Detecting Rater Centrality Effect
Using Simulation Methods and Rasch Measurement Analysis

Xiaohui Yue

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Educational Research and Evaluation

Dissertation Committee:
Edward W. Wolfe
Gary E. Skaggs
Yasuo Miyazaki
Elizabeth G. Creamer

July 14, 2011
Blacksburg, VA

Keywords: Performance assessment, rater effects, centrality, Rasch measurement, ANOVA, Type I and Type II errors, statistical power, logistic regression

Detecting Rater Centrality Effect Using Simulation Methods and Rasch Measurement Analysis

Xiaohui Yue

ABSTRACT

This dissertation illustrates how to detect the rater centrality effect in a simulation study that approximates data collected in large scale performance assessment settings. It addresses three research questions that: (1) which of several centrality-detection indices are most sensitive to the difference between effect raters and non-effect raters; (2) how accurate (and inaccurate), in terms of Type I error rate and statistical power, each centrality-detection index is in flagging effect raters; and (3) how the features of the data collection design (i.e., the independent variables including the level of centrality strength, the double-scoring rate, and the number of raters and ratees) influence the accuracy of rater classifications by these centrality-detection indices. The results reveal that the measure-residual correlation, the expected-residual correlation, and the standardized deviation of assigned scores perform better than the point-measure correlation. The mean-square fit statistics, traditionally viewed as potential indicators of rater centrality, perform poorly in terms of differentiating central raters from normal raters. Along with the rater slope index, the mean-square fit statistics did not appear to be sensitive to the rater centrality effect. All of these indices provided reasonable protection against Type I errors when all responses were double scored, and that higher statistical power was achieved when responses were 100% double scored in comparison to only 10% being double scored. With a consideration on balancing both Type I error and statistical power, I recommend the measure-residual correlation and the expected-residual correlation for detecting the centrality effect. I suggest using the point-measure correlation only when responses are 100% double scored. The four parameters evaluated in the experimental simulations had different impact on the accuracy of rater classification. The results show that improving the classification accuracy for non-effect raters may come at a cost of reducing the classification accuracy for effect raters. Some simple guidelines for the expected impact of classification accuracy when a higher-order interaction exists summarized from the analyses offer a glimpse of the “pros” and “cons” in adjusting the magnitude of the parameters when we evaluate the impact of the four experimental parameters on the outcomes of rater classification.

Acknowledgements

The printed pages of this dissertation hold far more than the culmination of years of study. The work reflects the relationships with many generous and inspiring people I have met since beginning my graduate work. I cherish each contribution from them to my development of career as a scholar:

My thanks and deep appreciation to Edward W. Wolfe for preserving with me as my advisor throughout the time it took me to complete this research and finish the dissertation. His commitment of continuing mentoring me after leaving academia has supported me to go through hard times in my life. Without his tremendous encouragement and thoughtful guidance, I would not have been able to make the achievements in the past few years.

The members of my dissertation committee, Gary E. Skaggs, Yasuo Miyazaki, and Elizabeth G. Creamer, have generously given their time and expertise to better my work. I thank them for their contribution and their good-natured support.

I am grateful to the group of psychometricians whom I worked with at Pearson during my summer internship in 2010, especially David Shin, Yuehmei Chien, Bob Parker and James Ingrisone. They generously shared their experience and insights, and gave me honest advice on both career and life wise matters.

I am grateful too to every member in the Educational Research and Evaluation program at Virginia Tech besides my dissertation committee members, Mido Chang and Serge Hein, the lovely staff, Karen Price, and my delightful colleagues, Youngyun Chung Baek, Leigh Harrell, Roofia Galeshi and Sabrina Simpson, for cheering me up and spurring me on.

My thanks must also go to my parents, my husband, Yong, and my 5-year-old daughter, Emily, for their love, support and understanding during the long years of my education.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Significance of Current Study

Performance assessments are used extensively in educational testing, particularly in content areas that are difficult to measure with multiple-choice items, and ratings assigned by a human are commonly used to judge the quality of the assessment products (Landy & Farr, 1980). Unfortunately, performance ratings are prone to various types of systematic and random error, potentially rendering the associated scores inaccurate as indicators of the student's true performance. That is, raters may exhibit *rater effects* that serve as sources of error in the performance ratings. Therefore, identifying aberrant raters is an important step in evaluating the psychometric qualities of ratings in terms of validity and reliability. Previous research has identified several ways that rater effects are manifested in ratings, and researchers have developed a variety of criteria under different methodological frameworks (e.g., classical test theory, analysis of variance, regression-based analysis, generalizability theory, and Rasch measurement and item response theory) to evaluate the psychometric qualities of performance ratings (Saal, Downey, & Lahey, 1980). However, supporting simulation studies and operational applications of these methods have remained sparse.

