity parameters (the severity refers to if a rater is severe or lenient) under high rater discrimination parameters (the discrimination parameter indicates a rater's ability to differentiate essays in adjacent categories). Second, the model only accounts for rater effects of severity or leniency, but other effects, such as the central tendency and halo effect, are not considered.

To solve these problems, DeCarlo (2010) proposed an HRM-SDT model that extended the traditional signal detection theory (SDT) model used in the first level of HRM. The HRM-SDT model not only considers the hierarchical structure of rating data but also deals with various rater effects beyond rater severity. In recent years, some studies have been conducted to evaluate the performance of the HRM-SDT. DeCarlo (2008, 2010) assessed the model performance with simulated data and several real-world essays from Educational Testing Service (ETS). The results showed that the distance parameter, *d*, and criterion parameter, *c*, offered essential information to classify students regarding writing quality and evaluate raters based on scoring precision. Besides assessing CR items alone, HRM-SDT can be implemented by combining CR items with MC items (Kim, 2009). Kim (2009) simulated several situations to evaluate model performance under different design types and different CR item numbers. The result showed the HRM-SDT model accurately recovered rater parameters with or without MC items. Also, increasing the number of CR items and adding MC items can improve the estimation.

However, up to this point, there has been no simulation studies that have examined to what extent the HRM-SDT separates rater effects (i.e., rater severity and rater central tendency) from item effects (i.e., item difficulty). In other words, the reason for an examinee obtaining a low score can be a low examinee proficiency, a severe rater or/and a difficult item. Since the purpose of testing is to measure examinee proficiency and to eliminate the influence of raters and items, a good statistical model needs to separate these effects to achieve precise measurement goals. Thus, the ability to separate rater and item effects from raw rating scores is an essential issue needed to be addressed when evaluating the performance of the HRM-SDT model. Accordingly, one goal of this study was to simulate various rater effects and item effects to investigate the performance of the HRM-SDT model with respect to separating these effects. The other goal was to compare the fit of the HRM-SDT model with one commonly used model in language assessments, the Rasch model, in different simulation conditions and to examine the difference between these two models in terms of segregating rater and item effects.

Given these points, the purpose of the current study consisted of two aspects, (1) generating different datasets to evaluate how well the HRM-SDT model segregates rater effects (i.e., severity and central tendency) and item effects (i.e., item difficulty) from measured latent ability and (2) to compare how different rater models (Rasch and HRM-SDT) perform in terms of separating rater and item effects. In summary, this paper addresses the following research questions.

1. How well does the HRM-SDT model detect rater severity?
2. How well does the HRM-SDT model detect rater central tendency?
3. How well does the HRM-SDT model detect item difficulty (in Level 2)?

4. How well does the HRM-SDT model detect item effects and rater effects simultaneously?
5. How do the results compare to those obtained with the Rasch Model?

The present study begins, in Chapter 2, with a review of rater effects, item effects, and statistical models used for handling these effects. Chapter 3 illustrates data simulation methods. Chapter 4 presents the simulation analysis results. Finally, Chapter 5 discusses the findings and conclusions, as well as limitations of the study.

Chapter II

LITERATURE REVIEW

This chapter begins with an overview of rater effects (i.e., severity, central tendency, and halo effects) and item effects (i.e., item difficulty and discrimination), then follows with the introduction of four statistical models (G-theory, MFRM, HRM, HRM-SDT) that have been applied to discern rater and item effects. Finally, estimation methods are reviewed briefly.

2.1 Rater effects

The use of human raters to determine the quality of constructed response items involves raters' subjective judgments and likely produces some rater biases that affect measurement accuracy. Even if two raters agree with each other concerning the score of an essay, there is no guarantee for a correct rating or a valid assessment. Thus, rater effects are a significant factor that influences measurement accuracy in CR items. To reduce rater biases and increase measurement accuracy and validity, researchers need to understand rater effects and implement appropriate statistical models. Numerous studies have found significant differences among scores assigned by different raters to the same performance (Braun, 1988; Lunz et al., 1990). One of the reasons for these differences is due to rater effects that cause construct-irrelevant sources of variance and bring a threat to validity (Messick, 1995). Such effects may be related to raters' background, professional training, and work experiences (Wolfe et al., 1998). Kingsbury (1992) and Wolfe (2004) summarized three main types of rater effects, including severity, central tendency, and halo effects. These effects are related to how raters use a rating scale, which refers to a measurement instrument for raters to assign examinees to a position along the continuum to denote examinees' relative proficiency (Rahman et al., 2017).

2.1.1 Rater severity

Rater severity effect describes raters' tendency to offer consistently high or low ratings that add noises in ability evaluation (Leckie & Baird, 2011). Researchers have shown that rater severity significantly introduces systematic errors in student ability measurement for constructed response items (Engelhard Jr, 1994). Among all rater effects, rater severity is the most studied since it highly impacts measurement. When raters are assigned to the same essay, severe raters tend to give lower scores, whereas lenient raters tend to offer higher scores. To investigate how much deviation can be caused by rater severity, Wolfe (2004) conducted a study with 28 essays written by 28 students and scored by 101 raters from the Advanced Placement (AP) courses of English Literature and Composition. The result showed that 75% of examinees would pass if they were graded by the most lenient rater, whereas only 14% of examinees would pass if they were graded by the most severe rater. Thus, to assess the true ability of an examinee, the influence of rater severity needs to be considered.

2.1.2 Rater central tendency

Central tendency describes raters' propensity to award a score around the middle categories of the scoring rubric (Saal, Downey, & Lahey, 1980). In other words, central tendency occurs when raters define their criterion for the highest proficient group far to the right and their criterion for the lowest proficient group far to the left on the scoring rubric (DeCarlo, 2008; DeCarlo et al., 2011). Researchers have found that central tendency existed in many contexts, including in the assessment of the Advanced Placement English Literature and Composition essays, in school writing examinations in Georgia, in writing and speaking in English as a second language test, and in writing and speaking in German as a foreign language (Leckie & Baird, 2011). One reason for the existence of central tendency is that some raters have insufficient

knowledge to distinguish examinees along with the scoring rubric (Liu & Xie, 2014). The other reason is that raters may use a "play it safe" strategy to prevent from being too lenient or too severe (Rahman et al., 2017; Wolfe, 2004). Wolfe et al. (1998) analyzed the essay ratings from raters' cognitive behaviors and argued that expert raters are more likely to offer a broader range of scorings than less experienced ones. In an analysis of a large-scale writing assessment, however, DeCarlo et al. (2011) found evidence of central tendency. Accordingly, the detection of central tendency is relevant to the training of raters.

2.1.3 Halo effects

For an analytic scoring rubric, raters are required to provide ratings for several facets and to assign a separate score to each, which may be subject to halo effects (Lai et al., 2012). Halo effects refer to raters' preference to give an overall evaluative impression to each examinee rather than carefully distinguish levels of performance in multicomponent ratings (Saal, Downey, & Lahey, 1980). Halo effects cause several problems, such as reducing the number of independent constructs for examinees to demonstrate their proficiency, and resulting in a correlation between the scores of different facets (Lai et al., 2012; Viswesvaran et al., 2005).

Since halo effects only occur with an analytic scoring rubric and this study assumes a holistic scoring rubric where raters provide an overall score, the simulation studies only consider the first two rater effects (i.e., severity and central tendency).

2.2 Item characteristics

Besides rater effects, an item effect (the quality of an item) plays a vital role in constructing precise measurement. Without a qualified item, no matter how skilled a rater is, the measurement of examinee's ability can hardly reach a satisfactory level. Thus, the identification of variance due to item effects is equally as important as detecting rater effects. Under the IRT

framework, various item effects can be represented by the item characteristic function, which determines the probability of endorsing an item together with the examinee's ability (Hambleton & Swaminathan, 1985). In the item characteristic function, item difficulty and item discrimination are two parameters frequently used to evaluate item quality (Linacre, 1989). This function can be converted to an S-shaped curve, known as the item characteristic curve (ICC), that is used to define the relationship between a respondent's latent ability and his or her performance on a test item.

The shape of an ICC is identified by many item characteristics, among which are item difficulty, *b*, and item discrimination, *a*. Item difficulty is a location parameter that is defined as the amount of ability required to reach a 50% chance of getting an item correct. An item with a higher difficulty parameter requires a higher level of expertise to answer the item correctly. The typical range for item difficulty is from -3 to 3. The item discrimination parameter defines the steepness of the ICC, which indicates the extent that an item measures the trait. With higher item discrimination, even small changes in the examinee's ability can result in substantial changes in the probability of answering the item correctly. In other words, an item with higher discrimination performs better in terms of differentiating examinees.

Another item characteristic is the guessing parameter, *c,* which indicates respondents with a very low trait who still get the correct answer in multiple-choice questions. Since guessing is not an issue for constructed response items, this parameter is not examined in detail.

2.3 Item Response Theory models for Polytomous Responses

Since many CR scoring models are extensions of polytomous IRT models, this section briefly reviews IRT polytomous models. Generally, ratings can be generated in a dichotomous manner where one indicates correct and zero incorrect, or in a polytomous way where more than

two scoring categories are used. Accordingly, IRT models are classified by the number of categories in the rating scale, where dichotomous IRT models deal with the scale with two categories, and polytomous IRT models handle data with more than two categories.

Among polytomous IRT models, there are mainly two branches, including models for ordered response items and models for nominal response items. Ordered response items refer to items with options ordered in a prespecified way corresponding to the extent of skill completeness or the degree of approval, such as a five-level Likert-type scale item using five ordered categorical values to indicate five levels of agreement (i.e., strongly disagree, disagree, neither agree nor disagree, agree and strongly agree) (Naumenko, 2014). Other than ordered response items with ordered distractors, another type of polytomous response items, nominal response items, describe items with mutually exclusive and exhaustive categories without a specific order (Bock, 1997; Naumenko, 2014). An example of a nominal response item is a multiple-choice item, with no particular order among different options regarding the trait being measured. Since CR item scores have more than two categories in order to represent the extent to which an examinee endorses a skill, models dealing with CR items are derived from polytomous IRT models for ordered responses, such as the graded response model (GRM), the partial credit model (PCM) and the generalized partial credit model (GPCM). The following sections review these IRT models briefly.

2.3.1 Graded Response Model

Samejima (1969) proposed the Graded Response Model (GRM) as an extension of the two-parameter logistic (2PL) IRT model that assumes an examinee's probability of endorsing an item depends on his or her ability together with the item difficulty and discrimination. Similar to the 2PL IRT, the GRM models the probability of scoring in a given category *k* or higher by the

item difficulty and discrimination as well, where $k = 1, \ldots, K$ represents the response categories. The GRM models the probability of endorsing a response category or higher category using cumulative logits, or the logarithm of the odds, written as follows,

$$log\left[\frac{P(Y_{il}\geq k|\theta_i)}{P(Y_{il}<k|\theta_i)}\right] = a_l(\theta_i - b_{lk}), \quad (1)$$

where Y_{il} denotes the score of an examinee i (i=1, …, N) on item l (l=1, …, L); θ_i represents the proficiency of examinee i on the measured trait; a_l indicates the discrimination for item l; and b_{lk} refers to the threshold parameter for each response category k. Note that a_l is assumed to be the same for all categories and b_{lk} is strictly ordered (i.e., $b_{l1} < \cdots < b_{lK}$).

2.3.2 Generalized Partial Credit Model and Partial Credit Model

The other two commonly used polytomous IRT models are the partial credit model (PCM) proposed by Masters (1982) and the generalized partial credit model (GPCM) developed by Muraki (1992). The PCM is a special case of the GPCM with a uniform discrimination parameter ($a_l = 1$). Both PCM and GPCM use *adjacent-category logits* rather than the cumulative logits used in the GRM (Agresti, 1990).

The generalized partial credit model (GPCM) is,

$$\text{Log}\left[\frac{P(\text{Y}_{il}=\text{k}+1|\theta_i)}{P(\text{Y}_{il}=\text{k}|\theta_i)}\right] = a_l(\theta_i - b_{lk}), \quad (2)$$

where Y_{il} is the score of examinee i on item l; θ_i is the ability of examinee i; a_l is the discrimination for item l; and b_{lk} is the increase in item difficulty from category k to category k+1. Masters (1982) refers to b_{lk} as the "step" or transition probabilities between adjacent ratings, k and k+1. For example, the first step parameter b_{l1} determines the probability of getting a rating of 1 to a rating of 2; the second step parameter b_{l2} determines the probability of getting a

rating of 2 to a rating of 3, and so on. Same as in the GRM, b_{lk} is strictly ordered (i.e., $b_{l1} < \cdots < b_{lK}$).

The GPCM assumes a_l to be the same across all categories, but it can vary across different items. As a particular case of GPCM, PCM not only assumes the same a_l for all categories, but also for all items ($a_l = 1$).

2.4 Models for Constructed Response Items

Regarding statistical models, generalizability theory (G-theory; Brennan, 1992; Koretz, Stecher, Klein, & McCaffrey, 1994) and item response theory (IRT; Linacre, John, & Wright, 2002) are two main psychometric domains implemented to analyze constructed response items.

2.4.1 Generalizability Theory

Generalizability theory (G-theory) arises from classical test theory (CTT) as a traditional method to model constructed response items through decomposing the error components into systematic variability and random error variability (Cronbach, Rajaratnam, & Gleser, 1963; Brennan, 1992; Shavelson & Webb, 1991; Lynch & McNamara, 1998). To be specific, G-theory partitions an observed score into different sources of main effects and all kinds of interactions.

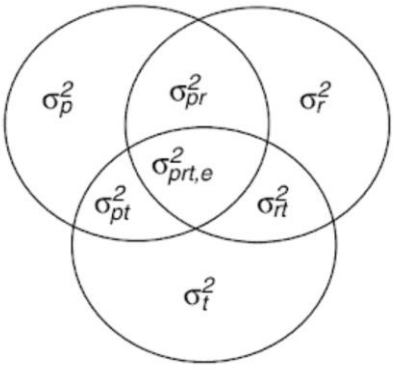

Figure 1. A Venn diagram showing the partitioning of variance in G-theory (Bandalos, 2018).

Figure 1 is a Venn diagram illustrating how the total variance partitions into additive variance components, such as three main effects (i.e., examinee, item and rater), three two-way interactions (person × task, person × rater, rater × task), and a three-way interaction (person × task × rater) (Bandalos, 2018; Robitzsch & Steinfeld, 2018). Equation 3 explains the idea of variance partitioning in a function as follows

$$\sigma^2(X_{prt}) = \sigma_p^2 + \sigma_r^2 + \sigma_t^2 + \sigma_{pr}^2 + \sigma_{pt}^2 + \sigma_{rt}^2 + \sigma_{prt,e}^2, \quad (3)$$

where σ_p^2 represents the variance in observed scores due to person facet; σ_r^2 denotes the variance due to rater facet; σ_t^2 refers to the variance due to test facet; σ_{pr}^2 represents the variance due to the interaction between person and rater; σ_{pt}^2 indicates the variance due to the interaction between person and test; σ_{rt}^2 refers to the variance due to the interaction between rater and test; and $\sigma_{prt,e}^2$ indicates the variance related to the interaction between person, rater, test, and error.

In the context of G-theory, two approaches are considered to improve measurement precision, including adding more items or/and adding more raters for each item (Patz et al., 2002). The choice between the two or both of the two can be determined by two steps (1) a generalizability study (G-study) conducted to estimate different types of variance components; (2) a decision study (D-study) implemented to minimize the influence from error components. The G-theory functions well in detecting the formulation of the total variance, but it inappropriately regards the relationship to be additive among all components. Another drawback of G-theory is that it assumes the scores to be on a continuous scale; however, constructed response items are on a discrete scale in most cases (Smith Jr & Kulikowich, 2004). To overcome these limitations, researchers prefer IRT models, which have been more widely employed than G-Theory, such as in the Program for International Student Assessment (PISA)

and US National Center for Education Statistics (NCES), and are used in the products of large testing companies such as Educational Testing Services (ETS) (Choi & Wilson, 2018).

2.4.2 Many-Facet Rasch Measurement

The Many-Facet Rasch Measurement (MFRM) model extends the IRT model by incorporating more facets into IRT framework such as raters and scoring criteria, in addition to examinees and items (Linacre, 1989; John M. Linacre & Wright, 2002). In general, a facet represents any factor that affects test scores including those of substantive interest (i.e., examiner ability, item difficuley, rater severity) (Eckes, 2009, 2011). Besides the main effect brought by each facet, there are interaction effects among facets, including two-way interactions, such as an interaction between examinees and raters, and three-way interaction, such as an interaction among examinees, raters, and items. Accordingly, researchers need to build hypotheses to decide possible facets before analyzing data under the MFRM model, which is similar to the D-study under the G-theory framework illustrated in the previous section.

To compare G-theory and the MFRM model in terms of detecting rater effects, Lynch and McNamara (1998) investigated these two models with four raters scoring 83 essays in a fully crossed design, which means each rater scored all 83 essays. The authors used the software FACETS (Linacre & Wright, 1993) to fit the MFRM model and the software GENOVA (Crick & Brennan, 1984) to perform the G-theory analysis. The results showed that the MFRM model and G-theory varied in recognizing rater effects, especially the interaction between raters and examinees. To be specific, the MFRM model detected 36% of the variance attributed to the interaction between raters and examinees, whereas G-theory only accounted for 3% of the variance related to the rater and examinee interaction. Lynch and McNamara (1998) concluded that the MFRM model tended to magnify rater effects, but can provide detailed information for

test design, while G-theory showed aggregated information about possible factors and gave an overall suggestion about test design.

The MFRM model is derived from the Partial Credit Model (PCM), but adds rater effects. Given the latent ability θ_i of each examinee i, the probability that rater j assigns examinee i's response to item l in category m can be written as

$$log\left[\frac{P(Y_{ilj}=k+1|\theta_i)}{P(Y_{ilj}=k|\theta_i)}\right] = \theta_i - b_l - \gamma_k - c_j, \quad (4)$$

where Y_{ilj} is the score of examinee i on item l for rater j; θ_i denotes the ability of examinee i; b_l represents the item difficulty of item l; γ_k refers to the item step parameter that indicates the change in item difficulty of receiving a rating of $k+1$ relative to a rating of k; c_j denotes rater severity, the tendency of the rater to be lenient or strict.

Note that the step parameter γ_k is a transition point, where the probability is 50% that an examinee is being rated in two adjacent categories, k+1, and k. This point is also called *Rasch-Andrich thresholds* (Andrich, 1998). Moreover, the sum of item difficulty and item step parameter equals the threshold parameter ($b_l + \gamma_k = b_{lk}$) in the graded response model shown in Equation 1 and the generalized partial credit model shown in Equation 2. Furthermore, the rater severity parameter c_j shifts the ICC up and down on the examinee ability scale.