Previous research concerning rater effects has identified several ways that raters introduce error into ratings, including severity/leniency, centrality/extremism, and inaccuracy (Saal, et al., 1980). This dissertation describes a three-study research series that focused on detection of a relatively less studied rater effect, the centrality effect, within the context of the Rasch measurement framework. A rater who exhibits the centrality effect commonly overuses

the middle categories of the rating scale while avoiding extreme categories. Hence, the centrality effect induces a reduction of the variation of assigned ratings. In addition, the centrality effect results in accurate ratings in the central range of the ability continuum, but overestimates of ratee proficiency for non-proficient ratees and underestimates of ratee proficiency for highly proficient ratees.

Clearly, the centrality effect will manifest itself in the standard deviation of observed ratings. However, the use of the standard deviation of ratings as a rater effect index is problematic because the standard deviation is inflated when random error exists in the ratings. That is, the elimination of either random error or the existence of centrality can decrease the value of the standard deviation. As a result, it would be unclear whether raters who produce ratings with small standard deviations are engaging in centrality or are simply accurate raters. In addition, many operational analyses of rater effects employ latent trait modeling procedures (e.g., item response theory), and numerous indices exist within those frameworks that might be well suited as indicators of the centrality effect.

One family of indices that has been investigated as a potential indicator of the centrality effect includes the various mean-square fit indices that are commonly reported by commercial latent trait software, such as *Winsteps* (Engelhard, 1992, 1994). However, prior research has indicated that rater centrality may manifest itself inconsistently in these indices, and may, therefore, cause analysts to incorrectly conclude that accurate raters are engaging in central rating patterns or that centrality exists when indeed it does not exist in the data in question (Wolfe, Chiu, & Myford, 2000).

Hence, in this dissertation, in addition to the standard deviation of ratings and mean-square fit indices, I evaluated four other indices to detect the centrality effect: the score-measure correlation (also known as the point-measure correlation), the measure-residual correlation, the expected-residual correlation (Wolfe, 2004b, 2005), and the rater slope index (Wolfe, 1998a). I examined the performance of these indices as indicators of the centrality effect as a function of several factors assumed to influence their accuracy in a typical case of operational scoring setting: the number of raters and ratees involved, the level of the centrality effect, and the amount of missing data brought by double-scoring rate (further details provided in Section 3.1).

Thus far, very little research has been done to examine the relative strengths of these indices, the accuracy of these indices, or the conditions under which each of these indices perform best for detecting the centrality effect. The purpose of this research is to tackle these unresolved problems and to provide analysts with reference tools to evaluate the indicators of less studied rater effects. With an eye toward detecting effect raters and promoting the work of non-effect raters, this dissertation also focuses on the relationship between index performance and working conditions. The findings may lend assistance in planning scoring project, such as determining acceptable double-scoring rates, the number of raters to recruit, and the number of ratees allocated to each scoring group where raters are divided up into multiple groups.

1.2 Research questions

The series of analyses in the ensuing discussion aim to address the following research questions:

1) Which centrality-detection indices are most sensitive to the difference between effect raters and non-effect raters?
2) How accurate (and inaccurate), in terms of Type I error rate and statistical power, is each centrality-detection index in flagging effect raters?
3) How do the features of the data collection design (i.e., the independent variables including the level of centrality strength, the double-scoring rate, and the number of raters and ratees) influence the accuracy of rater classifications by these centrality-detection indices?

Each of these research questions was addressed separately in one of the three manuscripts presented in this dissertation. The three studies shared a common data generation process, but each focused on a different aspect of centrality effect detection.

1.3 Overview of Methodology

This project was simulation-based, consisting of a series of three sub studies. One study focused on the relationship between raw rating scores and several latent-trait indices of the centrality effect; one evaluated the accuracy of those centrality indices in identifying aberrant raters; and another examined rater classification accuracy for the centrality indices. All three studies share the same sets of simulated data with full combinations of four experimental parameters: ratee sample size (1000, 2000, and 3000), rater sample size (50, 100 and 250), level of centrality effect (0.1, 0.35, 0.6 and 0.85) and double-scoring rate (0.1 and 1.0). There were 1000 replications for each cell of the experimental design. Data generation was based on a Rasch model using *ConQuest* (Wu, Adams, Wilson, & Haldane, 2007).

I first scaled the simulated data to the Rasch rating scale model (Andrich, 1978) using *Winsteps* (Linacre, 2009b). Based on the estimates from the Rasch rating scale model, I calculated values for six sets of centrality-detection indices. Then, I used analysis of variance (ANOVA) method to identify indices that were potentially sensitive to difference between effect raters and non-effect raters. Next, within each cell bearing with practical importance in the above ANOVA procedure, I computed and compared the proportion of incorrectly detected and incorrectly nominated raters (i.e., Type I and Type II error) for the 1000 replications among the centrality indices and across experimental conditions. Finally, I used logistic regression to assess the relationship between the rater-classification accuracy by each centrality index and the four experimental parameters.