The MFRM model provides a way to examine the effect of rater subjective judgment within an IRT framework; however, it regards rater severity as the only type of rater effect. In reality, raters can vary concerning their ability to discriminate between different essays and different categories of CR items. MFRM tends to have three disadvantages. First, a fundamental flaw argued by Patz et al. (2002) is that the MFRM model tends to give an infinitely precise prediction of examinees' latent proficiency as the number of raters per item increases. Second, MFRM assumes ratings among different raters to be independent, but ignores the correlation

among multiple raters when assigned to one common item. Since assigning more than one rater to an item is a standard method of reducing rating bias and controlling rating quality in CR designs, the correlation among ratings needs to be handled carefully; otherwise, ignoring this likely dependency results in an underestimation of standard errors (Wilson & Hoskens, 2001). Third, the MFRM model assumes that all raters have the same discrimination, but ignores the fact that raters' discrimination may differ due to their expertise and training.

2.4.3 The Hierarchical Rater Model (HRM) of Patz et al. (2002)

To address problems in MFRM, the hierarchical rater model (HRM) accounts for individual rater effects when evaluating examinee proficiency with multiple ratings. In the HRM framework, ratings do not directly indicate examinee proficiency, but reflect the latent quality of each essay. Since the true category for an essay cannot be directly observed, a rater's task is a signal detection process to assess the true category, and so the rater's judgment is an unreliable indicator of examinees' ability.

Patz (1996) recognized the hierarchical structure existing in rating data. In the first level, a discrete signal detection model holds between ratings and the true category of an essay. In the second level of HRM, the estimated true category of the essay from the first level serves as an indicator of examinee proficiency via an IRT model. The curved arrows represent a nonlinear relationship. Different from visible indicators in the traditional IRT model, indicators in the second level of HRM are unobservable or latent. The following figure represents the two-level structure in the HRM.

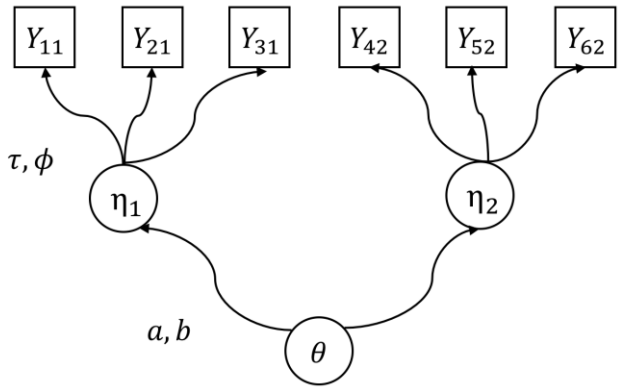

Figure 2. A representation of the HRM of Patz et al. (2002); from DeCarlo et al. (2011).

Level 1: The rater model

Figure 2 illustrates an HRM example, where Y_{jl} represents the score from the *j*th rater on the *l*th item (DeCarlo et al., 2011). In the first level, the first three raters are assigned to the first item, denoted as Y_{11} , Y_{21} and Y_{31} , and the second three raters are allocated to the second item, denoted as Y_{42} , Y_{52} and Y_{62} . These six raters attempt to detect the true latent category for each writing sample, indicated by η_1 for the first item and η_2 for the second item. The Level 1 parameters are τ_{jl} denoting rater precision and ϕ_{jl} representing rater severity. Note that the arrows from η_l to Y_{jl} are curved indicating nonlinear relations. The first level of the HRM of Patz et al. (2002) can be written as,

$$P(Y_{jl} = k|\eta_l = \eta) \propto exp\left\{-\frac{1}{2\psi_{jl}^2}[k - (\eta - \phi_{jl})]^2\right\}, \quad (5)$$

where Y_{jl} is the response given by the *j*th rater to the *l*th item with ordinal scores from 1 to *K* and η_l is a latent categorical variable representing the examinee proficiency for the *l*th item. As noted by Patz (2002), ψ_{jl}^2 is a variance parameter for rater *j* on item *l* measuring the extent to which a rater shows a lack of reliability; its inverse, $\tau_{jl} = \frac{1}{(2\Psi_{jl}^2)}$, is a measure of rater precision. The parameter ϕ_{jl} represents rater severity, with higher values indicating a more severe rater and vice

versa. Additionally, the probabilities for each response category are assumed to be approximately normally distributed in the first level of the HRM.

Level 2: IRT model for polytomous responses

The second level of the HRM models the relationship between the estimated true category of an essay from the first level and examinee's true proficiency. In this level, Patz (1996) utilized IRT polytomous models, including the PCM (Masters, 1982) and the GPCM (Muraki, 1992).

In Figure 2, the true category for the first essay (η_1) and that for the second essay (η_2) serve as indicators of an examinee's latent proficiency θ. The GPCM can be written as

$$\log\left[\frac{P(\eta_{il}=\eta+1)}{P(\eta_{il}=\eta)}\right] = a_l(\theta_i - b_{lm}), \quad (6)$$

where η_{il} denotes the latent categorical variable representing the examinee $i's$ (i=1, …, N) proficiency for the lth item (l=1, …, L) on values η from 0 to $M-1$; θ_i refers to the proficiency of examinee i; a_l indicates item discrimination for item l; and b_{lm} is a threshold parameter for each ordered response category, from 1 to M, in Item l. Note that the PCM is a special case of the GPCM with a uniform discrimination parameter ($a_l = 1$). Furthermore, an assumption is that the number of latent classes, M, in the second level (or in Equation 6) is equal to the number of scoring categories, K, in the first level (in Equation 5).

HRM explicitly recognizes the hierarchical structure of the data and handles the nesting issue between raters and items and so it reduces standard error estimation problems, as in the MFRM. However, there are some limitations. First, the model has only one rater parameter, ϕ_{jl}, to measure rater severity or leniency (DeCarlo et al., 2011). In the real-world, there are many factors other than rater severity, such as central tendency, where raters give scores in the middle of a rating scale, and Patz et al.s' HRT doesn't recognize these factors. Second, when rater

discrimination is high (i.e., raters with small values of ψ_{jl}^2), there are problems obtaining estimates of the rater severity parameter (DeCarlo et al., 2011; Patz et al., 2002). As shown in Equation 5, with relatively small ψ_{jl}^2, the likelihood function of ϕ_{jl} is nearly a constant value in the range of -.5 to .5, which results in difficulty in estimating the value of ϕ_{jl}.

2.4.4 The Hierarchical Rater Model (HRM) with a Latent Class SDT by DeCarlo et al. (2011)

DeCarlo et al. (2011) incorporated a latent class extension of the signal detection theory (SDT) model (DeCarlo, 2002; 2005) in Level 1 of the HRM, and referred to the model as the HRM-SDT. The model addressed the limitations of the traditional HRM approach with only a single rater effect parameter. Under the HRM-SDT framework, the process of scoring CR items reflects two aspects of the rater's psychological processes, including a perceptual aspect (a rater's perception of the quality of a CR item) and a decision aspect (a rater's decision criteria for the scoring rubric categories; see DeCarlo, 2010).

To illustrate the complete HRM-SDT model, DeCarlo et al. (2011) presented a path diagram with a latent class SDT in the first level and an IRT model in the second level, as shown in Figure 3. In the path diagram, observable variables are placed in boxes, and latent variables are placed in ovals (Bollen, 1989; Schumacker & Lomax, 2012). All independent variables have arrows pointing to dependent variables with the weighting coefficient above the arrow. Figure 3 illustrates the HRM-SDT with six raters scoring two CR items for each examinee. In the first level, Y_{jl} represents the observed responses from rater *j* on item *l* with *K* categories, and Ψ_{jl} indicates the rater's perception of the overall quality of the *l*th CR item. In the second level, η_l denotes the true category for the *l*th CR item with *M* categories, and θ represents the latent proficiency of the examinee.

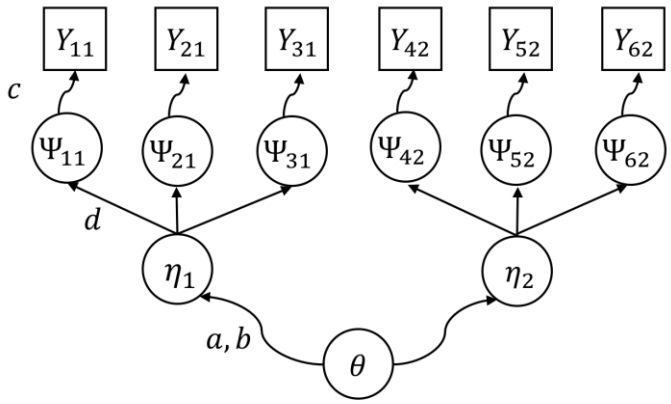

Figure 3. The HRM-SDT model (from DeCarlo et al., 2011)

Level 1: The rater model

The first level of the HRM-SDT models a rater's decision as to which category an essay belongs to, which serves as an indicator of an examinee's proficiency in the second level. A rater's task is a signal detection process, which depends on his or her perception of essay quality and his or her criteria for a decision. The perception of the quality of an essay is a latent continuous variable denoted by Ψ from a location-family probability distribution, such as logistic or normal (DeCarlo, 1998; Peterson, Birdsall, & Fox, 1954). For each category of rating, the distribution Ψ has different locations. Figure 4 displays an example of the signal detection theory with four latent classes and a 1-4 rating response.

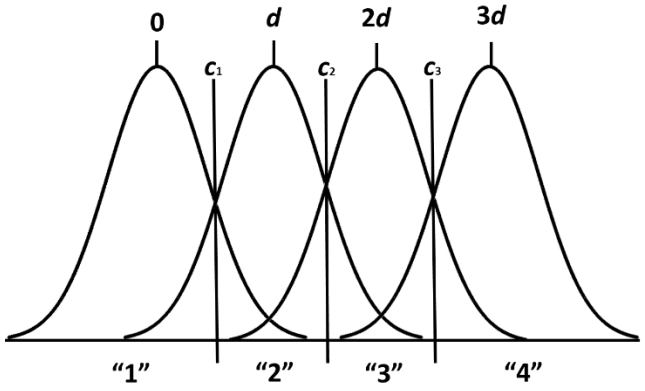

Figure 4. An example of SDT with four response categories (from DeCarlo et al., 2011)

Figure 4 illustrates the equal distance model where an assumption is made that the distance parameter for item *l*, d_{jl}, is identical across all response categories, but varies across the *j* raters (DeCarlo, 2002). The distance between each perceptual category describes a rater's ability to distinguish essays in adjacent categories, with a larger value indicating higher discrimination. Accordingly, the distance parameter functions similarly to the item discrimination parameter in traditional IRT. If the distance parameter is smaller, the overlap between adjacent distributions is greater and the misclassification probability is greater.

The other parameter in SDT is a location parameter for item *l*, c_{jkl}, which is a rater's subjective criteria for assigning a score to an essay, and is responsible for rater effects, if any. That is to say, raters' arbitrary criteria for using the scoring rubric determine if they prefer to use higher scores (lenient raters) or lower scores (severe raters). For example, if a rater perceives an essay between c_1 and c_2, the rater scores the essay as being in the 2nd category. When a rater's criterion c_1 shifts up along the x-axis, the rater becomes more severe and is more likely to assign a lower "1" response than a "2" response. That is to say, moving the location parameter, c_{jkl}, to the right of the x-axis indicates a more severe rater, whereas shifting the location to the left reflects a more lenient rater.

With the distance parameter, d_{jl}, and the criterion parameter, c_{jkl}, the HRM-SDT can account for many rater effects, such as rater severity, central tendency, restriction of rating range, and even unequal discrimination between different categories (DeCarlo & Zhou, 2020). The latent class signal detection model in Figure 4 can be written as:

$$P\left(Y_{jl} \leq k \middle| \eta_l = \eta\right) = F\left(\mathrm{c}_{jkl} - \mathrm{d}_{jl}\eta_l\right), \tag{7}$$

where Y_{jl} refers to the response given by the *j*th rater on the *l*th item and takes *k* discrete scores (*k* = 1, …, and *K*); η_l represents the latent category for item *l* and takes *m* discrete scores (*m* = 1, …, and *M*); *F* indicates a cumulative location-family distribution function, such as a logistic or normal distribution; d_{jl} denotes a distance parameter for the *j*th rater on the *l*th item; c_{jkl} represents a response criteria for the *j*th rater on the *l*th item in the *k* category, which is strictly ordered $c_{j1l} < c_{j2l} < \cdots < c_{j,K-1,l}$ with $c_{j0l} = -\infty$ and $c_{jKl} = \infty$.

In recent years, there have been some studies evaluating the performance of the HRM-SDT regarding how precise the model can recover rater parameters; however, HRM-SDT has been insufficiently studied compared to MFRM and HRM. DeCarlo (2008) assessed the model with simulated data in a fully-crossed design and a Balanced Incomplete (BIB) design, and then applied it to several real datasets from the essay writing section of tests from Educational Testing Service (ETS). The results obtained from the simulation study showed that the distance parameter, d_{jl}, and criterion parameter, c_{jkl}, offer essential information for classifying students in terms of writing quality and for evaluating raters based on scoring precision. Concerning the real-world dataset from ETS, the HRM-SDT provided an informative evaluation of rater behavior.

To account for more study designs used in practice, DeCarlo (2010) examined the HRM-SDT performance in incomplete and unbalanced designs. The "incomplete" means that each essay is not scored by all raters, and the "unbalanced" indicates that raters are assigned a different number of essays. Besides simulated data analysis, DeCarlo (2010) applied the HRM-SDT to a large-scale language test with two writing tasks and a mathematics test with three problem-solving questions (ETS tests). The results showed that the HRM-SDT could be easily

incorporated into different types of study designs, and it provides useful information about the ability of examinees, the performance of raters, and the characteristics of items.

DeCarlo et al. (2011) reviewed the limitations of the HRM (Patz et al., 2002) and explained the rationale for developing the HRM-SDT in detail. Also, they applied the HRM-SDT to a real-world dataset and found that it provided a valuable way to summarize raters' behavior and estimate examinees' ability. Besides assessing CR items, HRM-SDT can also incorporate MC items in addition to CR items (Kim, 2009). Kim examined several situations in order to evaluate model performance under different design types and different numbers of CR items. The results showed that the HRM-SDT model accurately recovered rater parameters with or without MC items. Also, increasing the number of CR items and adding MC items improved estimation. Compared to other models that combine CR items and MC items, HRM-SDT provided the most accurate ability estimates.

2.5 Estimation methods

In general, there are two main approaches to parameter estimation in rater models: maximum likelihood estimation (MLE) and Bayesian estimation. The fundamental idea under MLE is to find the parameter values that maximize the probability of obtaining the observed data (Fisher, 1925). MLE has many advantages: (1) sufficiency (MLE estimator provides complete information about the estimated parameters); (2) consistency (true parameter values are recovered asymptotically with sufficiently large sample size); (3) efficiency (MLE offers relatively low parameter estimation variance); (4) parameter invariance (the MLE solution is independent of the parameterization) (Myung, 2003). Robitzsch and Steinfeld (2018) reviewed marginal maximum likelihood (MML), joint maximum likelihood (JML), conditional maximum likelihood (CML), and presented R syntax to apply these estimation methods to rater models.

The other parameter estimation method, Bayesian estimation, such as Markov chain Monte Carlo (MCMC), has become popular recently due to the development of Bayesian software such as WinBUGS (Lunn et al., 2000), OpenBUGS (Sturtz et al., 2005), JAGS (Plummer, 2003), and Stan (Gelman et al., 2015). The MCMC approach is a sampling method to characterize a distribution by randomly drawing latent variables and parameters conditional on the information from the observed data (van Ravenzwaaij et al., 2018). When dealing with many latent variables such as IRT models, Patz & Junker (1999) claimed that MCMC is computationally superior to MLE, and they utilized MCMC methods to estimate HRM.

However, boundary problems often occur; a boundary problem occurs when a parameter estimate is large or indeterminate or a latent class size equals zero or unity, which presents problems (DeCarlo et al., 2011; Garre & Vermunt, 2006; Maris, 1999; Schafer, 1997). To handle boundary problems, DeCarlo et al. (2011) implemented posterior mode estimation (PME), a partially Bayesian approach, with the HRM-SDT model. Instead of maximizing the log-likelihood, PME maximizes the log posterior function, which can be written as follows

$$\text{posterior} \propto \text{prior} \times \text{liklihood}. \tag{8}$$

Equation 8 shows that the posterior is proportional to the prior. In this way, the prior adds a penalty for solutions close to the boundary. DeCarlo et al. (2011) discussed how to place priors on the conditional probabilities for responses and latent classes rather than directly on the model parameters. The PME method can be implemented via software such as Latent Gold 4.5 (Vermunt & Magidson, 2003, 2005).

In summary, this chapter discussed rater effects and item effects, different models for constructed response items, and common estimation methods in rater models. In the following

chapter, data generation methods for different combinations of rater effects and item effects are explained in detail.

Chapter III

METHODS

This chapter begins with an outline of the data generation procedure for the HRM-SDT, where the first stage simulates examinees' latent category data corresponding to the IRT framework (the second level in HRM-SDT) and the second stage simulates ratings utilizing the signal detection theory framework (the first level in the HRM-SDT). In Section 3.2, I show how seven sets of parameters were varied in the first set of simulations, referred to as Simulation A. In Section 3.3, a second set of simulations, Simulation B, are presented. The second set addressed some questions we had in a slightly different way and included some conditions of particular interest, where both the rater and item parameters were simultaneously varied. In Section 3.4, we discuss the criteria to evaluate model performance for the HRM-SDT and the Rasch model.

3.1 Data generation procedures

Data were generated using a SAS macro written by DeCarlo (2010), with some modifications. Rater parameters and item parameters were varied across the different conditions. The following aspects of the conditions remained the same for each simulation:

(1) Three CR items were assigned to 1000 examinees, with each examinee responding to all three items.

(2) Nine raters were assigned to score these three CR items with each item scored by three out of nine raters.

(3) All the ratings were scored using a 1-4 scale.

(4) For each simulation condition, 100 replications were conducted.

With the above specified experimental conditions, data generation followed two main stages, reflecting the two levels of the HRM-SDT that are shown in Figure 5. In Figure 5, boxes show observable variables, and ovals show latent or unobservable variables (Bollen, 1989; Schumacker & Lomax, 2012). Figure 5 illustrates an example with nine raters scoring three CR items for each examinee. In the first level, Y_{jl} represents the scoring responses given by rater *j* on item *l*, and Ψ_{jl} indicates the rater *j*'s recognition of the quality of the *l*th item. In the second level, η_l denotes the true category of the *l*th item and θ represents the latent proficiency of the examinee.