1.4 Organization of Dissertation

The first chapter of the dissertation serves as an introduction of the problem, research questions and methodology used in the three studies. The second chapter contains the literature review. Chapters three, four and five are publishable-quality articles focusing on the relationship between raw rating scores and latent-trait indices, the determination of critical values for each rater-centrality index and accuracy of each index, and the impact of various scoring-setting variables on the rater-classification accuracy by these indices respectively. The final chapter summarizes the results from the three manuscripts, discusses the implications, impact of the research, limitations of current studies and future direction of this research.

2 LITERATURE REVIEW

2.1 Introduction

Performance assessments are used extensively in educational testing, particularly in content areas that are difficult to measure with multiple-choice items. Ratings assigned by human are commonly used to judge the quality of the assessment products (Landy & Farr, 1980). Unfortunately, performance ratings are prone to various types of systematic and random error, potentially rendering the associated scores inaccurate as indicators of the student's true performance. Due to the complex nature of performance ratings that involves human judgment, *the Standards for Educational and Psychological Testing* (1999) explicitly requires performance assessment developers to monitor and report scoring errors, and clearly states that any systematic source of scoring errors should be corrected to assure the accuracy of scoring. Previous research has identified several ways that rater effects are manifested in ratings, and researchers have developed a variety of criteria under different methodological frameworks to evaluate the psychometric quality of performance ratings (Saal, et al., 1980).

Several types of rating errors have been proposed, and some of those are well studied, while others lack thorough evidence-based research. The fact that there are certain rating errors left only partially explored is not due to a lack of importance. Rather, these errors may manifest themselves in ways that are difficult to predict, or the error that they contribute to scores may be confounded with error from other sources, making them difficult to be detected. In addition to the unbalanced attention directed toward different types of rater errors in the literature, another challenge that hampers understanding research results concerning rater errors is a general lack of

agreement among researchers regarding the conceptual and operational definitions of any given rater effect (Saal, et al., 1980). Conceptually, most rater effects are defined in multiple ways. Operationally, researchers use different indices to detect the presence of a particular rater effect in the data. This lack of uniformity extends to the rating designs, the data collection procedures, and the data analysis approaches for investigations on each rater effect (Myford & Wolfe, 2003). Despite of these difficulties, researchers have taken up the challenge of aggregating research results in this field and have prepared comprehensive literature reviews (Landy & Farr, 1980; Myford & Wolfe, 2003; Saal, et al., 1980) and meta-analyses (Hoyt & Kerns, 1999; Murphy & Balzer, 1989) concerning rater effects.

The purpose of this chapter is to summarize the research literature on rater effects and to provide a backdrop for a focused study on the rater-centrality effect. This chapter first summarizes the context of existing research relating to rater effects. Then it provides an overview of the existing psychometric approaches to detect rater effects. Finally, it focuses on the centrality effect and identifies potential indices to be evaluated through simulation methods in the following studies.

2.2 Rater Effects

As discussed in Myford and Wolfe's work (2003), "rater effects", "rater bias", and "rater errors" have been largely undifferentiated in the research literature. For simplicity, I follow the term "rater effects" as defined by Scullen, Mount and Goff (2000), who define them as a "broad category of effects [resulting in] systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (p. 957).

2.2.1 Two Research Subdivisions on Rater Effects

When raters formulate judgments, many factors come into play that may influence their ratings, and researchers have developed different approaches to study these influences. Reviewing the research literature on rater effects, I divide the research into two areas.

One subdivision arises from the perspective of the psychological, behavioral, and personal aspects of human cognitive processing. Research that falls into this category often focuses on questions like how do raters read and judge students writing (Pula & Huot, 1993)? What part do rating procedures play in the process through which raters arrive at judgments about writing quality (Huot, 1990)? What are the contexts and interrelationships among psychological process and social systems in understanding writing? As quoted by Pula and Huot (1993), this type of research falls into functional research (Williamson, 1988) that seeks "an explanation of human behavior against the background of a social system or culture as whole… term[ed]…holistic, because it preserves the unity of personal, psychological events and their public, social functions…[I]t is the meaning that humans themselves impute to their behavior" (pp.90-96). The depth of understanding in these studies of rater effects depends on the sophistication of the model of the psychological process involved in human scoring behavior and rater cognition (Wolfe & Feltovich, 1994).