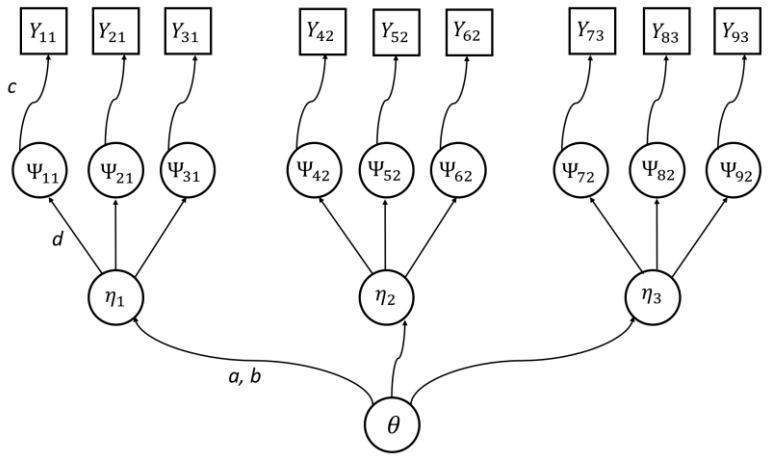

Figure 5. The HRM-SDT model with three essays scored by nine raters (DeCarlo et al., 2011)

The first stage corresponds to the examinee level or the second level of HRM-SDT and involves the following four steps. First, 1000 examinee latent abilities were generated from a normal distribution with a mean of zero and a standard deviation of unity, denoted as $\theta_i=\theta_1, \theta_2, \ldots, \theta_{1000}$. Second, with the generated examine abilities, θ_i's, the probability of an examinee belonging to a particular latent class for each CR item was obtained using the GPCM of Equation 8 as follows,

$$\log\left[\frac{P(\eta_{il}=\eta+1)}{P(\eta_{il}=\eta)}\right] = a_l(\theta_i - b_{lm}), \quad (8)$$

where η_{il} represents the categorical latent proficiency of examinee i (i=1, …, N) on the lth item (l=1, …, L), with values η from 0 to $M-1$, θ_i is the latent proficiency of examinee i, a_l is the discrimination of item l, and b_{lm} is the threshold for each response category from 1 to M on item l. In the current simulations, i=1000 represents 1000 examinees, l=3 denotes three items, and m=3 indicates three thresholds to give four latent categories. Based on different rater effects and item effects, the three item discrimination parameters (a_1, a_2, a_3) and nine item difficulty parameters ($b_{11}, b_{12}, b_{13}, b_{21}, b_{22}, b_{23}, b_{31}, b_{32}, b_{33}$) are shown in the next section. Third, the cumulative probability for each scoring category was calculated, such as $P(\eta_{il} \leq 1)$, which refers to the probability that a student was scored as being in class one or less. Accordingly, three cumulative probabilities are obtained for each CR item, denoted as $P(\eta_{i1} \leq 1)$, $P(\eta_{i1} \leq 2)$, and $P(\eta_{i1} \leq 3)$ for the first item, $P(\eta_{i2} \leq 1)$, $P(\eta_{i2} \leq 2)$, and $P(\eta_{i2} \leq 3)$ for the second item, and $P(\eta_{i3} \leq 1)$, $P(\eta_{i3} \leq 2)$, and $P(\eta_{i3} \leq 3)$ for the third item. Fourth, 3000 values were randomly generated from a uniform distribution ranging from 0 to 1 and then were compared to the cumulative probabilities for the three CR item scores obtained from the third step. For example, for the first item answered by the ith examinee, if the uniform value was less than $P(\eta_{i1} \leq 1)$, a value of one was assigned as the latent category and was denoted as $\eta_1 = 1$. If the uniform value was less than $P(\eta_{i1} \geq 2)$ but greater than $P(\eta_{i1} \geq 1)$, then the latent category for η_1 was 2; if the uniform value was less than $P(\eta_{i1} \geq 3)$ but greater than $P(\eta_{i1} \geq 2)$, $\eta_1 = 3$; and if the uniform value was greater than $P(\eta_{i1} \geq 3)$, $\eta_1 = 4$. The same steps were used for the second and the third CR item as well.

The second stage of the data simulation involved the first level of the HRM-SDT. First, raters' probabilities for each response category were calculated based on the SDT model illustrated in Equation 9 as follows,

$$P\left(Y_{jl} \leq k \middle| \eta_l = \eta\right) = F\left(c_{jkl} - d_{jl}\eta_l\right), \quad (9)$$

where Y_{jl} represents the score of the *j*th rater on the *l*th item using *k* discrete scores, η_l denotes the true latent categories for each examinee, *F* denotes a cumulative distribution function, d_{jl} is the distance parameter for the *j*th rater on the *l*th item, and c_{jkl} is the response criteria for the *j*th rater on the *l*th item in the *k*th category. The rater criteria parameter, c_{jkl}, and discrimination parameter, d_{jl}, are shown in the next section for the different simulation conditions.

In the first step, cumulative probabilities for each scoring rubric were obtained as $P\left(Y_{jl} \leq 1 \middle| \eta_l = \eta\right)$, $P\left(Y_{jl} \leq 2 \middle| \eta_l = \eta\right)$, and $P\left(Y_{jl} \leq 3 \middle| \eta_l = \eta\right)$ using three thresholds for four response categories. Second, 9000 values were randomly generated from a uniform distribution ranging from 0 to 1 and were compared to the cumulative probabilities in the previous step to get ratings from 1 to 4. For example, for the first item rated by the first rater, if the uniform value was less than $P(Y_{11} \leq 1|\eta_l = \eta)$, then a value of one was assigned as the observed score for the examinee, denoted as $Y_{11} = 1$. If the uniform value was less than $P(Y_{11} \leq 2|\eta_l = \eta)$ but greater than $P(Y_{11} \leq 1|\eta_l = \eta)$, then $Y_{11} = 2$; if the uniform value was less than $P(Y_{11} \leq 3|\eta_l = \eta)$ but greater than $P(Y_{11} \leq 2|\eta_l = \eta)$, then $Y_{11} = 3$; and if the uniform value was greater than $P(Y_{11} \leq 3|\eta_l = \eta)$, then $Y_{11} = 4$. The same steps were used for all three raters for the three CR items.

3.2 Simulation A

Table A1 shows the rater population parameters for the seven simulation conditions of Simulation A. Condition A1 served as a baseline that was used to make comparisons with the

other six simulation conditions. Table A2 shows the item population parameters for seven simulation conditions where item step parameters were varied to indicate more difficult and easier items.

Simulation Condition A1 (Baseline)

Condition A1 served as a baseline used to make comparisons with the other six conditions in Simulation A. For the baseline condition, the estimated rater and item parameters from previous studies with real-world data were used for the population values (Kim, 2009; DeCarlo, 2008). The values for the item discrimination parameters, a_l, were 0.5, 1, and 2, where *l* denotes different items. The range of item difficulty parameters, b_{lm}, was from −3 to 3, which covers the range found in practice (Kim, 2009). The population rater discrimination parameters, d_{jl}, referring to the *j*th rater on the *l*th item, were varied from 2 to 6, which also reflects the range found in real-world data analyses (DeCarlo, 2008). The rater criteria parameters, c_{jkl}, represent the intersection points for adjacent distributions, where *j* stands for nine raters, *k* represents four different score categories, and *l* refers to three different items. In a symmetric distribution, these intersection points are located at the midway points between distributions. For example, Rater 1's discrimination parameter for Item 1, d_{11}, was 2, indicating the distance between two perceptual categories was 2. Under an equal distance model, the rater's criteria parameters are midway between each of the perceptual distributions. Accordingly, Rater 1's first criteria parameter, c_{111}, was 1, which is the midway of the first distribution at zero and the second distribution at 2; the second criteria, c_{121}, is at 3, which is midway between the second distribution at 2 and the third at 4; and the third criteria, c_{131}, was 5, and so on. These locations are optimal (with other assumptions) in terms of maximizing the proportion of correct responses (DeCarlo, 2010).

Simulation Condition A2 (Rater Severity)

Condition A2 investigated the effects of changing rater severity in the first level of HRM-SDT. In this condition, only rater severity varied, and other parameters were kept the same as in baseline. In Condition A2, the rater criteria parameters, c_{jkl}, for Rater 1, Rater 4, and Rater 7 were increased by one compared with those in Condition A1 indicating that these three raters are more severe; the criteria parameters for Rater 2, Rater 5, and Rater 8 were decreased by one, indicating that these three raters are more lenient; the criteria parameters for Rater 3, Rater 6, and Rater 9 were kept the same as in Condition A1. Furthermore, the item parameters were the same as in the baseline. In short, one more severe rater, one more lenient rater, and one same rater were assigned to each of three items.

Simulation Condition A3 (Rater Central Tendency)

Condition A3 examined parameter recovery when raters show 'central tendency'. To generate this scenario, the first and the third rater criteria parameters, c_{j1l} and c_{j3l}, for Raters 1, 4, and 7 were shifted 0.5 down and up compared to the baseline criteria parameters; c_{j1l} and c_{j3l} for Rater 2, Rater 5, and Rater 8 were shifted one unit down and up; c_{j1l} and c_{j3l}, for Rater 3, Rater 6, and Rater 9 were kept the same. Accordingly, shifting the criteria in this way reduces the frequency of the usage of the end categories, which is the central tendency effect. Item parameters for Condition A3 were kept the same as in the baseline.

Simulation Condition A4 (Item Difficulty)

In Condition A4, Item 1 was made one unit more difficult than the baseline, and Item 3 was made one unit more lenient than the baseline. To be specific, the item difficulty parameters, b_{lm}, were shifted one unit upward in the scoring rubric for Item 1 and were shifted one unit

downward in the scoring rubric for Item 3. All the rater-level parameters were kept the same as in the baseline.

Simulation Condition A5 (Rater Severity and Item Difficulty)

The purpose of Condition A5 was to examine the effects of simultaneously varying rater effects and item difficulty. Accordingly, rater criteria parameters were manipulated in the same way as those in Condition A2, where Rater 1, 4, and 7 became more severe, and Rater 2, 5, and 8 became more lenient. Additionally, item difficulty parameters were altered in the same way as those in Condition A4, where Item 1 became one unit more difficult, and Item 3 became one unit easier.

Simulation Condition A6 (Two More Lenient Raters Assigned to One More Difficult Item and Two More Severe Raters Assigned to One Easier Item)

In Condition A6, parameter recovery was examined when the rater parameters and the item parameters were shifted in opposite directions. To be specific, two more lenient raters (Rater 1 and Rater 2) were associated with a more difficult item (Item 1) and two more severe raters (Rater 7 and Rater 8) were associated with an easier item (Item 3). Note that lenient raters tend to give higher scores, whereas difficult items usually result in lower scores. As a result, combining lenient raters and difficult items might create problems with parameter recovery (it's a kind of confounding); a basic and important question is whether rater effects and item effects can be separated by the HRM-SDT, and this and the next condition examine this question.

Simulation Condition A7 (Three More Lenient Raters Assigned to One More Difficult Item and Three More Severe Raters Assigned to One Easier Item)

In Condition A7, three more lenient raters (Rater 1, Rater 2, and Rater 3) were assigned to a more difficult item (Item 1) and three more severe raters (Rater 7, Rater 8, and Rater 9) were assigned to an easier item (Item 3). In Condition A6, we kept two raters, Rater 3 and Rater 9, the

same in each item as in the baseline. In Condition A7, we did not keep these 'reference' raters, but instead made all three raters within Item 1 more lenient and all three raters within Item 3 more severe. This might create estimation problems, and could raise identification issues, and so we wanted to see whether the recovered parameters reflected the complex changes.

3.3 Simulation B

Simulation B consisted of four conditions and was added after completing Simulation A in order to answer questions about parameter recovery in some (simpler) situations. The conditions add to the information given by the conditions in Simulation A. The simulation is called 'B' because it examines variations on what was done in 'A'. The main difference is that, whereas the A conditions examined picking up rater effects 'between' items, the B conditions examine the situation, where the effects vary 'within' items, to see if we can detect rater differences within items.

Table B1, which is after all of the A tables, shows the rater population parameters for four simulation conditions where item criteria parameters were manipulated to indicate more lenient and more severe raters. Table B2 shows the item population parameters for four simulation conditions where item step parameters were varied to indicate more difficult and easier items. Condition B1 performed as a baseline in Simulation B to make comparisons with the other three conditions.

Simulation Condition B1 (Baseline)

In Simulation B, we utilized a baseline with no variation among raters to see the effects on parameter recovery under a relatively 'extreme' scenario. We examined if the HRM-SDT accurately recovered rater and item parameters. Using a baseline with raters and items all the

same could help us to further understand the effects of rater severity and item difficulty within each condition.

In Simulation A, we mainly compared raters 'between' conditions to see if the model could detect the change between each simulation condition and the baseline. However, in reality, there might be no "baseline" to compare with. Suppose there is only one test with three questions, it is important to distinguish which rater is the most severe one and which item is the most difficult one. Accordingly, we examine if this can be done in simple situations.

Concerning item-level parameters, we simulated all item discrimination parameters, a_l, to be 1, and item difficulty parameters, b_{lm}, to be from −2 to 2, which is common found in real-world datasets. Regarding rater-level parameters, all rater discrimination parameters, d_{jl}, were 2 and the rater criteria parameters, c_{jkl}, 1, 3, 5, were at the optimum locations.

Simulation Condition B2 (Rater Severity and Item Difficulty)

In Condition B2, we did the same manipulation on rater parameters and item parameters as Condition A5, where three raters (Rater 1, Rater 4, and Rater 7) became one unit more severe, three raters (Rater 2, Rater 5, and Rater 8) became one unit more lenient, Item 1 became one unit more difficult, and Item 3 became one unit more lenient.

In this condition, we mainly compared rater severity within each item, between three items within Condition B2, and between Condition B2 and Condition B1. For example, within Item 1, Rater 1 should be the most severe, followed by Rater 3, and Rater 2 should be the most lenient. Moreover, in Condition B2, Rater 1 should be estimated to be the same as Rater 4 and Rater 7 if the model perfectly segregates item effects from rater effects. Also, between conditions, Rater 1 should be more severe in Condition B2 than that in Condition B1.

Simulation Condition B3 (Two More Lenient Raters Assigned to One More Difficult Item and Two More Severe Raters Assigned to One Easier Item)

In Condition B3, we followed the same steps to change rater and item parameters as Condition A6 but incorporated a different baseline, where two more lenient raters (Rater 1 and Rater 2) were assigned to one more difficult item (Item 1) and two more severe raters (Rater 7 and Rater 8) were assigned to an easier item (Item 3).

In this condition, rater severity was compared within each item, between three items within Condition B3, and between Condition B3 and Condition B1. For example, for Item 1, Rater 3 should be the most severe, followed by Rater 1, and Rater 2 should be the most lenient. Within Condition B3, Rater 3 should be the same as Rater 6 and Rater 9. Regarding between conditions, Rater 1 should be more lenient in Condition B3 than in Condition B1.

Simulation Condition B4 (Three More Lenient Raters Assigned to One More Difficult Item and Three More Severe Raters Assigned to One Easier Item)

In Condition B4, we followed the same steps to change rater and item parameters as Condition A7 but incorporated a different baseline, where three more lenient raters (Rater 1, Rater 2, and Rater 3) assigned to one more difficult item (Item 1) and three more severe raters (Rater 7, Rater 8, and Rater 9) assigned to an easier item (Item 3).

In this condition, rater severity was compared within each item, between three items within Condition B4, and between Condition B4 and Condition B1. For example, within Item 1, Rater 1 should be the same as Rater 3, and more severe than Rater 2. For Condition B4, the last three raters should be the most severe ones, followed by the second three raters, and the first three raters should be the most lenient ones. Concerning between conditions, the first three raters should be more lenient in Condition B4 than in Condition B1.

3.4 Parameter Recovery

In terms of evaluating the performance of the HRM-SDT, parameter recovery was examined in three ways: with respect to bias, the percentage of bias, and the mean square error (MSE). First, bias represents the difference between the estimated value and the true value of a parameter, where a positive value indicates overestimation, and a negative value indicates underestimation. Second, the absolute percentage of bias is calculated by dividing the bias by the true parameter, taking the absolute value, and then multiplying by 100%. An absolute percentage of bias with a value of 10% or less can be viewed as small, and a value greater than 10% can be regarded as substantial (Curran, West & Finch, 1996; Flora & Curran, 2004). Third, MSE measures bias and variance of estimators and is calculated by averaging the squared differences between the estimated parameters and the true parameters. The MSE evaluates the quality of estimation with values closer to zero, indicating a more accurate estimation.

With respect to the Rasch model, the difference between the estimated values in each simulation condition and the baseline was compared to check whether the Rasch model recognized the correct pattern of change, such as an increase/decrease in rater severity and an increase/decrease in item difficulty.

In summary, this chapter described how the data in each of seven conditions in Simulation A and four conditions in Simulation B was generated, and what criteria was used to evaluate model performance of the HRM-SDT. In next chapter, the results for each of these simulation conditions are explained.

Chapter IV

RESULTS

This chapter presents the results for the conditions of Simulations A and B. For Simulation A, population *rater* parameters for each of the seven conditions are shown in Table A1, and the population *item* parameters are shown in Table A2.

4.1 Results for Simulation A

4.1.1 Results for Condition A1 (Baseline)

As shown in Table A3, the HRM-SDT accurately recovered all rater parameters (first-level) in the between baseline study, including the 27 rater criteria parameters, c_{jkl}, and the nine rater discrimination parameters, d_{jl}. Table A4 shows that the HRM-SDT also precisely recovered all item parameters (second-level), including the three item discrimination parameters, a_l, and the nine item step parameters, b_{lm}. The percent bias for the rater parameters and the item parameters was small, with a value of less than 10%. To sum up, parameter estimation in Condition A1, the baseline, was very good.

4.1.2 Results for Condition A2 (Rater Severity)

In Condition A2, the rater criteria parameters were varied from the baseline so that there were three more severe raters (Rater 1, Rater 4, and Rater 7) and three more lenient raters (Rater 2, Rater 5, and Rater 8). The purpose was to see if the parameter estimates for the HRM-SDT would detect these changes.

As shown in Table A5, all of the rater-level parameters were accurately recovered in Condition A2. For example, the bias for the discrimination parameters, d_{jl}, and criteria parameters, c_{jkl}, was small, with a percent bias of less than 10%. Moreover, the changes in rater

severity from baseline were detected, that is, the criteria for more severe raters (Rater 1, Rater 4, and Rater 7) were all shifted upward by around one unit, indicating a one-unit increase in severity, and the criteria for less severe raters (Rater 2, Rater 5, and Rater 8) were all shifted downward by around one unit, indicating a one-unit decrease in severity. Table A6 presents the recovery of the second-level or item-level parameters in Condition A2. All item discrimination parameters, a_l, and item step parameters, b_{lm}, were accurately recovered, with a percent bias of less than 10%. Generally, all rater parameters and item parameters were precisely recovered, and so the model detected changes in rater severity between Condition A2 and Condition A1.