The research within this subdivision[1] has documented rater effects arising from rater psychological traits such as motivation, anxiety, motivational intensity, achievement, self-

[1] Because the research literature under this division is not the central point of the dissertation, I do not summarize the findings for each of the listed articles. If interested in further details, please refer to the references.

efficacy and mood (AlFallay, 2004; Bernardin & Villanova, 2005; Fried, Levi, Ben-David, Tiegs, & Avital, 2000); personality traits (Wexley & Youtz, 1985); prior beliefs held by the raters (Cooper, 1981; Tziner, Murphy, Cleveland, Beaudin, & Marchand, 1998; Wexley & Youtz, 1985); rater's demographic backgrounds such as gender, race, and second language (Elder, Barkhuizen, Knoch, & von Randow, 2007; Mount & Systma, 1997; Scherer, Owen, & Brodzinski, 1991; Wen, 1979); the rater's job role (Van Iddekinge, Putka, Raymark, & Eidson, 2005); the rater's social style (May & Gueldenzoph, 2006); the rater's prior scoring experience (Weigle, 1998), training experience (Athey & McIntyre, 1987; Bernardin, 1978; Bernardin & Pence, 1980; Borman, 1979), etc. Unfortunately, due to the methodological difficulties of studying the interplay of these complex cognitive, experiential, and social variables, most studies in this area have focused on a small number of raters and a limited range of rater variables. Furthermore, there have been no broadly generalizable results drawn from studies such as these concerning rater behavior and their impact on rater effects, as documented by Landy and Farr (1980).

Another subdivision of research relating to rater effects focuses on particular patterns that emerge in rating data that are assumed to result from rater effects and which can be compared to expected "normal" patterns of ratings for the sake of diagnosis. Observed rating patterns may be labeled as evidence of severity, leniency, centrality, range restriction, or halo depending on the distribution of majority ratings on a rating scale. These patterns can be operationalized into values of statistical indices. Because they are directly derived from rating data, they have been extensively studied in functional research. The next sections of this chapter discuss rater effects

within this subdivision of rater effect research, followed by a review of existing measurement frameworks under which these rater effects are defined.

2.2.2 Types of Rater Effects

Traditionally, researchers have focused on describing and measuring four rater effects : (1) leniency/severity (i.e., a rater rates ratees above or below the values of true ratings for those ratees), (2) halo (i.e., a rater assigns ratings that are more highly correlated across a range of criteria than would be the true ratings across those criteria), (3) central tendency (i.e., a rater overuses the middle category of a rating scale while avoiding the extreme categories), and (4) restriction of range (i.e., a rater overuses any point on a rating continuum). Of these four "classic" rater effects, the halo effect has been the most widely studied and has received the widest attention in the research literature (Myford & Wolfe, 2003). The halo effect differs from the rest in that it is associated with the presence of multiple traits or dimensions in tasks or ratees (Fisicaro & Lance, 1990), while each of the other three rater effects remains on a single trait or dimension. This distinction separates the halo effect from the other "classic" rater effects.

Saal, et al. (1980) explain the connection among the other three effects: range restriction refers to situations in which ratings are clustered about any point on the rating continuum, be it a favorable (lenient) point, an unfavorable (severe) point, or the midpoint (central tendency) on the rating scale. Thus, range restriction may reflect leniency, severity, or central tendency. Hence, we can treat leniency, severity, and central tendency as special cases of restriction of range. As the leniency/severity effect was referred by Cronbach (1990) as the most serious error that a rater can introduce into a rating procedure, numerous studies have been conducted relating to that

effect (Athey & McIntyre, 1987; Bernardin & Walter, 1977; Borman, 1977; Congdon & McQueen, 2000; Engelhard, 1994). As a result, the detection and correction of that effect is fairly well understood. However, the central tendency is relatively less studied.

One noteworthy and lesser known rater effect is inaccuracy (Wolfe, 2004a). Inaccuracy refers to seemingly random discrepancies between a rater's ratings of examinee proficiency and the actual proficiency of those examinees. As a result, inaccuracy does not strictly adhere to the adopted definition of rater effects, which specifies that the error contributed to scores is systematic. In operational scoring settings, rater inaccuracy may occur when a rater has insufficient content background, insufficient training in applying the scoring rubric, or immutable biases or content-based beliefs that cause the rater to assign scores that are not consistent with the scoring rubric. When a rater assigns inaccurate ratings, any ratee scored by the rater could be misplaced on the underlying ability scale, and it is likely that the rank ordering of ratees scored by the effect rater disagrees with scores assigned by accurate raters. By definition, inaccuracy can occur alone or in conjunction with any of other types of rater effects, leaving it difficult for diagnosis and differentiation. Rater inaccuracy is introduced here because it is a common concern for those who evaluate ratings and because it may influence rater effect indices in a manner similar to other rater effects, particularly the centrality effect.

2.2.3 Existing Measurement Approaches

Researchers working from different psychometric perspectives have chosen different measurement frameworks to study rater effects. These approaches include: (a) applications of the classical test theory (CTT) in assessing the mean and the standard deviations of trait ratings

(Murphy & Anhalt, 1992); (b) applications of analysis of variance (ANOVA) method in assessing convergent and discriminant validity (Hedge & Kavanagh, 1988); (c) the use of multivariate analysis of variance (MANOVA) in assessing ratings on multiple performance dimensions (Saal, et al., 1980); (d) the use of regression-based procedures (LaHuis & Avis, 2007); (e) the use of the generalizability theory (G-Theory) in assessing group-level rater effects; and (f) applications of latent trait measurement models. Among these approaches, G-Theory and Multi-Faceted Rasch Measurement (MFRM) are dominant but contrasting approaches in the study of rater effects, and there are several empirical studies using each technique (Kim & Wilson, 2009; Linacre, 1996; Lynch & McNamara, 1998; Smith, 2004; Sudweeks, Reeve, & Bradshaw, 2005). Linacre (1996) compares these two competing methodologies and explains the fundamental differences: G-Theory takes the discrepancy (between the observed score and the mean score over all acceptable observations) into account in determining the reliability of measures while the MFRM applies a correction to the estimated measures. A common conclusion from those comparison studies is that G-Theory provides a general summary including an estimation of the relative influence of each facet on a measure and the reliability of a decision based on the data; while MFRM focuses on the individual examinee or rater and provides as fair of a measure as possible from the data in addition to summary information such as reliability indices. Moreover, the MFRM provides valuable diagnostic information about individual rater performance, allowing for the detection and potential correction of rater effects in operational settings (Myford & Wolfe, 2003). Because of the advantages in detecting both group- and individual-level rater effects, the MFRM approach seems to have gained preferential status over other measurement approaches.

The basic form of the MFRM is an extension of the Rasch rating scale model (Andrich, 1978). The rating scale model depicts the logit value of a ratee *n* being assigned a rating of *x* versus the next lower rating category by a particular rater *r* as a linear function of three parameters that locate the respondent ability θ_n, rater location λ_r, and rating scale category threshold τ_k onto the same latent trait continuum.

$$LN(\frac{p_x}{p_{x-1}}) = \theta_n - \lambda_r - \tau_k,$$

where *k* is the threshold between categories *x* and *x*-1. Parameters for this model are estimated using joint maximum likelihood estimation procedures as implemented in *Facets* (Linacre, 2009a), *Winsteps* (Linacre, 2009b), or *ConQuest* (Wu, et al., 2007).

This model is appropriate for the situation where raters rate ratee responses to a single item, which is a typical case of operational practices in most testing companies. Several statistical estimates associated with this model are powerful for evaluating rater effects. First, the rater severity or leniency effect is explicitly estimated by the λ_r component in the model. λ_r depicts the location of the mean score assigned by rater *r*. Second, four model-data fit statistics associated with the Rasch rating scale model can be used to evaluate rater effects. These include the weighted and unweighted mean-square fit statistics (also called infit and outfit respectively) and the standardized versions of these two fit statistics. The mean-square fit statistics (Wright & Masters, 1982) are based on the standardized residual of the observed response for each person-by-item combination from the modeled expectation, given the parameter estimates,

$$Z_{nr} = \frac{x_{nr} - E_{nr}}{\sqrt{W_{nr}}},$$

where x_{nr} = the score assigned to person *n* by rater *r*,

$E_{nr} = \sum_{k=0}^{m} kp_{nrk}$, the expected score assigned to person *n* by rater *r*,

$W_{nr} = \sum_{k=0}^{m} (k - E_{nr})$, the expected model variance of the scores assigned to all persons by rater *r*, *k* is the scored responses, ranging from 0 to *m*, and p_{nrk} is the model-based probability that person *n* is assigned a score in category *k* by rater *r*. A positive residual indicates that the observation is higher than expected. A negative residual indicates that the observation is lower than expected.

Unweighted mean-square fit statistics for raters are computed as the average of the squared standardized residuals across all persons scored by a rater,

$$UMS_r = \frac{\sum_{n=1}^{N} Z_{nr}^2}{N},$$

where *N* is the number of persons. Weighted mean-square fit statistics for raters are computed as the average of the squared standardized residuals across all persons rated by a rater, each weighted by its variance so that remote scores are given less weight than proximal scores,

$$WMS_r = \frac{\sum_{n=1}^{N} Z_{nr}^2 W_{nr}}{\sum_{n=1}^{N} W_{nr}}.$$

Each of these statistics can be standardized to obtain the standardized weighted and unweighted mean-square fit statistics, i.e., ZWMS and ZUMS (Wright & Masters, 1982). A rule of thumb for upper and lower limits of acceptable mean-square fit values have been suggested to flag items,

and these values have been adopted in operational settings for evaluating raters (i.e., 0.7 and 1.3 for multiple-choice items, 0.6 and 1.4 for rating scales, and $\pm$ 2.0 for the standardized versions (Wright & Linacre, 1994)).