Table A7 shows the parameter estimates for Condition A1 and Condition A2 for a fit of the Rasch model. As shown in the last column of Table A7, subtracting the estimated parameters in the baseline from those in Condition A2 gives us an idea about how well the Rasch model detected the pattern of changes in rater severity. To be specific, positive values indicate that raters are more severe, and negative values indicate that raters are more lenient. According to the Rasch model, Rater 1, Rater 4, and Rater 7 became more severe; Rater 2, Rater 5, and Rater 8 became more lenient; and Rater 3, Rater 6, and Rater 9 remained the same as the baseline. These results are all consistent with how the data were generated. Thus, the Rasch model performed well with respect to detecting changes in rater severity.

Therefore, both the HRM-SDT and the Rasch identified the three more severe raters and the three more lenient raters. Furthermore, the HRM-SDT precisely recovered the item-level parameters in addition to the rater-level parameters.

4.1.3 Results for Condition A3 (Rater Central Tendency)

Condition A3 involved the rater central-tendency effect, where three raters (Rater 1, Rater 4, and Rater 7) had ‘small’ central tendency, three raters (Rater 2, Rater 5, and Rater 8) had

larger central tendency, and three raters (Rater 3, Rater 6, and Rater 9) remained the same as the baseline.

As shown in Table A8, the HRM-SDT detected the central tendency effect by recovering the shift of the criteria at the endpoints and recovered all rater-level parameters with a percent bias less than 10%. Table A9 shows that, for the item-level parameters, recovery was again excellent with a percent bias of less than 10%.

Table A10 shows the results for fits of the Rasch model. In this case, the Rasch model accurately captured the central tendency effect as well. Subtracting the estimated parameters for the baseline from the estimated parameters in Condition A3, the first and the third parameters for Rater 1, 2, 4, 5, 7, and 8 are shifted down and up (negative and positive), respectively, which reflects central tendency. For example, Rater 1's first criteria parameter, b_{11}, shifted downward by -0.56 units and the third parameter, b_{13}, moved upward by 0.55 units, which implies that Rater 1 tended to avoid the use of Category 1 and Category 4. Furthermore, Raters 2, 5, and 8 have a greater central tendency than Raters 1, 4, and 7, and this result can also be seen in Table A10, in that the deviations are larger.

Consequently, both the HRM-SDT and the Rasch detected central tendency by revealing a small central tendency for three raters and a larger central tendency for three other raters. Additionally, the HRM-SDT accurately recovered both rater-level and item-level parameters, generally with the percentage of bias of less than 10% for all parameters.

4.1.4 Results for Condition A4 (Item Difficulty)

Condition A4 varied item difficulty from the baseline and kept all the rater parameters the same as in the baseline. To be specific, the first item was one unit more difficult; the second item remained the same; and the third item was one unit easier.

Table A11 and Table A12 show that the HRM-SDT precisely recovered the first-level parameters and the second-level parameters in all cases, with an absolute percentage of bias of less than 10%. As shown in Table A12, the difficulty parameters for Item 1 are all shifted around one unit upward, indicating a more difficult item; and the difficulty parameters for Item 3 are all shifted down by around one unit, indicating an easier item. Thus, the model shows the change in item difficulty and separates it from the rater parameters.

In contrast, for the Rasch model, Table A13 shows that the first three raters appear to be more severe, there is no change for the second three raters, and the last three raters all appear to be more lenient. Note, however, that there is actually no change in the rater parameters from baseline. Thus, when the first item was made more difficult, this was reflected in the Rasch model as having more severe raters, and when the last item was made easier, the Rasch model reflected this as having easier raters. This shows that, in this case the Rasch model is confounding item effects with rater effects. This means that, if item difficulty varies, the Rasch model will incorrectly attribute this to changes in the raters.

In summary, the HRM-SDT model gave more accurate results than the Rasch model when item difficulty parameters were varied. If, for example, examinees had lower scores because the items were more difficult, the Rasch model would lead to the conclusion that the raters were more severe and would fail to distinguish between more difficult items and more severe raters.

4.1.5 Results for Condition A5 (Rater Severity and Item Difficulty)

Simulation A5 involved changes in item difficulty and changes in rater severity at the same time. With respect to raters, Raters 1, 4, and 7 are one unit more severe; Raters 2, 5, and 8 are one unit more lenient; and Raters 3, 6, and 9 remain the same as in baseline. In the item part

of the model, the first item is one unit more difficult, the second item remains the same, and the third item is one unit easier than in the baseline.

As shown in Table A14, the HRM-SDT model again accurately recovered all of the rater parameters, where the bias for the discrimination parameters, d_{jl}, and the criteria, c_{jkl}, is consistently small (percent bias of less than 10%). The estimates of the criteria for Rater 1, Rater 4, and Rater 7 are all shifted up by one unit compared to baseline, and so the raters are more severe; the estimates of the criteria for Rater 2, Rater 5, and Rater 7 are all shifted down by one unit compared to baseline, and so the raters are more lenient; and criteria estimates for Rater 3, Rater 6, and Rater 9 are the same as in baseline. All these findings agree with how the rater parameters were specified in the simulation. Table A15 shows, for the item parameters, that the changes in item difficulty were detected, with all the estimates showing a percentage bias of less than 10%, except for b_{13} which had a percent bias of 10.74%. Accordingly, Item 1 was found to be more difficult; Item 2 was found to be about the same; and Item 3 was found to be easier than the baseline. These findings also match the way in which the item parameters were specified.

Table A16 shows that the Rasch model failed to identify the changes when both rater severity and item difficulty parameters were changed. To be specific, regarding the three raters assigned to Item 1, all the values in the last column of Table A16 are positive, which indicates that all three raters (Rater 1, Rater 2 and Rater 3) are more severe than the baseline; however, only Rater 1 was specified to be more severe. Concerning the three raters assigned to Item 2 (same as baseline), Rater 4 was estimated to be more severe, Rater 5 was more lenient, and Rater 6 was about the same as in baseline. In this case, the Rasch model accurately detected the pattern of changes for these three raters. For the three raters assigned to Item 3, all the values in the last column of Table A16 are negative indicating that all three raters (Rater 7, Rater 8 and Rater 9)

were estimated to be more lenient than the baseline; however, only Rater 8 was actually more lenient. Thus, the Rasch Model failed to recover the rater parameter changes when the item difficulty also changed.

In summary, the HRM-SDT segregated rater severity effects and item difficulty effects and gave good estimates of parameters in both levels. However, the Rasch model confounded item effects and rater effects to some extent. The model tended to show raters to be more severe when they were assigned to more difficult items, and more lenient when they were assigned to easier items.

4.1.6 Results for Condition A6 (Two More Lenient Raters Assigned to One More Difficult Item and Two More Severe Raters Assigned to One Easier Item)

Condition A6 incorporated item difficulty effects and rater severity effects at the same time. In particular, two lenient raters (Rater 1 and Rater 2) were associated with a more difficult item (Item 1), and two severe raters (Rater 7 and Rater 8) were associated with an easier item (Item 3).

Table A17 displays the rater-level parameter recovery results. A fit of the HRM-SDT model accurately recovered all discrimination parameters, d_{jl}, for all nine raters, and the criteria, c_{jkl}, for the first six raters, with a percent bias of less than 10%. However, there was some bias for the last three raters. The biases for Rater 7, Rater 8, and Rater 9 were all negative, indicating that the rater criteria parameters were underestimated when more severe raters were assigned to an easier item. Thus, how item difficulty and rater severity were changed affected parameter recovery. For example, Rater 9 was highly discriminating (d_{93}=6) and very severe (c_{913}=3, c_{923}=9, and c_{933}=15). When Rater 9 was assigned to an easy item (b_{31} =−4, b_{32} = −1.5, and b_{33} = 2), the criteria estimates incorrectly indicated that Rater 9 was less severe ($c_{913}{}'$=2.09, $c_{923}{}'$=8.06, and $c_{933}{}'$=14.37) than he or she actually was. Table 18 shows the item-level

recovery. There is some bias in recovering the item step parameters, b_{lm}, especially for Item 3, where b_{31} had a percent bias of 17.18% and b_{32} had a percent bias of 30.82%. In other words, Item 3 was estimated to be more difficult than it should be. Generally, the HRM-SDT separated item and rater effects, but in this case there was some bias when severe raters were assigned to easy items.

Table A19 shows the results for the Rasch Model. As before, the estimates did not capture the changes between the baseline study and Condition A6. As shown in the last column of Table A19, the model showed an increase in most criteria for the first three raters, indicating greater rater severity; however, none of these three raters were simulated to be more severe. The Rasch model suggested that the middle three raters are the same as in baseline, which is consistent with the way we simulated these raters. The Rasch model also did not correctly pick up the rater changes for the last three raters.

To sum up Condition A6, the HRM-SDT model performed well with respect to recovering parameters for the first six raters and the first two items, but had some slight bias for the last three raters and Item 3. This shows that, if the estimation of the item parameters is poor, as for Item 3 in this example, then the associated rater parameters will also be off. The Rasch model, on the other hand, consistently confounded item effects with rater effects; generally, raters were misrecognized as being more severe/lenient when they were associated with more difficult/easier items.

4.1.7 Results for Condition A7 (Three More Lenient Raters Assigned to One More Difficult Item and Three More Severe Raters Assigned to One Easier Item)

In Condition A7, three more lenient raters (Rater 1, Rater 2, and Rater 3) were assigned to grade one more difficult item (Item 1), and three more severe raters (Rater 7, Rater 8, and

Rater 9) were assigned to grade one easier item (Item 3). This condition was included because it seems that it might create problems with separating item effects from rater effects.

Table A20 shows the recovery of the rater-level parameters. The model accurately recovered all parameters associated with the first six raters, where the bias was small with a percent bias less than 10%. However, the HRM-SDT failed to accurately estimate parameters associated with the last three raters, in that the bias was large with a percent bias above 10%. Moreover, the biases associated with Rater 7, Rater 8 and Rater 9 were all negative, indicating that the model tended to underestimate rater severity parameters when three more severe raters were assigned to an easier item. Table A21 shows the item-level parameter recovery results. The bias for the third item was large with a percent bias above 10%. Moreover, Item 3 was simulated to be easier than in the baseline, but the results indicate that it is more difficult. Thus, assigning all three more severe raters to one easier item appears to confound estimation to an extent.

Table A22 shows the results for the Rasch Model. Once again, the Rasch model did not capture the changes between the baseline study and Condition A7. Concerning the three raters assigned to Item 1 (more difficult), almost all parameters were estimated as shifting upwards, indicating more severe raters, however, these three raters were actually simulated to be more lenient. Regarding the three raters assigned to Item 2, with no change from the baseline, the estimates were consistent with the simulation design. Concerning the three raters assigned to Item 3 (easier), all parameters were shifted downward indicating more lenient raters; however, they were simulated to be more severe. Thus, when assigning three more lenient raters to a more difficult item and three more severe raters to an easier item, the Rasch failed to segregate rater effects and item effects.

Overall, HRM-SDT accurately recovered the first six raters and the first two items; however, it failed to recover Item 3 (an easier item) and the parameters for the raters assigned to Item 3. This again shows that if there are problems estimating the item parameters in Level 2, there are also problems estimating the rater parameters in Level 1. The Rasch model again confounded rater effects and item effects.

4.2 Results for Simulation B

Table B1 presents the rater population parameter values for the four conditions of Simulation B, whereas Table B2 presents the item parameter values.

4.2.1 Results for Condition B1 (Baseline)

Condition B1 or the baseline investigated how the HRM-SDT recovered item and rater parameters in an extreme condition with no variation among raters and items.

As shown in Table B3, the HRM-SDT precisely recovered all rater discrimination parameters, d_{jl} and the majority of rater criteria parameters, c_{jkl}, with a percent bias less than 10%, except for the first rater criteria parameters for all nine raters. The first criteria parameters had a bias between -0.13 and -0.25, which was associated with a percent bias between 10% and 25%. Thus, the first criteria tended to be underestimated, although the underestimation is small in absolute value (an estimate of around 0.8 instead of 1.0). Table B4 shows the item-level parameter recovery results. There is some bias in recovering item-level parameters, with a percent bias ranging from 4.21% and 21.94%.

Overall, when there was no variation among raters within an item, the HRM-SDT accurately recovered the majority of rater parameters and had some slight bias in recovering the first rater criteria parameter. It is also interesting to note that the recovery in Condition B1 was

less precise than in Condition A1, which suggests that the HRM-SDT performs better with more variation among raters and items.

4.2.2 Results for Condition B2 (Rater Severity and Item Difficulty)

In Condition B2, rater severity and item difficulty parameters were both manipulated, and in particular, Item 1 was more difficult and Item 3 was easier. In the rater level, three raters (Rater 1, Rater 4, and Rater 7) were more severe, and three raters (Rater 2, Rater 5, and Rater 8) were more lenient.

As shown in Table B5, the HRM-SDT correctly captured the most severe rater and the most lenient one. For example, for Item 1, Rater 1 was estimated to be the most severe, followed by Rater 3, and Rater 2 was estimated to be the most lenient. This order was consistent with the way we simulated these rater parameters. Comparing the estimation results for three groups of raters assigned to three items, the first three raters (Rater 1, Rater 2 and Rater 3) assigned to a more difficult item were estimated to be the most severe, the last three raters (Rater 7, Rater 8, and Rater 9) assigned to an easier item were estimated to be the most lenient, and the middle three raters (Rater 4, Rater 5, and Rater 6) were in-between; however, all these three groups were simulated to be the same as the baseline. In terms of comparing the estimation results in Condition B2 with those in Condition B1, the HRM-SDT clearly recognized Rater 1 and Rater 4 to be more severe and Rater 2, Rater 5, and Rater 8 to be more lenient, but estimated Rater 7 to be roughly the same as the baseline and Rater 9 to be more lenient. In fact, Rater 7 was simulated to be more severe and Rater 9 was simulated to be the same as the baseline.

Table B6 shows the results for the item parameter estimates for Condition B2. In this case the percent bias tends to be larger, particularly for the third item. As shown above, when there

were difficulties estimating the item parameters, there were difficulties estimating the rater parameters.

Table B7 shows the results for a fit of the Rasch model. For the three raters within each item, the Rasch detected the most severe rater and the most lenient rater. Concerning between items comparisons, the Rasch model confounded item effects with rater effects by showing the first three raters to be the most severe and the last three raters to be the most lenient, whereas these three groups of raters should be the same. In other words, the Rasch model tended to estimate a rater as being more severe when a more challenging item was assigned and more lenient when an easier item was assigned. In terms of detecting the changes made from the baseline, the Rasch model did not perform as well as the HRM-SDT model, even though there were problems in both cases. The Rasch model made mistakes for the estimates of Rater 2, Rater 3, Rater 7, and Rater 9.

To sum up, both the HRM-SDT and the Rasch had errors with respect to the rater parameters. Overall, however, with respect to between item comparisons, the HRM-SDT performed better than the Rasch in detecting how rater severity changed from the baseline.

4.2.3 Results for Condition B3 (Two More Lenient Raters Assigned to One More Difficult Item and Two More Severe Raters Assigned to One Easier Item)

In Condition B3, two more lenient raters (Rater 1 and Rater 2) were assigned to a more difficult item (Item 1), and two more severe raters (Rater 7 and Rater 8) were assigned to an easier item (Item 3).

As shown in Table B8, the HRM-SDT correctly recovered the rater parameters within each item. For example, for Item 1, Rater 1 was the most severe, followed by Rater 3, and Rater 2 was the most lenient. This finding agrees with how the rater parameters were simulated. Concerning between items comparison, the HRM-SDT accurately detected that the first three

raters were the most lenient, followed by the second three raters, and the last three raters were the most severe. When comparing the results in Condition B3 and Condition B1, the HRM-SDT accurately detected the changes for all of the raters except for Rater 9. The model confounded item effects, and Rater 9 appears to be more lenient; however, Rater 9 was actually the same as in the baseline.

Table B9 shows the results for the item parameter estimates. The percent bias is larger, though the main problem appears to be with the first two steps of Item 3. This likely accounts for the larger bias found for the rater parameters for the three raters assigned to Item 3.

Table B10 shows the results for the Rasch model. The model successfully detected the sequence of rater severity within each item. However, when comparing raters between items, the Rasch model did not perform as well as the HRM-SDT model. The Rasch model indicated that Rater 1, Rater 4, and Rater 7 to be unchanged (the Condition B3−Condition B1 difference was close to zero), but they were simulated to be different. Thus, the Rasch model again confounded item effects with rater effects to a large extent when item difficulty parameters were varied.

In summary, both models captured different rater effects within an item in Condition B3. However, in terms of comparing raters between items and between conditions, the HRM-SDT model performed more accurately than the Rasch model. The Rasch model confounded rater-level and item-level parameters for the first three raters assigned to Item 1 and the last three raters assigned to Item 3 when an item effect was introduced.

4.2.4 Results for Condition B4 (Three More Lenient Raters Assigned to One More Difficult Item and Three More Severe Raters Assigned to One Easier Item)

In Condition B4, three more lenient raters (Rater 1, Rater 2 and Rater 3) were assigned to one more difficult item (Item 1), and three more severe raters (Rater 7, Rater 8 and Rater 9) were assigned to one easier item (Item 3).

As shown in Table B11, the HRM-SDT model can clearly distinguish raters within an item, but there were some problems. To be specific, the first three raters were correctly estimated to be the most lenient ones; the second three raters were correctly estimated to be unchanged; however, the change for the last three raters was mainly undetected, whereas they were the most severe among three items. Thus, when comparing Condition B4 and Condition B1, the model failed to detect the change for the last three raters; these three raters were estimated to be roughly the same as the baseline, but were simulated to be more severe. Table B12 shows that there were again problems with the item parameter estimates for Item 3, and this is again likely why there were problems with the rater parameters for this item.

As shown in Table B13, the Rasch model correctly detected the most severe rater and the most lenient rater within each item. However, the model failed to distinguish raters between items. Accordingly, except for Rater 2 and Rater 8, all of the other raters were estimated to be roughly the same, which contrasts with the way that the rater parameters were simulated. In terms of comparing Condition B4 and Condition B1, the Rasch model did not detect the changes in Rater 1, Rater 3, Rater 7, and Rater 9.

To sum up, the results for each condition in Simulation A were discussed in Section 4.1 and the results for each condition in Simulation B were explained in Section 4.2. A summary and discussion for these results were illustrated in the following chapter.

Chapter V

SUMMARY AND DISCUSSION

The purpose of this study was to examine to what extent that the HRM-SDT recovered rater and item population parameters, as well as separated rater effects, such as rater severity and rater central tendency, and item effects, such as item difficulty. Since the Rasch model has been one of the most commonly used techniques in language assessment (Fan et al., 2019; McNamara et al., 2019), it was included as a comparison with the HRM-SDT to investigate how much the HRM-SDT could improve upon the Rasch model in separating rater effects and item effects. Simulation A focused on 'between-item' differences, whereas Simulation B focused on 'within-item' differences.

5.1 Summary and Discussion

Simulation A compared how well the HRM-SDT model and the Rasch model recovered and separated rater effects and/or item effects across seven conditions. In the first five simulation conditions, the HRM-SDT accurately recovered parameters, and clearly detected and separated changes in rater severity, rater central tendency, and item difficulty. The last two conditions, A6 and A7, examined recovery in situations where both rater and item parameters were varied in opposite directions. The results showed that the HRM-SDT model accurately recovered most rater and item parameters, except for some parameters where estimation problems for an item (Item 3) appeared. In Condition A6, even though there was some underestimation in the rater level, the HRM-SDT successfully recognized the last three raters to be more severe and Item 3 to be easier than the baseline. However, in Condition A7, the model failed to detect the last three raters were simulated to be more severe and Item 3 to be easier compared to the baseline. Accordingly, recovery was also better in Condition A6 where two raters out of three were

changed within an item than in Condition A7 where all three raters were changed within an item. The problem was probably that there was less variation between raters in Condition A7 than in Condition A6. Additionally, the problem in item-level estimation and the rater-level estimation always affected each other and always came in pairs.

With respect to the Rasch model, it accurately detected the manipulation of rater severity in Condition A2 and rater central tendency in Condition A3, as long as the item-level parameters remained the same as in the baseline. However, when item effects were also introduced, as in Condition A4, A5, A6, and A7, the Rasch model confounded rater effects with item effects. For example, raters assigned to more difficult items appeared to be more severe and raters assigned to easier items appeared to be more lenient, whereas there was actually no change in rater severity. Furthermore, the Rasch model failed to compare raters within both Condition A6 and Condition A7. In both conditions, the last three raters were generated to be the most severe ones, but were estimated to be the most lenient ones among all nine raters.

To sum up, both the HRM-SDT and the Rasch model recovered rater effects when item effects were not involved. When item difficulty was also manipulated, the HRM-SDT had a much better recovery of rater effects than the Rasch model. An implication of this outcome is that one must take care with respect to the design used in any given situation, in that it should provide information about item difficulty along with rater effects, as in the type of design that the HRM-SDT model involves. Furthermore, the diversity among raters could potentially improve the HRM-SDT model performance in terms of recognizing item effects and rater effects. In other words, assigning some severe raters and some lenient raters to one item could increase the estimation precision. instead of training raters to be the same, we would like to have diversity in

using a scoring rubric with some more lenient or some more severe raters assigned to each item, which could potentially improve the HRM-SDT model performance.

Simulation B investigated how the HRM-SDT model and the Rasch model performed in some additional conditions. In Condition B1, where all the rater parameters and all the item parameters were equal, parameter recovery for the HRM-SDT was generally good for both the raters and items, however some problems did appear, such as underestimation of the first criterion. The percent bias at both the rater and item level was also larger, and so the model performed better when there was some variation across raters and items, as in Simulation A (to see this, compare Condition A1 to Condition B1). In Condition B2, where three groups of same raters assigned to different items, the model failed to recognize these three groups to be the same but estimated the first group to be more severe and the last group to be more lenient. That is to say, item-level effect confounded rater-level estimation when raters varied within each item but were the same between items. In Condition B3 and B4, the HRM-SDT precisely estimated the first three raters to be most lenient ones, the second three raters to be unchanged. However, the model undetected the changes of increasing rater severity in the last three raters and recovered these raters to be roughly the same as the baseline. One possible explanation is that the range of item difficulty parameters affects the performance of the HRM-SDT, where the model performs well with very difficult items and medium-difficult items, but had some trouble in recovering extreme easy items.

Regarding the Rasch model in Simulation B, when raters were simulated to be the same between items in Condition B2, the model mistakenly recovered the first three raters to be the most severe ones and the last three raters to be the most lenient ones. When raters were generated to be different between items in Condition B3 and B4, the Rasch model recovered them to be

roughly the same when both rater severity effect and item difficulty effect was manipulated in the opposite directions. That is to say, the Rater model always gives unstable estimation in raters with fluctuation in items.

5.2 Limitations and Future Research

For one limitation, this study simulated data under specific scenarios, such as a four-category scoring rubric, nine raters assigned to grade three items, each of the three items answered by 1000 students, and no missing data involved. For future studies, a different number of scoring categories other than the four-category scoring rubric can be examined to evaluate the model performance. Researchers can also investigate other study designs besides a fully-crossed design and a Balanced Incomplete (BIB) design in this study. It is more likely to have an unbalanced design in a reality, where all raters do not grade the same number of essays. Small sample sizes and missing data are possible directions for future studies.

Second, only rater severity and the central tendency were examined as rater effects, and only item difficulty was investigated as the item effect. However, there might be other effects worth considering. For example, raters might restrict the range of scores. Or, some raters might overuse just the highest score, and some might overuse just the lowest score. Accordingly, besides the three effects investigated in this study, there are more possibilities to explore in future research.

Third, this study generated data for a holistic scoring rubric, which consisted of a single score. Many second language testing uses an analytic scoring rubric to evaluate an essay on different dimensions, such as language use, grammar, and development of ideas (Frey, 2018; Purpura, 2016). A different level should be added in the HRM-SDT model to account for different dimensions in scoring an essay.

Finally, this study was based on the assumption of an equal distance model, where the distance parameters for all four categories were identical for each rater. However, in reality, some raters might have unequal perception distance across different scoring categories. That is to say, raters may have wider distributions for some categories than others. Future research studies can relax this assumption and try unequal distance models (DeCarlo & Zhou, 2020).

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Andrich, D. (1998). Thresholds, teps and rating scale conceptualization. *Rasch Measurement Transactions, 12*(3), 648–649.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Bock, R. D. (1997). The Nominal Categories Model. In Wim J. & Ronald K. *Handbook of Modern Item Response Theory* (pp. 33–49). Springer, New York, NY.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*(3), 303–316.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*(1), 1–18.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27–34.

Brennan, R. L. (2003). Generalizability theory. Journal of Educational Measurement, 40(1), 105-107.

Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modeling, 60*(1), 53–80.

Crick, J. E., & Brennan, R. L. (1984). *GENOVA*: A general purpose analysis of variance system. [*Computer software]. Iowa City, IA: American College Testing*.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Cumming, A. (2009). Assessing academic writing in foreign and second languages. *Language Teaching, 42*(1), 95-107. doi:10.1017/S0261444808005430

De Jong, J., & Linacre, J. M. (1993). Estimation methods, statistical independence, and global fit. *Rasch Measurement Transactions*, *7*(2), 296–297.

DeCarlo, L. (1998). Signal Detection Theory and Generalized Linear Models. *Psychological Methods, 3,* 186–205.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research*, 37(4), 423-451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, *42*(1), 53-76.

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

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).

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.

DeCarlo, L. T., & Zhou, X. (2020). A Latent Class Signal Detection Model for Rater Scoring with Ordered Perceptual Distributions. *Journal of Educational Measurement.* Advance online publication.

Donoghue, J. R., & Hombo, C. M. (2000). A comparison of different model assumptions about rater effects. *A paper presented at the Annual Meeting of the National Council on Measurement in Education* New Orleans, LA.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement* (New edition edition). Peter Lang GmbH.

Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171–191.

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.

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and Scoring of Tests With Multiple-Choice and Constructed-Response Item Types. *Journal of Educational Measurement*, *35*(2), 137–154.

Fan, J., Knoch, U., & Bond, T. (2019). Application of Rasch measurement theory in language assessment: Using measurement to enhance language assessment research and practice. *Papers in Language Testing and Assessment, 8*(2), III.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

Frey, B. (2018). The SAGE encyclopedia of educational research, measurement, and evaluation *(Vols. 1-4). Thousand Oaks,, CA: SAGE Publications, Inc. doi: 10.4135/9781506326139*

Garre, F. G., & Vermunt, J. K. (2006a). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*(1), 43–59.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics, 40*(5), 530–543.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309–334.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff Pub. ; Distributors for North America, Kluwer Boston.

Heine, J.-H., Gebhard, M., Schwab, S., Neumann, P., Gorges, J., & Wild, E. (2018). Testing psychometric properties of the CFT 1-R for students with special educational needs. *Psychological Test and Assessment Modeling, 60*(1). 3-27.

Kim, S., & Moses, T. (2013). Determining When Single Scoring for Constructed-Response Items Is as Effective as Double Scoring in Mixed-Format Licensure Tests. *International Journal of Testing, 13*(4), 314–328.

Kim, Y. (2009). *Combining Constructed Response Items and Multiple Choice Items Using a Hierarchical Rater Model*. Unpublished doctoral dissertation, Columbia University

Kingsbury, F. A. (1922). Analyzing Ratings and Training Raters. *Journal of Personnel Research, 1*, 377–383

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vemont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Kuo, B.-C., Chen, C.-H., Yang, C.-W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology, 36*(6), 1115–1133.

Lai, E. R., Wolfe, E. W., & Vickers, D. H. (2012). *Halo Effects and Analytic Scoring: A Summary of Two Empirical Studies Research Report*. New York: Pearson Research and Innovation Network.

Leckie, G., & Baird, J. A. (2011). Rater Effects on Essay Scoring: A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. *Journal of Educational Measurement*, *48*(4), 399–418.

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

Linacre, J.M., & Wright, B. (1993). *FACETS: Many-facet Rasch analysis (Version 2.68).* Chicago: MESA Press.

Linacre, John M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486–512.

Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing*, *4*(1), 50–65.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.

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

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment*. Oxford University Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Muraki, E. (1992). *A Generalized Partial Credit Model: Application of an EM Algorithm*. ETS Research Report Series, 1992(1).

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

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100.

Naumenko, O. (2014). *Comparison of various polytomous ıtem response theory modeling approaches for task based simulation cpa exam data*. AICPA 2014 Summer Internship Project.

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

Patz, R., Junker, B., & Johnson, M. (2000). 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.

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.

Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory, 4*(4), 171–212.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 124.*

Purpura, J. E. (2016). Second and foreign language assessment. Modern Language Journal, 100 (Supplement 2016), pp. 190-208.

Rahman, A. A., Ahmad, J., Yasin, R. M., & Hanafi, N. M. (2017). Investigating central tendency in competency assessment of design electronic circuit: Analysis using many facet Rasch measurement (MFRM). *Int. J. Inf. Educ. Technol., 7*(7), 525–528.

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.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Schumacker, R. E., & Lomax, R. G. (2012). *A beginner's guide to structural equation modeling*. Routledge Academic New York.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Smith Jr, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, *64*(4), 617–639.

Sturtz, S., Ligges, U., & Gelman, A. E. (2005). *R2WinBUGS: A package for running WinBUGS from R*.

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154.

Vermunt, J. K., & Magidson, J. (2003). *Latent Gold 3.0. B*elmont, MA. URL Http://Www. Statisticalinnovations. Com.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced.* Belmont Massachusetts: Statistical Innovations Inc.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*(1), 108-131.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35–51.

Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465–492.

APPENDIX
TABLES AND FIGURES

Table A1. Population Values of Rater Parameters for Nine Raters in Seven Conditions of Simulation A

			Simulation Conditions						
			A1	A2	A3	A4	A5	A6	A7
CR_1	Rater1	c_{111}	1	2	0.5	1	2	0	0
		c_{121}	3	4	3	3	4	2	2
		c_{131}	5	6	5.5	5	6	4	4
		d_{11}	2	2	2	2	2	2	2
	Rater2	c_{211}	1.5	0.5	0.5	1.5	0.5	-0.5	-0.5
		c_{221}	4.5	3.5	4.5	4.5	3.5	2.5	2.5
		c_{231}	7.5	6.5	8.5	7.5	6.5	5.5	5.5
		d_{21}	3	3	3	3	3	3	3
	Rater3	c_{311}	2	2	2	2	2	2	1
		c_{321}	6	6	6	6	6	6	5
		c_{331}	10	10	10	10	10	10	9
		d_{31}	4	4	4	4	4	4	4
CR_2	Rater4	c_{412}	1.5	2.5	1	1.5	2.5	1.5	1.5
		c_{422}	4.5	5.5	4.5	4.5	5.5	4.5	4.5
		c_{432}	7.5	8.5	8	7.5	8.5	7.5	7.5
		d_{42}	3	3	3	3	3	3	3
	Rater5	c_{512}	2	1	1	2	1	2	2
		c_{522}	6	5	6	6	5	6	6
		c_{532}	10	9	11	10	9	10	10
		d_{52}	4	4	4	4	4	4	4
	Rater6	c_{612}	2.5	2.5	2.5	2.5	2.5	2.5	2.5
		c_{622}	7.5	7.5	7.5	7.5	7.5	7.5	7.5
		c_{632}	12.5	12.5	12.5	12.5	12.5	12.5	12.5
		d_{62}	5	5	5	5	5	5	5
CR_3	Rater7	c_{713}	2	3	1.5	2	3	3	3
		c_{723}	6	7	6	6	7	7	7
		c_{733}	10	11	10.5	10	11	11	11
		d_{73}	4	4	4	4	4	4	4
	Rater8	c_{813}	2.5	1.5	1.5	2.5	1.5	4.5	4.5
		c_{823}	7.5	6.5	7.5	7.5	6.5	9.5	9.5
		c_{833}	12.5	11.5	13.5	12.5	11.5	14.5	14.5
		d_{83}	5	5	5	5	5	5	5
	Rater9	c_{913}	3	3	3	3	3	3	4
		c_{923}	9	9	9	9	9	9	10
		c_{933}	15	15	15	15	15	15	16
		d_{93}	6	6	6	6	6	6	6

Table A2. Population Values of Item Parameters for Three CR Items in Seven Conditions of Simulation A

		Simulation Conditions						
		A1	A2	A3	A4	A5	A6	A7
CR_1	b_{11}	-1	-1	-1	0	0	0	0
	b_{12}	0	0	0	1	1	1	1
	b_{13}	1	1	1	2	2	2	2
	a_1	0.5	0.5	0.5	0.5	0.5	0.5	0.5
CR_2	b_{21}	-2	-2	-2	-2	-2	-2	-2
	b_{22}	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	b_{23}	2	2	2	2	2	2	2
	a_2	1	1	1	1	1	1	1
CR_3	b_{31}	-3	-3	-3	-4	-4	-4	-4
	b_{32}	-0.5	-0.5	-0.5	-1.5	-1.5	-1.5	-1.5
	b_{33}	3	3	3	2	2	2	2
	a_3	2	2	2	2	2	2	2

Table A3. HRM-SDT Results for Rater Parameters in Condition A1 of Simulation A

			Value	Estimate	Bias	%Bias	MSE
CR_1	Rater1	c_{111}	1	0.99	-0.01	1.30	0.03
		c_{121}	3	2.99	-0.01	0.48	0.05
		c_{131}	5	4.99	-0.01	0.10	0.07
		d_{11}	2	1.99	-0.01	0.38	0.02
	Rater2	c_{211}	1.5	1.46	-0.04	2.52	0.09
		c_{221}	4.5	4.45	-0.05	1.11	0.16
		c_{231}	7.5	7.46	-0.04	0.51	0.24
		d_{21}	3	2.96	-0.04	1.18	0.05
	Rater3	c_{311}	2	1.99	-0.01	0.60	0.16
		c_{321}	6	6.04	0.04	0.59	0.24
		c_{331}	10	10.16	0.16	1.56	0.51
		d_{31}	4	4.04	0.04	1.04	0.09
CR_2	Rater4	c_{412}	1.5	1.51	0.01	0.57	0.03
		c_{422}	4.5	4.52	0.02	0.38	0.06
		c_{432}	7.5	7.56	0.06	0.74	0.11
		d_{42}	3	3.02	0.02	0.62	0.02
	Rater5	c_{512}	2	2.00	0.00	0.19	0.06
		c_{522}	6	5.98	-0.02	0.26	0.10
		c_{532}	10	10.02	0.02	0.23	0.24
		d_{52}	4	4.01	0.01	0.21	0.05
	Rater6	c_{612}	2.5	2.54	0.04	1.48	0.15
		c_{622}	7.5	7.55	0.05	0.70	0.29
		c_{632}	12.5	12.62	0.12	0.99	0.54
		d_{62}	5	5.05	0.05	0.97	0.11
CR_3	Rater7	c_{713}	2	1.99	-0.01	0.35	0.05
		c_{723}	6	6.01	0.01	0.18	0.09
		c_{733}	10	10.05	0.05	0.49	0.14
		d_{73}	4	4.02	0.02	0.40	0.03
	Rater8	c_{813}	2.5	2.46	-0.04	1.59	0.08
		c_{823}	7.5	7.45	-0.05	0.70	0.15
		c_{833}	12.5	12.46	-0.04	0.36	0.21
		d_{83}	5	4.97	-0.03	0.69	0.05
	Rater9	c_{913}	3	3.02	0.02	0.51	0.16
		c_{923}	9	9.02	0.02	0.20	0.36
		c_{933}	15	15.09	0.09	0.57	0.58
		d_{93}	6	6.03	0.03	0.56	0.12

Table A4. HRM-SDT Results for Item Parameters in Condition A1 in Simulation A

		Value	Estimate	Bias	%Bias	MSE
CR_1	b_{11}	-1	-0.99	0.01	0.89	0.03
	b_{12}	0	0.00	0.00	—	0.02
	b_{13}	1	0.99	-0.01	0.86	0.04
	a_1	0.5	0.51	0.01	2.13	0.00
CR_2	b_{21}	-2	-2.02	-0.02	1.19	0.04
	b_{22}	0.5	0.51	0.01	1.91	0.01
	b_{23}	2	2.08	0.08	4.05	0.07
	a_2	1	1.05	0.05	5.30	0.03
CR_3	b_{31}	-3	-2.89	0.11	3.78	0.29
	b_{32}	-0.5	-0.49	0.01	1.58	0.02
	b_{33}	3	2.93	-0.07	2.18	0.24
	a_3	2	1.91	-0.09	4.68	0.18