A third potentially powerful index in evaluating rater effects within the MFRM context is the score-measure correlation, $r_{score,measure}$ (also known as the point-measure correlation). In rating data, it correlates Rasch measures of ratees with ratings assigned by raters, and is computed by the formula

$$r_{score,measure} = \frac{1}{n-1}\sum_{n=1}^{N}\left(\frac{X_{nr}-\overline{X}_r}{S_{X_r}}\right)\left(\frac{\theta_n-\overline{\theta}}{S_\theta}\right),$$

where $\frac{X_{nr}-\overline{X}_r}{S_{X_r}}$ is the standard score of rating assigned to ratee *n* by rater *r*, $\overline{X}_r$ is the sample mean of ratings assigned by rater r, S_{X_r} is the sample standard deviation of ratings assigned by rater *r*, $\frac{\theta_n-\overline{\theta}}{S_\theta}$ is the standard score of Rasch measure of ratee *n*, and S_θ is the standard deviation of Rasch measure of all ratees. The index value depicts the consistency between the rank ordering of the examinees by a particular rater and the rank ordering of those examinees by composite scores assigned by all other raters. Therefore, it should be sensitive to rater effects, such as inaccuracy, that create inconsistencies between this pair of measures. On the other hand, the score-measure correlation should not be strongly influenced by rater effects that preserve the rank ordering of examinees (e.g., when centrality exists).

2.3 Rater centrality and indices used for detecting the effect

Researchers have suggested several definitions of central tendency (Cascio, 1982; DeCotiis, 1977; Korman, 1971; Landy & Farr, 1983). Korman (1971) defined central tendency as the tendency to rate all rating objects around the 'middle' or mean of a rating continuum and not to use the extremes. DeCotiis (1977) defined central tendency as a rater's unwillingness to go out on the proverbial limb in either the favorable or unfavorable direction. Most of these definitions are based upon the notion that raters exhibiting this effect overuse the middle categories of a rating scale while avoiding the extreme categories.

By definition, the centrality effect induces a reduction of the variation in assigned ratings. In addition, the centrality effect results in accurate ratings in the central range of the ability continuum, but overestimates of ratee proficiency for non-proficient ratees and underestimates of ratee proficiency for highly proficient ratees. This rater effect also introduces an artificial dependency in the ratings that leads to overly consistent response patterns that can be detected with rater fit statistics.

One approach to detecting the centrality effect relies on the mean and the standard deviation of the ratings assigned to all ratees by a rater on a particular performance dimension or trait (Borman & Dunnette, 1975). Smaller standard deviations reflect greater range restriction. Whether such range restriction represents leniency, severity or central tendency is determined by the mean rating. In other words, a center-regressing mean of the ratings and a relatively small standard deviation would lead to the conclusion of central tendency. Similarly, DeCotiis (1977) and Korman (1971) concentrated specifically on the proximity of the mean dimension ratings to the midpoint of the scale. Statistically, one would expect central tendency to be indicated by neutral ratings with little variability. Therefore, for a given instrument, a displaced mean rating

in conjunction with a relatively large standard deviation would lead to the conclusion that the instrument is sensitive to criterion performance, and therefore resistant to central tendency. Unfortunately, the use of the standard deviation of ratings as a rater effect index is problematic because the standard deviation is inflated when random error exists in the ratings. As a result, it would be unclear whether raters who produce ratings with small standard deviations are engaging in centrality or are simply accurate raters in a sample of inaccurate raters.

Another family of indices that has been investigated as a potential indicator of centrality effect includes the various mean-square fit statistics that are commonly reported by commercial latent trait software, such as *Winsteps* (Engelhard, 1992, 1994). Mean-square fit statistics can be used to detect rater effects, as every rating contributes to both weighted and unweighted mean-square fit statistics. These residual-based statistics capture deviations from expected ratings. Weighted mean-square fit statistic are more influenced by the overall pattern of assigned ratings than unweighted mean-square fit statistic are, because they weight the assigned ratings by their statistical information, which is higher in the center of the test and lower at the extremes. Unweighted mean-square fit statistics are highly influenced by extremely unexpected ratings when the elements of the measurement context are mismatched (i.e., a high ability examinee is rated by a lenient rater). Mean-square fit statistics are chi-square statistics divided by their degrees of freedom. Consequently the indices have a lower bound of zero and are positively unbounded. Values greater than 1.0 indicate unmodeled noise. A 0.1 increase in a fit statistic is associated with a 10% increase in unmodeled error. Values less than 1.0 indicate that the model predicts the data better than the assumed model-based error, which could result in inflated statistics for summary statistics, such as reliability statistics. In general, elements with fit statistic

values ranging from 0.6 to 1.5 are considered to show adequate fit to the model (Wright & Linacre, 1994). Standardized mean-square fit statistics work in the same way as the non-standardized version except that they report the statistical probability of the mean-square statistics occurring by chance when the data fit the Rasch model.