Table A5. HRM-SDT Results for Rater Parameters in Condition A2 of Simulation A (Rater Severity)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1	Rater 1 (one unit more severe)	c_{111}	2	1	2.00	0.00	0.03	0.05
		c_{121}	4	3	3.99	-0.01	0.14	0.07
		c_{131}	6	5	6.01	0.01	0.17	0.10
		d_{11}	2	2	1.99	-0.01	0.53	0.02
	Rater 2 (one unit more lenient)	c_{211}	0.5	1.5	0.51	0.01	1.78	0.05
		c_{221}	3.5	4.5	3.50	0.00	0.07	0.09
		c_{231}	6.5	7.5	6.51	0.01	0.11	0.19
		d_{21}	3	3	2.99	-0.01	0.38	0.05
	Rater 3	c_{311}	2	2	2.02	0.02	0.76	0.16
		c_{321}	6	6	6.07	0.07	1.09	0.32
		c_{331}	10	10	10.10	0.10	1.02	0.65
		d_{31}	4	4	4.01	0.01	0.37	0.12
CR_2	Rater 4 (one unit more severe)	c_{412}	2.5	1.5	2.51	0.01	0.25	0.04
		c_{422}	5.5	4.5	5.54	0.04	0.67	0.10
		c_{432}	8.5	7.5	8.56	0.06	0.69	0.17
		d_{42}	3	3	3.02	0.02	0.54	0.03
	Rater 5 (one unit more lenient)	c_{512}	1	2	0.98	-0.02	2.00	0.05
		c_{522}	5	6	4.98	-0.02	0.30	0.08
		c_{532}	9	10	9.01	0.01	0.06	0.25
		d_{52}	4	4	4.00	0.00	0.01	0.06
	Rater 6	c_{612}	2.5	2.5	2.49	-0.01	0.30	0.16
		c_{622}	7.5	7.5	7.49	-0.01	0.13	0.26
		c_{632}	12.5	12.5	12.55	0.05	0.42	0.56
		d_{62}	5	5	5.00	0.00	0.01	0.12
CR_3	Rater 7 (one unit more severe)	c_{713}	3	2	3.04	0.04	1.21	0.07
		c_{723}	7	6	7.06	0.06	0.93	0.15
		c_{733}	11	10	11.12	0.12	1.10	0.28
		d_{73}	4	4	4.04	0.04	0.99	0.04
	Rater 8 (one unit more lenient)	c_{813}	1.5	2.5	1.50	0.00	0.15	0.08
		c_{823}	6.5	7.5	6.55	0.05	0.83	0.15
		c_{833}	11.5	12.5	11.58	0.08	0.67	0.37
		d_{83}	5	5	5.04	0.04	0.73	0.09
	Rater 9	c_{913}	3	3	2.99	-0.01	0.30	0.15
		c_{923}	9	9	9.03	0.03	0.32	0.29
		c_{933}	15	15	15.03	0.03	0.19	0.49
		d_{93}	6	6	6.01	0.01	0.12	0.10

Table A6. HRM-SDT Results for Item Parameters in Condition A2 of Simulation A (Rater Severity)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1	b_{11}	-1	-1	-1.02	-0.02	2.49	0.04
	b_{12}	0	0	-0.02	-0.02	—	0.01
	b_{13}	1	1	1.00	0.00	0.00	0.03
	a_1	0.5	0.5	0.52	0.02	3.47	0.01
CR_2	b_{21}	-2	-2	-2.06	-0.06	2.98	0.06
	b_{22}	0.5	0.5	0.51	0.01	1.71	0.01
	b_{23}	2	2	2.06	0.06	3.08	0.07
	a_2	1	1	1.06	0.06	6.33	0.03
CR_3	b_{31}	-3	-3	-2.83	0.17	5.79	0.24
	b_{32}	-0.5	-0.5	-0.48	0.02	3.56	0.01
	b_{33}	3	3	2.84	-0.16	5.46	0.19
	a_3	2	2	1.84	-0.16	7.90	0.15

Table A7. Rasch Model Results for Rater Parameters in Condition A2 of Simulation A (Rater Severity)

			Condition A2	Condition A1 (Baseline)	Condition A2 – Condition A1
CR_1	Rater 1 (one unit more severe)	b_{11}	-0.15	-0.76	0.60
		b_{12}	0.46	0.00	0.46
		b_{13}	1.37	0.76	0.60
	Rater 2 (one unit more lenient)	b_{21}	-1.48	-1.02	-0.46
		b_{22}	-0.35	-0.01	-0.35
		b_{23}	0.54	1.04	-0.51
	Rater 3	b_{31}	-1.16	-1.16	0.00
		b_{32}	0.00	-0.02	0.01
		b_{33}	1.16	1.18	-0.03
CR_2	Rater 4 (one unit more severe)	b_{41}	-0.60	-1.22	0.62
		b_{42}	0.71	0.30	0.41
		b_{43}	1.87	1.46	0.41
	Rater 5 (one unit more lenient)	b_{51}	-1.88	-1.46	-0.42
		b_{52}	0.02	0.36	-0.35
		b_{53}	1.22	1.65	-0.43
	Rater 6	b_{61}	-1.63	-1.63	-0.01
		b_{62}	0.41	0.41	0.00
		b_{63}	1.73	1.77	-0.04
CR_3	Rater 7 (one unit more severe)	b_{71}	-1.11	-1.50	0.40
		b_{72}	0.04	-0.29	0.34
		b_{73}	1.87	1.50	0.37
	Rater 8 (one unit more lenient)	b_{81}	-1.84	-1.61	-0.23
		b_{82}	-0.49	-0.32	-0.17
		b_{83}	1.22	1.67	-0.45
	Rater 9	b_{91}	-1.67	-1.67	0.00
		b_{92}	-0.34	-0.35	0.02
		b_{93}	1.73	1.77	-0.04

Table A8. HRM-SDT Results for Rater Parameters in Condition A3 of Simulation A (Rater Central Tendency)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1	Rater 1 (slight central tendency)	c_{111}	0.5	1	0.50	0.00	0.95	0.03
		c_{121}	3	3	3.00	0.00	0.08	0.05
		c_{131}	5.5	5	5.52	0.02	0.35	0.08
		d_{11}	2	2	2.01	0.01	0.28	0.02
	Rater 2 (obvious central tendency)	c_{211}	0.5	1.5	0.49	-0.01	1.19	0.05
		c_{221}	4.5	4.5	4.57	0.07	1.60	0.13
		c_{231}	8.5	7.5	8.59	0.09	1.08	0.26
		d_{21}	3	3	3.05	0.05	1.53	0.04
	Rater 3	c_{311}	2	2	1.99	-0.01	0.70	0.11
		c_{321}	6	6	6.03	0.03	0.54	0.18
		c_{331}	10	10	10.05	0.05	0.46	0.37
		d_{31}	4	4	4.03	0.03	0.65	0.09
CR_2	Rater 4 (slight central tendency)	c_{412}	1	1.5	1.03	0.03	2.57	0.03
		c_{422}	4.5	4.5	4.54	0.04	0.80	0.06
		c_{432}	8	7.5	8.07	0.07	0.91	0.13
		d_{42}	3	3	3.04	0.04	1.21	0.02
	Rater 5 (obvious central tendency)	c_{512}	1	2	1.00	0.00	0.44	0.05
		c_{522}	6	6	6.01	0.01	0.11	0.09
		c_{532}	11	10	11.03	0.03	0.30	0.33
		d_{52}	4	4	4.01	0.01	0.32	0.04
	Rater 6	c_{612}	2.5	2.5	2.50	0.00	0.09	0.10
		c_{622}	7.5	7.5	7.52	0.02	0.25	0.21
		c_{632}	12.5	12.5	12.55	0.05	0.44	0.50
		d_{62}	5	5	5.03	0.03	0.56	0.10
CR_3	Rater 7 (slight central tendency)	c_{713}	1.5	2	1.45	-0.05	3.05	0.05
		c_{723}	6	6	5.97	-0.03	0.58	0.08
		c_{733}	10.5	10	10.49	-0.01	0.14	0.17
		d_{73}	4	4	3.99	-0.01	0.35	0.03
	Rater 8 (obvious central tendency)	c_{813}	1.5	2.5	1.46	-0.04	2.72	0.07
		c_{823}	7.5	7.5	7.48	-0.02	0.29	0.13
		c_{833}	13.5	12.5	13.48	-0.02	0.15	0.30
		d_{83}	5	5	5.00	0.00	0.08	0.04
	Rater 9	c_{913}	3	3	2.99	-0.01	0.20	0.13
		c_{923}	9	9	9.02	0.02	0.19	0.36
		c_{933}	15	15	15.10	0.10	0.67	0.57
		d_{93}	6	6	6.04	0.04	0.73	0.12

Table A9. HRM-SDT Results for Item Parameters in Condition A3 of Simulation A (Rater Central Tendency)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1	b_{11}	-1	-1	-1.01	-0.01	0.82	0.04
	b_{12}	0	0	-0.01	-0.01	—	0.02
	b_{13}	1	1	1.02	0.02	2.36	0.05
	a_1	0.5	0.5	0.52	0.02	3.84	0.01
CR_2	b_{21}	-2	-2	-2.07	-0.07	3.42	0.05
	b_{22}	0.5	0.5	0.51	0.01	2.09	0.01
	b_{23}	2	2	2.07	0.07	3.37	0.08
	a_2	1	1	1.05	0.05	5.25	0.03
CR_3	b_{31}	-3	-3	-2.82	0.18	5.87	0.36
	b_{32}	-0.5	-0.5	-0.48	0.02	4.36	0.02
	b_{33}	3	3	2.81	-0.19	6.21	0.27
	a_3	2	2	1.84	-0.16	7.90	0.23

Table A10. Rasch Model Results for Rater Parameters in Condition A3 of Simulation A (Rater Central Tendency)

			Condition A3	Condition A1 (Baseline)	Condition A3 – Condition A1
CR_1	Rater1 (slight central tendency)	b_{11}	-1.31	-0.76	-0.56
		b_{12}	-0.01	0.00	-0.01
		b_{13}	1.31	0.76	0.55
	Rater2 (obvious central tendency)	b_{21}	-1.81	-1.02	-0.79
		b_{22}	0.00	-0.01	0.01
		b_{23}	1.77	1.04	0.73
	Rater3	b_{31}	-1.23	-1.16	-0.07
		b_{32}	0.00	-0.02	0.01
		b_{33}	1.21	1.18	0.03
CR_2	Rater4 (slight central tendency)	b_{41}	-1.69	-1.22	-0.47
		b_{42}	0.30	0.30	0.00
		b_{43}	1.87	1.46	0.41
	Rater5 (obvious central tendency)	b_{51}	-2.11	-1.46	-0.65
		b_{52}	0.37	0.36	0.01
		b_{53}	2.23	1.65	0.58
	Rater 6	b_{61}	-1.73	-1.63	-0.10
		b_{62}	0.42	0.41	0.01
		b_{63}	1.81	1.77	0.05
CR_3	Rater7 (slight central tendency)	b_{71}	-1.85	-1.50	-0.35
		b_{72}	-0.29	-0.29	0.00
		b_{73}	1.81	1.50	0.31
	Rater8 (obvious central tendency)	b_{81}	-2.05	-1.61	-0.44
		b_{82}	-0.32	-0.32	0.00
		b_{83}	2.06	1.67	0.39
	Rater9	b_{91}	-1.75	-1.67	-0.08
		b_{92}	-0.36	-0.35	0.00
		b_{93}	1.79	1.77	0.02

Table A11. HRM-SDT Results for Rater Parameters in Condition A4 of Simulation A (Item Difficulty)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	Rater1	c_{111}	1	1	1.03	0.03	2.51	0.02
		c_{121}	3	3	3.05	0.05	1.54	0.04
		c_{131}	5	5	5.07	0.07	1.49	0.07
		d_{11}	2	2	1.96	-0.04	2.10	0.03
	Rater2	c_{211}	1.5	1.5	1.53	0.03	1.95	0.07
		c_{221}	4.5	4.5	4.54	0.04	0.92	0.15
		c_{231}	7.5	7.5	7.55	0.05	0.68	0.29
		d_{21}	3	3	2.91	-0.09	3.12	0.06
	Rater3	c_{311}	2	2	2.04	0.04	2.19	0.12
		c_{321}	6	6	6.17	0.17	2.87	0.35
		c_{331}	10	10	10.21	0.21	2.12	0.65
		d_{31}	4	4	3.94	-0.06	1.58	0.11
CR_2	Rater4	c_{412}	1.5	1.5	1.50	0.00	0.26	0.04
		c_{422}	4.5	4.5	4.49	-0.01	0.12	0.07
		c_{432}	7.5	7.5	7.50	0.00	0.01	0.13
		d_{42}	3	3	3.00	0.00	0.09	0.03
	Rater5	c_{512}	2	2	2.04	0.04	2.19	0.06
		c_{522}	6	6	6.04	0.04	0.60	0.11
		c_{532}	10	10	10.07	0.07	0.70	0.24
		d_{52}	4	4	4.01	0.01	0.28	0.05
	Rater6	c_{612}	2.5	2.5	2.53	0.03	1.40	0.13
		c_{622}	7.5	7.5	7.54	0.04	0.53	0.20
		c_{632}	12.5	12.5	12.58	0.08	0.65	0.44
		d_{62}	5	5	5.02	0.02	0.42	0.09
CR_3 (one unit easier)	Rater7	c_{713}	2	2	1.95	-0.05	2.65	0.18
		c_{723}	6	6	5.99	-0.01	0.10	0.26
		c_{733}	10	10	10.02	0.02	0.19	0.34
		d_{73}	4	4	4.02	0.02	0.46	0.03
	Rater8	c_{813}	2.5	2.5	2.40	-0.10	4.11	0.29
		c_{823}	7.5	7.5	7.42	-0.08	1.08	0.45
		c_{833}	12.5	12.5	12.47	-0.03	0.22	0.65
		d_{83}	5	5	5.00	0.00	0.06	0.07
	Rater9	c_{913}	3	3	2.79	-0.21	7.13	0.50
		c_{923}	9	9	8.83	-0.17	1.89	0.83
		c_{933}	15	15	14.87	-0.13	0.88	1.03
		d_{93}	6	6	5.96	-0.04	0.69	0.14

Table A12. HRM-SDT Results for Item Parameters in Condition A4 of Simulation A (Item Difficulty)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	b_{11}	0	-1	0.01	0.01	—	0.02
	b_{12}	1	0	0.94	-0.06	6.34	0.04
	b_{13}	2	1	1.81	-0.19	9.67	0.17
	a_1	0.5	0.5	0.51	0.01	1.41	0.01
CR_2	b_{21}	-2	-2	-2.08	-0.08	4.07	0.06
	b_{22}	0.5	0.5	0.50	0.00	0.81	0.01
	b_{23}	2	2	2.05	0.05	2.69	0.07
	a_2	1	1	1.06	0.06	6.01	0.03
CR_3 (one unit easier)	b_{31}	-4	-3	-3.82	0.18	4.41	0.49
	b_{32}	-1.5	-0.5	-1.45	0.05	3.22	0.15
	b_{33}	2	3	1.97	-0.03	1.73	0.10
	a_3	2	2	1.93	-0.07	3.40	0.15

Table A13. Rasch Model Results for Rater Parameters in Condition A4 of Simulation A (Item Difficulty)

			Condition A4	Condition A1 (Baseline)	Condition A4-Condition A1
CR_1 (one unit more difficult)	Rater1	b_{11}	0.05	-0.76	0.81
		b_{12}	0.85	0.00	0.86
		b_{13}	1.65	0.76	0.89
	Rater2	b_{21}	-0.04	-1.02	0.98
		b_{22}	1.04	-0.01	1.04
		b_{23}	2.05	1.04	1.01
	Rater3	b_{31}	-0.09	-1.16	1.07
		b_{32}	1.15	-0.02	1.16
		b_{33}	2.26	1.18	1.07
CR_2	Rater4	b_{41}	-1.25	-1.22	-0.03
		b_{42}	0.31	0.30	0.01
		b_{43}	1.44	1.46	-0.02
	Rater5	b_{51}	-1.49	-1.46	-0.03
		b_{52}	0.38	0.36	0.01
		b_{53}	1.64	1.65	0.00
	Rater6	b_{61}	-1.67	-1.63	-0.05
		b_{62}	0.42	0.41	0.01
		b_{63}	1.75	1.77	-0.02
CR_3 (one unit easier)	Rater7	b_{71}	-2.07	-1.50	-0.57
		b_{72}	-0.82	-0.29	-0.52
		b_{73}	0.98	1.50	-0.52
	Rater8	b_{81}	-2.20	-1.61	-0.59
		b_{82}	-0.91	-0.32	-0.59
		b_{83}	1.09	1.67	-0.58
	Rater9	b_{91}	-2.30	-1.67	-0.64
		b_{92}	-0.96	-0.35	-0.60
		b_{93}	1.17	1.77	-0.61

Table A14. HRM-SDT Results for Rater Parameters in Condition A5 of Simulation A (Rater Severity and Item Difficulty)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	Rater 1 (one unit more severe)	c_{111}	2	1	2.02	0.02	1.16	0.03
		c_{121}	4	3	4.03	0.03	0.75	0.05
		c_{131}	6	5	6.02	0.02	0.38	0.11
		d_{11}	2	2	1.94	-0.06	2.92	0.04
	Rater 2 (one unit more lenient)	c_{211}	0.5	1.5	0.51	0.01	2.83	0.02
		c_{221}	3.5	4.5	3.51	0.01	0.19	0.06
		c_{231}	6.5	7.5	6.55	0.05	0.71	0.19
		d_{21}	3	3	2.89	-0.11	3.62	0.09
	Rater 3	c_{311}	2	2	2.05	0.05	2.31	0.09
		c_{321}	6	6	6.20	0.20	3.29	0.42
		c_{331}	10	10	10.31	0.31	3.06	0.90
		d_{31}	4	4	3.94	-0.06	1.45	0.13
CR_2	Rater 4 (one unit more severe)	c_{412}	2.5	1.5	2.48	-0.02	0.74	0.05
		c_{422}	5.5	4.5	5.48	-0.02	0.30	0.13
		c_{432}	8.5	7.5	8.48	-0.02	0.22	0.24
		d_{42}	3	3	2.98	-0.02	0.63	0.04
	Rater 5 (one unit more lenient)	c_{512}	1	2	0.96	-0.04	4.25	0.08
		c_{522}	5	6	4.97	-0.03	0.54	0.09
		c_{532}	9	10	8.96	-0.04	0.40	0.26
		d_{52}	4	4	3.97	-0.03	0.84	0.06
	Rater 6	c_{612}	2.5	2.5	2.50	0.00	0.04	0.14
		c_{622}	7.5	7.5	7.48	-0.02	0.24	0.24
		c_{632}	12.5	12.5	12.53	0.03	0.26	0.58
		d_{62}	5	5	4.98	-0.02	0.50	0.12
CR_3 (one unit easier)	Rater 7 (one unit more severe)	c_{713}	3	2	2.92	-0.08	2.66	0.29
		c_{723}	7	6	6.97	-0.03	0.43	0.41
		c_{733}	11	10	10.98	-0.02	0.18	0.50
		d_{73}	4	4	4.01	0.01	0.37	0.05
	Rater 8 (one unit more lenient)	c_{813}	1.5	2.5	1.41	-0.09	5.88	0.30
		c_{823}	6.5	7.5	6.43	-0.07	1.02	0.58
		c_{833}	11.5	12.5	11.50	0.00	0.02	0.57
		d_{83}	5	5	5.03	0.03	0.57	0.09
	Rater 9	c_{913}	3	3	2.93	-0.07	2.33	0.67
		c_{923}	9	9	8.97	-0.03	0.39	1.12
		c_{933}	15	15	15.03	0.03	0.21	1.14
		d_{93}	6	6	6.04	0.04	0.64	0.16

Table A15. HRM-SDT Results for Item Parameters in Condition A5 of Simulation A (Rater Severity and Item Difficulty)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	b_{11}	0	-1	0.01	0.01	—	0.02
	b_{12}	1	0	0.92	-0.08	7.66	0.04
	b_{13}	2	1	1.79	-0.21	10.74	0.28
	a_1	0.5	0.5	0.49	-0.01	1.48	0.01
CR_2	b_{21}	-2	-2	-2.09	-0.09	4.54	0.06
	b_{22}	0.5	0.5	0.51	0.01	2.08	0.01
	b_{23}	2	2	2.05	0.05	2.25	0.07
	a_2	1	1	1.05	0.05	5.47	0.03
CR_3 (one unit easier)	b_{31}	-4	-3	-3.76	0.24	5.92	0.65
	b_{32}	-1.5	-0.5	-1.40	0.10	6.52	0.17
	b_{33}	2	3	1.93	-0.07	3.25	0.20
	a_3	2	2	1.87	-0.13	6.74	0.17

Table A16. Rasch Model Results for Rater Parameters in Condition A5 of Simulation A (Rater Severity and Item Difficulty)

			Condition A5	Condition A1 (Baseline)	Condition A5-Condition A1
CR_1 (one unit more difficult)	Rater 1 (one unit more severe)	b_{11}	0.77	-0.76	1.52
		b_{12}	1.28	0.00	1.29
		b_{13}	2.08	0.76	1.31
	Rater 2 (one unit more lenient)	b_{21}	-0.62	-1.02	0.41
		b_{22}	0.64	-0.01	0.64
		b_{23}	1.60	1.04	0.56
	Rater 3	b_{31}	-0.09	-1.16	1.07
		b_{32}	1.12	-0.02	1.13
		b_{33}	2.24	1.18	1.06
CR_2	Rater 4 (one unit more severe)	b_{41}	-0.60	-1.22	0.61
		b_{42}	0.72	0.30	0.42
		b_{43}	1.86	1.46	0.40
	Rater 5 (one unit more lenient)	b_{51}	-1.92	-1.46	-0.46
		b_{52}	0.03	0.36	-0.33
		b_{53}	1.23	1.65	-0.42
	Rater 6	b_{61}	-1.66	-1.63	-0.03
		b_{62}	0.42	0.41	0.01
		b_{63}	1.74	1.77	-0.03
CR_3 (one unit easier)	Rater 7 (one unit more severe)	b_{71}	-1.68	-1.50	-0.17
		b_{72}	-0.41	-0.29	-0.12
		b_{73}	1.33	1.50	-0.18
	Rater 8 (one unit more lenient)	b_{81}	-2.43	-1.61	-0.83
		b_{82}	-1.08	-0.32	-0.76
		b_{83}	0.72	1.67	-0.95
	Rater 9	b_{91}	-2.27	-1.67	-0.61
		b_{92}	-0.93	-0.35	-0.58
		b_{93}	1.14	1.77	-0.63

Table A17. HRM-SDT Results for Rater Parameters in Condition A6 of Simulation A (Assign Two More Lenient Raters to One More Difficult Item and Two More Severe Raters to One Easier Item)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	Rater 1 (one unit more lenient)	c_{111}	0	1	0.01	0.01	—	0.01
		c_{121}	2	3	2.01	0.01	0.66	0.02
		c_{131}	4	5	4.03	0.03	0.83	0.05
		d_{11}	2	2	1.89	-0.11	5.70	0.05
	Rater 2 (two units more lenient)	c_{211}	-0.5	1.5	-0.49	0.01	2.47	0.02
		c_{221}	2.5	4.5	2.52	0.02	0.67	0.05
		c_{231}	5.5	7.5	5.58	0.08	1.47	0.15
		d_{21}	3	3	2.87	-0.13	4.48	0.13
	Rater 3	c_{311}	2	2	2.05	0.05	2.41	0.09
		c_{321}	6	6	6.22	0.22	3.61	0.47
		c_{331}	10	10	10.31	0.31	3.06	0.98
		d_{31}	4	4	3.84	-0.16	4.04	0.15
CR_2	Rater 4	c_{412}	1.5	1.5	1.52	0.02	1.20	0.04
		c_{422}	4.5	4.5	4.51	0.01	0.26	0.05
		c_{432}	7.5	7.5	7.54	0.04	0.51	0.09
		d_{42}	3	3	3.01	0.01	0.45	0.02
	Rater 5	c_{512}	2	2	1.97	-0.03	1.55	0.06
		c_{522}	6	6	6.00	0.00	0.07	0.12
		c_{532}	10	10	10.01	0.01	0.12	0.27
		d_{52}	4	4	3.99	-0.01	0.14	0.05
	Rater 6	c_{612}	2.5	2.5	2.46	-0.04	1.59	0.14
		c_{622}	7.5	7.5	7.48	-0.02	0.31	0.23
		c_{632}	12.5	12.5	12.46	-0.04	0.28	0.45
		d_{62}	5	5	4.97	-0.03	0.52	0.10
CR_3 (one unit easier)	Rater 7 (one unit more severe)	c_{713}	3	2	2.36	-0.64	21.42	2.07
		c_{723}	7	6	6.32	-0.68	9.70	2.66
		c_{733}	11	10	10.41	-0.59	5.33	2.00
		d_{73}	4	4	3.88	-0.12	2.95	0.14
	Rater 8 (two units more severe)	c_{813}	4.5	2.5	3.69	-0.81	17.89	3.56
		c_{823}	9.5	7.5	8.65	-0.85	8.91	4.70
		c_{833}	14.5	12.5	13.79	-0.71	4.88	3.86
		d_{83}	5	5	4.85	-0.15	3.09	0.32
	Rater 9	c_{913}	3	3	2.09	-0.91	30.38	3.60
		c_{923}	9	9	8.06	-0.94	10.49	5.28
		c_{933}	15	15	14.37	-0.63	4.22	3.28
		d_{93}	6	6	5.91	-0.09	1.56	0.23

Table A18. HRM-SDT Results for Item Parameters in Condition A6 of Simulation A (Assign Two More Lenient Raters to One More Difficult Item and Two More Severe Raters to One Easier Item)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	b_{11}	0	-1	0.03	0.03	—	0.02
	b_{12}	1	0	0.88	-0.12	12.38	0.08
	b_{13}	2	1	1.62	-0.38	18.94	0.49
	a_1	0.5	0.5	0.48	-0.02	3.81	0.01
CR_2	b_{21}	-2	-2	-2.06	-0.06	3.24	0.05
	b_{22}	0.5	0.5	0.52	0.02	3.55	0.01
	b_{23}	2	2	2.06	0.06	3.05	0.08
	a_2	1	1	1.07	0.07	6.59	0.04
CR_3 (one unit easier)	b_{31}	-4	-3	-3.31	0.69	17.18	1.53
	b_{32}	-1.5	-0.5	-1.04	0.46	30.82	1.27
	b_{33}	2	3	1.95	-0.05	2.65	0.48
	a_3	2	2	1.84	-0.16	7.80	0.19

Table A19. Rasch Model Results for Rater Parameters in Condition A6 of Simulation A (Assign Two More Lenient Raters to One More Difficult Item and Two More Severe Raters to One Easier Item)

			Condition A6	Condition A1 (Baseline)	Condition A6 - Condition A1
CR_1 (one unit more difficult)	Rater 1 (one unit more lenient)	b_{11}	-0.65	-0.76	0.11
		b_{12}	0.31	0.00	0.31
		b_{13}	1.02	0.76	0.26
	Rater 2 (two units more lenient)	b_{21}	-1.31	-1.02	-0.29
		b_{22}	0.21	-0.01	0.21
		b_{23}	1.13	1.04	0.09
	Rater 3	b_{31}	-0.47	-1.16	0.69
		b_{32}	0.80	-0.02	0.82
		b_{33}	1.77	1.18	0.59
CR_2	Rater 4	b_{41}	-1.18	-1.22	0.04
		b_{42}	0.30	0.30	0.00
		b_{43}	1.37	1.46	-0.09
	Rater 5	b_{51}	-1.43	-1.46	0.03
		b_{52}	0.37	0.36	0.00
		b_{53}	1.55	1.65	-0.10
	Rater 6	b_{61}	-1.60	-1.63	0.03
		b_{62}	0.41	0.41	-0.01
		b_{63}	1.67	1.77	-0.09
CR_3 (one unit easier)	Rater 7 (one unit more severe)	b_{71}	-1.58	-1.50	-0.08
		b_{72}	-0.42	-0.29	-0.13
		b_{73}	1.28	1.50	-0.22
	Rater 8 (two units more severe)	b_{81}	-1.42	-1.61	0.19
		b_{82}	-0.19	-0.32	0.13
		b_{83}	1.58	1.67	-0.09
	Rater 9	b_{91}	-1.93	-1.67	-0.26
		b_{92}	-0.74	-0.35	-0.38
		b_{93}	1.26	1.77	-0.51

Table A20. HRM-SDT Results for Rater Parameters in Condition A7 of Simulation A (Assign Three More Lenient Raters to One More Difficult Item and Three More Severe Raters to One Easier Item)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	Rater1 (one unit more lenient)	c_{111}	0	1	0.02	0.02	—	0.01
		c_{121}	2	3	2.01	0.01	0.69	0.02
		c_{131}	4	5	4.03	0.03	0.80	0.05
		d_{11}	2	2	1.89	-0.11	5.46	0.06
	Rater2 (two units more lenient)	c_{211}	-0.5	1.5	-0.49	0.01	1.21	0.02
		c_{221}	2.5	4.5	2.50	0.00	0.05	0.06
		c_{231}	5.5	7.5	5.55	0.05	0.82	0.16
		d_{21}	3	3	2.83	-0.17	5.57	0.16
	Rater3 (one unit more lenient)	c_{311}	1	2	1.04	0.04	4.17	0.03
		c_{321}	5	6	4.94	-0.06	1.15	0.14
		c_{331}	9	10	8.90	-0.10	1.16	0.51
		d_{31}	4	4	3.68	-0.32	8.11	0.32
CR_2	Rater4	c_{412}	1.5	1.5	1.51	0.01	0.45	0.03
		c_{422}	4.5	4.5	4.53	0.03	0.67	0.05
		c_{432}	7.5	7.5	7.55	0.05	0.72	0.09
		d_{42}	3	3	3.01	0.01	0.50	0.02
	Rater5	c_{512}	2	2	1.99	-0.01	0.53	0.06
		c_{522}	6	6	6.03	0.03	0.48	0.10
		c_{532}	10	10	10.04	0.04	0.39	0.20
		d_{52}	4	4	4.01	0.01	0.37	0.04
	Rater6	c_{612}	2.5	2.5	2.45	-0.05	1.92	0.11
		c_{622}	7.5	7.5	7.49	-0.01	0.16	0.19
		c_{632}	12.5	12.5	12.51	0.01	0.07	0.37
		d_{62}	5	5	4.99	-0.01	0.24	0.08
CR_3 (one unit easier)	Rater7 (one unit more severe)	c_{713}	3	2	0.99	-2.01	66.98	6.66
		c_{723}	7	6	4.80	-2.20	31.36	8.82
		c_{733}	11	10	9.16	-1.84	16.77	6.00
		d_{73}	4	4	3.44	-0.56	14.10	0.57
	Rater8 (two units more severe)	c_{813}	4.5	2.5	1.90	-2.60	57.76	11.13
		c_{823}	9.5	7.5	6.64	-2.86	30.14	14.45
		c_{833}	14.5	12.5	12.30	-2.20	15.15	9.27
		d_{83}	5	5	4.31	-0.69	13.89	0.96
	Rater9 (one unit more severe)	c_{913}	4	3	1.02	-2.98	74.61	13.78
		c_{923}	10	9	6.93	-3.07	30.75	17.96
		c_{933}	16	15	13.98	-2.02	12.65	8.33
		d_{93}	6	6	5.45	-0.55	9.16	0.65

Table A21. HRM-SDT Results for Item Parameters in Condition A7 of Simulation A (Assign Three More Lenient Raters to One More Difficult Item and Three More Severe Raters to One Easier Item)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (on unit more difficult)	b_{11}	0	-1	-0.02	-0.02	—	0.01
	b_{12}	1	0	0.94	-0.06	5.79	0.04
	b_{13}	2	1	1.55	-0.45	22.73	0.64
	a_1	0.5	0.5	0.48	-0.02	4.53	0.01
CR_2	b_{21}	-2	-2	-2.03	-0.03	1.31	0.05
	b_{22}	0.5	0.5	0.51	0.01	1.09	0.01
	b_{23}	2	2	2.03	0.03	1.29	0.08
	a_2	1	1	1.02	0.02	2.04	0.03
CR_3 (on unit easier)	b_{31}	-4	-3	-2.20	1.80	45.05	4.90
	b_{32}	-1.5	-0.5	-0.25	1.25	83.54	3.70
	b_{33}	2	3	1.55	-0.45	22.55	0.66
	a_3	2	2	1.73	-0.27	13.26	0.25

Table A22. Rasch Model Results for Rater Parameters in Condition A7 of Simulation A (Assign Three More Lenient Raters to One More Difficult Item and Three More Severe Raters to One Easier Item)

			Condition A7	Condition A1 (Baseline)	Condition A7-Condition A1
CR_1 (one unit more difficult)	Rater 1 (one unit more lenient)	b_{11}	-0.65	-0.76	0.10
		b_{12}	0.32	0.00	0.32
		b_{13}	1.05	0.76	0.28
	Rater 2 (two units more lenient)	b_{21}	-1.31	-1.02	-0.28
		b_{22}	0.21	-0.01	0.22
		b_{23}	1.14	1.04	0.10
	Rater 3 (one unit more lenient)	b_{31}	-0.08	-1.16	1.08
		b_{32}	1.08	-0.02	1.10
		b_{33}	2.20	1.18	1.01
CR_2	Rater 4	b_{41}	-1.18	-1.22	0.04
		b_{42}	0.29	0.30	-0.01
		b_{43}	1.38	1.46	-0.08
	Rater 5	b_{51}	-1.45	-1.46	0.01
		b_{52}	0.37	0.36	0.01
		b_{53}	1.57	1.65	-0.08
	Rater 6	b_{61}	-1.61	-1.63	0.02
		b_{62}	0.42	0.41	0.01
		b_{63}	1.67	1.77	-0.09
CR_3 (one unit easier)	Rater 7 (one unit more severe)	b_{71}	-1.61	-1.50	-0.11
		b_{72}	-0.43	-0.29	-0.14
		b_{73}	1.29	1.50	-0.22
	Rater 8 (two units more severe)	b_{81}	-1.45	-1.61	0.16
		b_{82}	-0.18	-0.32	0.14
		b_{83}	1.58	1.67	-0.09
	Rater 9 (one unit more severe)	b_{91}	-2.22	-1.67	-0.55
		b_{92}	-0.92	-0.35	-0.57
		b_{93}	1.12	1.77	-0.66

Table B1. Population Values of Rater Parameters for Nine Raters in Four Conditions of Simulation B

			Simulation Conditions			
			B1	B2	B3	B4
CR_1	Rater1	c_{111}	1	2	0	0
		c_{121}	3	4	2	2
		c_{131}	5	6	4	4
		d_{11}	2	2	2	2
	Rater2	c_{211}	1	0	-1	-1
		c_{221}	3	2	1	1
		c_{231}	5	4	3	3
		d_{21}	2	2	2	2
	Rater3	c_{311}	1	1	1	0
		c_{321}	3	3	3	2
		c_{331}	5	5	5	4
		d_{31}	2	2	2	2
CR_2	Rater4	c_{412}	1	2	1	1
		c_{422}	3	4	3	3
		c_{432}	5	6	5	5
		d_{42}	2	2	2	2
	Rater5	c_{512}	1	0	1	1
		c_{522}	3	2	3	3
		c_{532}	5	4	5	5
		d_{52}	2	2	2	2
	Rater6	c_{612}	1	1	1	1
		c_{622}	3	3	3	3
		c_{632}	5	5	5	5
		d_{62}	2	2	2	2
CR_3	Rater7	c_{713}	1	2	2	2
		c_{723}	3	4	4	4
		c_{733}	5	6	6	6
		d_{73}	2	2	2	2
	Rater8	c_{813}	1	0	3	3
		c_{823}	3	2	5	5
		c_{833}	5	4	7	7
		d_{83}	2	2	2	2
	Rater9	c_{913}	1	1	1	2
		c_{923}	3	3	3	4
		c_{933}	5	5	5	6
		d_{93}	2	2	2	2

Table B2. Population Values of Item Parameters for Three CR Items in Four Conditions of Simulation B

		Simulation Conditions			
		B1	B2	B3	B4
CR_1	b_{11}	-2	-1	-1	-1
	b_{12}	0.5	1.5	1.5	1.5
	b_{13}	2	3	3	3
	a_1	1	1	1	1
CR_2	b_{21}	-2	-2	-2	-2
	b_{22}	0.5	0.5	0.5	0.5
	b_{23}	2	2	2	2
	a_2	1	1	1	1
CR_3	b_{31}	-2	-3	-3	-3
	b_{32}	0.5	-0.5	-0.5	-0.5
	b_{33}	2	1	1	1
	a_3	1	1	1	1

Table B3. HRM-SDT Results for Rater Parameters in Condition B1 of Simulation B

			Value	Estimate	Bias	%Bias	MSE
CR_1	Rater1	c_{111}	1	0.76	-0.24	24.02	0.22
		c_{121}	3	2.78	-0.22	7.33	0.23
		c_{131}	5	4.78	-0.22	4.44	0.27
		d_{11}	2	1.90	-0.10	4.98	0.07
	Rater2	c_{211}	1	0.79	-0.21	21.47	0.23
		c_{221}	3	2.78	-0.22	7.29	0.24
		c_{231}	5	4.81	-0.19	3.81	0.29
		d_{21}	2	1.92	-0.08	4.07	0.07
	Rater3	c_{311}	1	0.77	-0.23	22.72	0.21
		c_{321}	3	2.77	-0.23	7.62	0.23
		c_{331}	5	4.78	-0.22	4.40	0.29
		d_{31}	2	1.90	-0.10	4.90	0.08
CR_2	Rater4	c_{412}	1	0.87	-0.13	13.43	0.15
		c_{422}	3	2.89	-0.11	3.76	0.17
		c_{432}	5	4.91	-0.09	1.84	0.24
		d_{42}	2	1.95	-0.05	2.29	0.07
	Rater5	c_{512}	1	0.87	-0.13	12.95	0.12
		c_{522}	3	2.88	-0.12	4.10	0.15
		c_{532}	5	4.88	-0.12	2.39	0.20
		d_{52}	2	1.95	-0.05	2.57	0.06
	Rater6	c_{612}	1	0.90	-0.10	10.08	0.12
		c_{622}	3	2.90	-0.10	3.22	0.15
		c_{632}	5	4.90	-0.10	1.92	0.19
		d_{62}	2	1.97	-0.03	1.31	0.06
CR_3	Rater7	c_{713}	1	0.76	-0.24	24.15	0.22
		c_{723}	3	2.76	-0.24	7.94	0.22
		c_{733}	5	4.77	-0.23	4.55	0.25
		d_{73}	2	1.90	-0.10	5.11	0.07
	Rater8	c_{813}	1	0.80	-0.20	19.94	0.22
		c_{823}	3	2.84	-0.16	5.39	0.22
		c_{833}	5	4.84	-0.16	3.15	0.25
		d_{83}	2	1.94	-0.06	2.98	0.07
	Rater9	c_{913}	1	0.75	-0.25	24.88	0.24
		c_{923}	3	2.73	-0.27	9.00	0.28
		c_{933}	5	4.73	-0.27	5.42	0.32
		d_{93}	2	1.88	-0.12	6.01	0.09