When a rater exhibits the centrality effect in ratings, we may expect the fit statistics to deviate away from 1.0 and the absolute value of standardized fit statistics to be larger than 2.0. We might expect the unweighted mean-square fit statistic to be more sensitive to centrality because it is more heavily influenced by large residuals for mismatched raters and ratees. We could expect these cases to produce larger residuals (i.e., a normal rater would produce ratings of high and low ability ratees that have large residuals). However, prior research has indicated that rater centrality may manifest itself inconsistently in these indices, and may, therefore, cause analysts to incorrectly conclude that accurate raters are engaging in central rating patterns (Wolfe, et al., 2000).

Another index used to detect rater centrality is the expected-residual correlation – the correlation between estimated ratings and residuals, $r_{\exp,res}$ from Rasch measurement models (Wolfe, 2004a, 2005). Specifically, the expected-residual correlation is based on the notion that the raw residual, $x_{nr} - E_{nr}$, produced by all ratings assigned by rater *r* who exhibits the centrality effect is positive for ratees of low ability (i.e., x_{nr} is greater than E_{nr}) and negative for ratees of high ability(i.e., x_{nr} is less than E_{nr}). As a result, when the centrality effect exists, a scatter plot of the expected scores (X axis) and residuals (Y axis) should have a negative relationship. A

value of the expected-residual correlation that approaches its upper limit of 1.00 indicates the way a rater utilizes the rating scale is uniform across ability continuum.

The relationship between the MFRM ratee measures and residuals can also be expected to be influenced by the existence of rater centrality. That is, the correlation between ratee measures and model-based residuals, depicted by $r_{measure,res}$, may also be a powerful rater centrality-detection index. Again, when centrality presents in the rating data, the residuals for low proficiency ratees should be positive (i.e., x_{nr} is greater than E_{nr}) and the residual for high proficiency ratees should be negative (i.e., x_{nr} is less than E_{nr}). As a result, the scatter plot of the ratee measures should display a negative relationship between ratee measures and residuals. That means that the value of $r_{measure,res}$ should be negative when centrality exists in the rating data. A highly positive value of measure-residual correlation indicates the way a rater utilizes the rating scale is uniform across ability continuum. $r_{\exp,res}$ and $r_{measure,res}$ summarize the relationship of residuals with expected scores and ratee calibrations respectively. The only difference between expected score and ratee calibration is that expected score is in response-level score units and ratee calibration is in logits. That is, ratee calibration is response-level score divided by the modeled score variance of observed scores around expected scores, and the modeled score variance is in fact the squared observed scores. Therefore the difference between $r_{\exp,res}$ and $r_{measure,res}$ should be subtle.

Another index that can be used to detect rater centrality is the correlation of Rasch measures of ratees with assigned ratings by a rater, also called the score-measure correlation or

the point-measure correlation, denoted as "PTMEA" in *Winsteps* output. In typical item analyses using Rasch measurement, we use item correlations as an immediate check that the response level scoring is consistent with the scoring of other items (i.e., the rank ordering of ratees is the same for the item as for the estimated ratee measures). In case of study for rater centrality, we can evaluate the score-measure correlation to see if assigned ratings to ratees by a particular rater are consistent with the estimated ratee measures, which would be based on the ratings assigned by all raters. The value of score-measure correlation from raters with centrality should be less than one obtained from raters without centrality. That is, a highly positive value of the score-measure correlation indicates a strong association between ratees' ability measure and ratings assigned by a rater, and raters with centrality effect are expected to exhibit weaker or even negative correlation on this index. However, score-measure correlation along with point-biserial correlation is difficult to interpret because they are influenced by the predictability of the data, the targeting of the rater on the person sample, and the distribution of the personal sample. These factors may cause limited power for the score-measure correlation in detecting the centrality effect.