Table B4. HRM-SDT Results for Item Parameters in Condition B1 of Simulation B

		Value	Estimate	Bias	%Bias	MSE
CR_1	b_{11}	-2	-1.67	0.33	16.70	0.47
	b_{12}	0.5	0.57	0.07	14.13	0.08
	b_{13}	2	1.84	-0.16	8.19	0.40
	a_1	1	1.00	0.00	0.06	0.04
CR_2	b_{21}	-2	-1.86	0.14	7.11	0.28
	b_{22}	0.5	0.59	0.09	18.27	0.08
	b_{23}	2	1.92	-0.08	4.21	0.44
	a_2	1	1.01	0.01	1.07	0.06
CR_3	b_{31}	-2	-1.66	0.34	16.89	0.55
	b_{32}	0.5	0.61	0.11	21.94	0.08
	b_{33}	2	1.77	-0.23	11.37	0.49
	a_3	1	0.96	-0.04	4.40	0.05

Table B5. HRM-SDT Results for Rater Parameters in Condition B2 of Simulation B (Rater Severity and Item Difficulty)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	Rater 1 (one unit more severe)	c_{111}	2	1	2.06	0.06	2.96	0.21
		c_{121}	4	3	4.07	0.07	1.75	0.27
		c_{131}	6	5	6.09	0.09	1.46	0.33
		d_{11}	2	2	1.88	-0.12	6.17	0.08
	Rater 2 (one unit more lenient)	c_{211}	0	1	0.08	0.08	—	0.13
		c_{221}	2	3	2.08	0.08	4.06	0.16
		c_{231}	4	5	4.10	0.10	2.55	0.20
		d_{21}	2	2	1.89	-0.11	5.27	0.07
	Rater3	c_{311}	1	1	1.04	0.04	3.76	0.12
		c_{321}	3	3	3.05	0.05	1.61	0.15
		c_{331}	5	5	5.06	0.06	1.14	0.20
		d_{31}	2	2	1.87	-0.13	6.72	0.07
CR_2	Rater 4 (one unit more severe)	c_{412}	2	1	1.78	-0.22	11.01	0.25
		c_{422}	4	3	3.79	-0.21	5.28	0.28
		c_{432}	6	5	5.78	-0.22	3.64	0.36
		d_{42}	2	2	1.91	-0.09	4.46	0.07
	Rater 5 (one unit more lenient)	c_{512}	0	1	-0.20	-0.20	—	0.20
		c_{522}	2	3	1.79	-0.21	10.40	0.23
		c_{532}	4	5	3.80	-0.20	4.88	0.25
		d_{52}	2	2	1.92	-0.08	3.81	0.08
	Rater6	c_{612}	1	1	0.80	-0.20	19.74	0.24
		c_{622}	3	3	2.82	-0.18	5.88	0.28
		c_{632}	5	5	4.86	-0.14	2.87	0.34
		d_{62}	2	2	1.94	-0.06	2.97	0.08
CR_3 (one unit easier)	Rater7 (one unit more severe)	c_{713}	2	1	1.10	-0.90	45.05	1.10
		c_{723}	4	3	3.10	-0.90	22.58	1.14
		c_{733}	6	5	5.08	-0.92	15.25	1.20
		d_{73}	2	2	1.73	-0.27	13.44	0.13
	Rater8 (one unit more lenient)	c_{813}	0	1	-0.92	-0.92	—	1.10
		c_{823}	2	3	1.04	-0.96	47.85	1.18
		c_{833}	4	5	3.04	-0.96	24.09	1.24
		d_{83}	2	2	1.69	-0.31	15.37	0.17
	Rater9	c_{913}	1	1	0.10	-0.90	89.85	1.04
		c_{923}	3	3	2.13	-0.87	29.05	1.01
		c_{933}	5	5	4.17	-0.83	16.59	0.98
		d_{93}	2	2	1.77	-0.23	11.31	0.10

Table B6. HRM-SDT Results for Item Parameters in Condition B2 of Simulation B (Rater Severity and Item Difficulty)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	b_{11}	-1	-2	-1.08	-0.08	8.26	0.21
	b_{12}	1.5	0.5	1.25	-0.25	16.81	0.38
	b_{13}	3	2	2.37	-0.63	21.05	1.05
	a_1	1	1	0.96	-0.04	3.63	0.05
CR_2	b_{21}	-2	-2	-1.73	0.27	13.46	0.56
	b_{22}	0.5	0.5	0.60	0.10	20.04	0.12
	b_{23}	2	2	1.80	-0.20	10.06	0.53
	a_2	1	1	0.98	-0.02	1.70	0.07
CR_3 (one unit easier)	b_{31}	-3	-2	-1.55	1.45	48.31	2.77
	b_{32}	-0.5	0.5	-0.07	0.43	86.87	0.31
	b_{33}	1	2	0.95	-0.05	5.24	0.39
	a_3	1	1	0.89	-0.11	10.59	0.07

Table B7. Rasch Model Results for Rater Parameters in Condition B2 of Simulation B (Rater Severity and Item Difficulty)

			Condition B2	Condition B1 (Baseline)	Condition B2-Condition B1
CR_1 (one unit more difficult)	Rater 1 (one unit more severe)	b_{11}	0.51	-0.56	1.07
		b_{12}	0.99	0.18	0.82
		b_{13}	1.59	0.71	0.88
	Rater 2 (one unit more lenient)	b_{21}	-0.70	-0.54	-0.16
		b_{22}	0.09	0.15	-0.05
		b_{23}	0.77	0.72	0.05
	Rater 3	b_{31}	-0.13	-0.54	0.41
		b_{32}	0.59	0.16	0.43
		b_{33}	1.20	0.72	0.48
CR_2	Rater 4 (one unit more severe)	b_{41}	0.03	-0.56	0.60
		b_{42}	0.57	0.18	0.40
		b_{43}	1.15	0.73	0.41
	Rater 5 (one unit more lenient)	b_{51}	-1.08	-0.55	-0.53
		b_{52}	-0.30	0.17	-0.47
		b_{53}	0.25	0.72	-0.47
	Rater 6	b_{61}	-0.56	-0.54	-0.02
		b_{62}	0.16	0.17	-0.01
		b_{63}	0.72	0.70	0.02
CR_3 (one unit easier)	Rater 7 (one unit more severe)	b_{71}	-0.42	-0.55	0.12
		b_{72}	0.12	0.18	-0.06
		b_{73}	0.74	0.73	0.02
	Rater 8 (one unit more lenient)	b_{81}	-1.44	-0.55	-0.89
		b_{82}	-0.70	0.20	-0.90
		b_{83}	-0.29	0.70	-1.00
	Rater 9	b_{91}	-0.97	-0.54	-0.43
		b_{92}	-0.28	0.16	-0.45
		b_{93}	0.22	0.72	-0.51

Table B8. HRM-SDT Results for Rater Parameters in Condition B3 of Simulation B (Assign Two More Lenient Raters to One More Difficult Item and Two More Severe Raters to One Easier Item)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	Rater1 (one unit more lenient)	c_{111}	0	1	0.13	0.13	—	0.18
		c_{121}	2	3	2.15	0.15	7.63	0.23
		c_{131}	4	5	4.17	0.17	4.34	0.28
		d_{11}	2	2	1.89	-0.11	5.68	0.09
	Rater2 (two units more lenient)	c_{211}	-1	1	-0.89	0.11	11.05	0.13
		c_{221}	1	3	1.13	0.13	12.68	0.15
		c_{231}	3	5	3.14	0.14	4.55	0.18
		d_{21}	2	2	1.87	-0.13	6.64	0.08
	Rater3	c_{311}	1	1	1.13	0.13	13.30	0.25
		c_{321}	3	3	3.16	0.16	5.30	0.29
		c_{331}	5	5	5.17	0.17	3.31	0.33
		d_{31}	2	2	1.87	-0.13	6.30	0.09
CR_2	Rater4	c_{412}	1	1	0.86	-0.14	14.42	0.18
		c_{422}	3	3	2.85	-0.15	5.05	0.21
		c_{432}	5	5	4.86	-0.14	2.85	0.25
		d_{42}	2	2	1.94	-0.06	3.22	0.07
	Rater5	c_{512}	1	1	0.85	-0.15	15.28	0.18
		c_{522}	3	3	2.87	-0.13	4.41	0.19
		c_{532}	5	5	4.88	-0.12	2.45	0.24
		d_{52}	2	2	1.94	-0.06	2.77	0.08
	Rater6	c_{612}	1	1	0.83	-0.17	17.28	0.15
		c_{622}	3	3	2.83	-0.17	5.58	0.19
		c_{632}	5	5	4.84	-0.16	3.24	0.21
		d_{62}	2	2	1.92	-0.08	4.01	0.05
CR_3 (one unit easier)	Rater7 (one unit more severe)	c_{713}	2	1	0.98	-1.02	50.84	1.34
		c_{723}	4	3	3.02	-0.98	24.57	1.32
		c_{733}	6	5	5.03	-0.97	16.23	1.37
		d_{73}	2	2	1.73	-0.27	13.41	0.13
	Rater8 (two units more severe)	c_{813}	3	1	1.92	-1.08	35.86	1.39
		c_{823}	5	3	3.93	-1.07	21.34	1.40
		c_{833}	7	5	5.92	-1.08	15.37	1.46
		d_{83}	2	2	1.69	-0.31	15.33	0.14
	Rater9	c_{913}	1	1	-0.07	-1.07	106.69	1.36
		c_{923}	3	3	1.95	-1.05	35.04	1.38
		c_{933}	5	5	3.96	-1.04	20.76	1.39
		d_{93}	2	2	1.71	-0.29	14.68	0.14

Table B9. HRM-SDT Results for Item Parameters in Condition B3 of Simulation B (Assign Two More Lenient Raters to One More Difficult Item and Two More Severe Raters to One Easier Item)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	b_{11}	-1	-2	-1.15	-0.15	14.59	0.38
	b_{12}	1.5	0.5	1.13	-0.37	24.34	0.59
	b_{13}	3	2	2.23	-0.77	25.57	1.14
	a_1	1	1	0.92	-0.08	8.46	0.05
CR_2	b_{21}	-2	-2	-1.76	0.24	11.89	0.40
	b_{22}	0.5	0.5	0.57	0.07	14.18	0.07
	b_{23}	2	2	1.86	-0.14	6.90	0.43
	a_2	1	1	0.98	-0.02	1.57	0.05
CR_3 (one unit easier)	b_{31}	-3	-2	-1.25	1.75	58.50	3.59
	b_{32}	-0.5	0.5	-0.10	0.40	80.51	0.26
	b_{33}	1	2	1.07	0.07	7.43	0.22
	a_3	1	1	0.88	-0.12	11.86	0.07

Table B10. Rasch Model Results for Rater Parameters in Condition B3 of Simulation B Studies (Assign Two More Lenient Raters to One More Difficult Item and Two More Severe Raters to One Easier Item)

			Condition B3	Condition B1 (Baseline)	Condition B3-Condition B1
CR_1 (one unit more difficult)	Rater 1 (one unit more lenient)	b_{11}	-0.70	-0.56	-0.14
		b_{12}	0.09	0.18	-0.08
		b_{13}	0.76	0.71	0.05
	Rater 2 (two units more lenient)	b_{21}	-1.26	-0.54	-0.72
		b_{22}	-0.40	0.15	-0.55
		b_{23}	0.27	0.72	-0.45
	Rater 3	b_{31}	-0.13	-0.54	0.41
		b_{32}	0.60	0.16	0.44
		b_{33}	1.19	0.72	0.47
CR_2	Rater 4	b_{41}	-0.52	-0.56	0.04
		b_{42}	0.16	0.18	-0.01
		b_{43}	0.71	0.73	-0.02
	Rater 5	b_{51}	-0.54	-0.55	0.01
		b_{52}	0.18	0.17	0.01
		b_{53}	0.71	0.72	-0.01
	Rater 6	b_{61}	-0.54	-0.54	0.00
		b_{62}	0.17	0.17	0.00
		b_{63}	0.72	0.70	0.02
CR_3 (one unit easier)	Rater 7 (one unit more severe)	b_{71}	-0.42	-0.55	0.13
		b_{72}	0.13	0.18	-0.05
		b_{73}	0.74	0.73	0.02
	Rater 8 (two units more severe)	b_{81}	0.10	-0.55	0.65
		b_{82}	0.51	0.20	0.32
		b_{83}	1.27	0.70	0.57
	Rater 9	b_{91}	-0.95	-0.54	-0.41
		b_{92}	-0.27	0.16	-0.43
		b_{93}	0.22	0.72	-0.50

Table B11. HRM-SDT Results for Rater Parameters in Condition B4 of Simulation B (Assign Three More Lenient Raters to One More Difficult Item and Three More Severe Raters to One Easier Item)

			Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	Rater1 (one unit more lenient)	c_{111}	0	1	0.12	0.12	—	0.14
		c_{121}	2	3	2.11	0.11	5.31	0.17
		c_{131}	4	5	4.12	0.12	2.89	0.21
		d_{11}	2	2	1.89	-0.11	5.43	0.08
	Rater2 (two units more lenient)	c_{211}	-1	1	-0.89	0.11	11.28	0.09
		c_{221}	1	3	1.11	0.11	10.60	0.10
		c_{231}	3	5	3.11	0.11	3.66	0.14
		d_{21}	2	2	1.89	-0.11	5.40	0.09
	Rater3 (one unit more lenient)	c_{311}	0	1	0.13	0.13	—	0.14
		c_{321}	2	3	2.14	0.14	6.79	0.18
		c_{331}	4	5	4.16	0.16	3.92	0.23
		d_{31}	2	2	1.92	-0.08	4.06	0.08
CR_2	Rater4	c_{412}	1	1	0.86	-0.14	14.07	0.18
		c_{422}	3	3	2.86	-0.14	4.70	0.20
		c_{432}	5	5	4.87	-0.13	2.58	0.23
		d_{42}	2	2	1.91	-0.09	4.58	0.06
	Rater5	c_{512}	1	1	0.88	-0.12	12.18	0.16
		c_{522}	3	3	2.91	-0.09	3.09	0.17
		c_{532}	5	5	4.93	-0.07	1.46	0.21
		d_{52}	2	2	1.93	-0.07	3.54	0.06
	Rater6	c_{612}	1	1	0.89	-0.11	11.09	0.19
		c_{622}	3	3	2.89	-0.11	3.63	0.21
		c_{632}	5	5	4.91	-0.09	1.80	0.24
		d_{62}	2	2	1.93	-0.07	3.41	0.07
CR_3 (one unit easier)	Rater7 (one unit more severe)	c_{713}	2	1	0.84	-1.16	58.20	1.58
		c_{723}	4	3	2.85	-1.15	28.79	1.58
		c_{733}	6	5	4.85	-1.15	19.14	1.64
		d_{73}	2	2	1.65	-0.35	17.32	0.16
	Rater8 (two units more severe)	c_{813}	3	1	1.84	-1.16	38.72	1.60
		c_{823}	5	3	3.85	-1.15	22.94	1.58
		c_{833}	7	5	5.85	-1.15	16.36	1.61
		d_{83}	2	2	1.65	-0.35	17.42	0.16
	Rater9 (one unit more severe)	c_{913}	2	1	0.82	-1.18	58.80	1.62
		c_{923}	4	3	2.84	-1.16	29.12	1.63
		c_{933}	6	5	4.85	-1.15	19.15	1.61
		d_{93}	2	2	1.65	-0.35	17.66	0.17

Table B12. HRM-SDT Results for Item Parameters in Condition B4 of Simulation B (Assign Three More Lenient Raters to One More Difficult Item and Three More Severe Raters to One Easier Item)

		Value	Baseline	Estimate	Bias	%Bias	MSE
CR_1 (one unit more difficult)	b_{11}	-1	-2	-1.14	-0.14	13.97	0.19
	b_{12}	1.5	0.5	1.21	-0.29	19.27	0.49
	b_{13}	3	2	2.31	-0.69	23.13	1.04
	a_1	1	1	0.91	-0.09	8.97	0.05
CR_2	b_{21}	-2	-2	-1.82	0.18	8.95	0.41
	b_{22}	0.5	0.5	0.52	0.02	3.38	0.08
	b_{23}	2	2	1.77	-0.23	11.50	0.45
	a_2	1	1	0.95	-0.05	4.64	0.05
CR_3 (one unit easier)	b_{31}	-3	-2	-1.13	1.87	62.40	4.10
	b_{32}	-0.5	0.5	-0.06	0.44	87.37	0.32
	b_{33}	1	2	0.95	-0.05	4.73	0.18
	a_3	1	1	0.89	-0.11	11.38	0.06

Table B13. The Rasch Model Results for Rater Parameters in Condition B4 of Simulation B (Assign Three More Lenient Raters to One More Difficult Item and Three More Severe Raters to One Easier Item)

			Condition B4	Condition B1 (Baseline)	Condition B4 - Condition B1
CR_1 (one unit more difficult)	Rater 1 (one unit more lenient)	b_{11}	-0.69	-0.56	-0.13
		b_{12}	0.08	0.18	-0.10
		b_{13}	0.76	0.71	0.05
	Rater 2 (two units more lenient)	b_{21}	-1.23	-0.54	-0.70
		b_{22}	-0.41	0.15	-0.56
		b_{23}	0.26	0.72	-0.45
	Rater 3 (one unit more lenient)	b_{31}	-0.69	-0.54	-0.15
		b_{32}	0.08	0.16	-0.08
		b_{33}	0.76	0.72	0.04
CR_2	Rater 4	b_{41}	-0.54	-0.56	0.02
		b_{42}	0.16	0.18	-0.01
		b_{43}	0.72	0.73	-0.01
	Rater 5	b_{51}	-0.55	-0.55	0.01
		b_{52}	0.18	0.17	0.01
		b_{53}	0.72	0.72	0.01
	Rater 6	b_{61}	-0.54	-0.54	0.00
		b_{62}	0.16	0.17	-0.01
		b_{63}	0.72	0.70	0.02
CR_3 (one unit easier)	Rater 7 (one unit more severe)	b_{71}	-0.42	-0.55	0.13
		b_{72}	0.11	0.18	-0.06
		b_{73}	0.73	0.73	0.01
	Rater 8 (two units more severe)	b_{81}	0.10	-0.55	0.66
		b_{82}	0.51	0.20	0.31
		b_{83}	1.27	0.70	0.57
	Rater 9 (one unit more severe)	b_{91}	-0.43	-0.54	0.11
		b_{92}	0.11	0.16	-0.05
		b_{93}	0.75	0.72	0.03