Wolfe (1998b) proposed using rater slope, denoted as "DISCRM" in *Winsteps*, to detect the centrality effect. The Rasch model specifies that item discrimination, also called the item slope. In case of study for rater effects, it is rater discrimination and rater slope. In *Winsteps*, it is not a parameter but a descriptive statistic. *Winsteps* estimates what the rater slope parameter would have been if it had been parameterized. The Rasch slope is set at 1.0. The empirical slope is estimated through the model that has the appearance of a 2-PL IRT model but differs because the slope parameter is not used in the estimation of other parameters. The reported values of rater

slope are a first approximation to the precise value of a_r obtained from the Newton-Raphson estimation equation (Wright & Masters, 1982, pp. 72-77):

$$\hat{a}_r = 1 + \left[\frac{\sum_{n=1}^{N} (X_{nr} - P_{nr})(\theta_n - b_r)}{\sum_{n=1}^{N} P_{nr}(1 - P_{nr})(\theta_n - b_r)^2} \right],$$

where X_{nr} is the rating assigned to ratee *n* by rater *r*, P_{nr} is the probability for ratee *n* rated by rater *r* to get the item correct, θ_n is the Rasch measure for ratee *n*, and b_r is the severity of ratee *r*. Because a rater exhibiting the centrality effect may assign accurate ratings in the central range of the ability continuum and will overestimate ratee proficiency for non-proficient ratees and underestimate ratee proficiency for highly proficient ratees, the centrality effect rater could have, on average, lower discriminability than non-effect raters. Therefore, small values of the slope index could be a symptom of the centrality effect. The commonly acceptable value for the slope index is 0.5 or above, and raters with lower values are candidates for the centrality effect.

Other approaches suggested for detecting the rater-centrality effect, as summarized by Myford and Wolfe (2004), include checking (a) frequency counts that indicate how many times each rater used each category on each trait scale, (b) table of misfitting ratings, (c) category probability curves for each rater, (d) rating scale category unweighted mean-square indices, and (e) rating scale category thresholds. However, all these solutions remain experimental and are subject to judgmental interpretation because they lack clear-cut standards for analysts to follow. Therefore, they are not considered in this dissertation project.

In summary, in the ensuing studies, I evaluated six types of indices that should provide power in detecting the centrality effect: the standard deviation of assigned scores, the rater fit statistics (including weighted and unweighted, standardized and non-standardized mean-square fit statistics), the point-measure correlation, the expected-residual correlation, the measure-residual correlation and the rater-slope index. Each of these indices has been proposed as a potential index for evaluating rater centrality, and a rationale can be provided for why that index may be a good candidate for detecting centrality. On the other hand, because detailed studies have not been conducted on any of these indices, simulation studies that compare the detection accuracy of all these indices are warranted and should contribute to the practice of detecting rater effects in operational settings.

3 METHODOLOGY

3.1 General methods

The three studies conducted for this dissertation share common elements, particularly in the areas of data generation, parameter estimation, and index computation. Hence, for the sake of being concise, I organized this chapter as follows. The methods that are common to all three studies are described first, followed by sections containing the unique details about each of the three sub-studies.

Ratings on a six-point scale were generated using *ConQuest* 2.0 (Wu, et al., 2007) based on a unidimensional Rasch model that contains a single measurement facet (i.e., ratings assigned by raters to a single assessment item). Rating scale step threshold parameter values were specified to generate an approximately normal distribution of ratings for each simulated rater. In order to avoid confounding the detection of rater centrality with rater severity/leniency effects, all rater severity parameters were set to zero for all raters during data generation. That is, data sets were generated according to a Rasch rating scale model in which the item "difficulty" parameter was replaced by a rater "severity" parameter which was set to a constant value of zero.

Starting with these "non-effect" data, a single simulated rater exhibiting a centrality effect was generated by shrinking the underlying ratee ability to four ***centrality-strength*** levels (0.10, 0.35, 0.60, and 0.85; i.e., the variance of the ability distribution is shrunk to a minimum of 10% and to a maximum of 85% of that of the non-effect rater), and these effect ratings were added into the simulated data file in order to create a dichotomous ***rater-type*** variable (namely

focused non-effect, or "normal", rater and manipulated effect rater). The relationship between the centrality strength and the ability distribution can be expressed as a variance ratio defined as,

$$\text{Variance ratio} = \frac{\sigma^2_{effect}}{\sigma^2_{normal}},$$

where σ^2_{effect} is the variance of ability distribution for the ratees whose responses are rated by effect raters and σ^2_{normal} the variance of ability distribution for the ratees whose responses are rated by normal raters. I generated data according to these procedures for three ***rater-sample sizes*** (50, 100, and 250) and three ***ratee-sample sizes*** (1000, 2000, and 3000). In all data sets, single raters were randomly assigned to ratees, and a second rating was assigned to either 10% or 100% of the ratees (i.e., the ***double-scoring rate***), again, chosen randomly. Accordingly, the number of ratees per rater relies on the random assignment achieved by randomly selecting either 10% or 100% of ratees to assign a second score in the data generating process. The resulting data is not a full rater-by-ratee design (i.e., every rater does not rate every ratee), and with random rater-response assignment, raters would "give" different numbers of ratings across ratees. These designs are typical of those implemented in operational settings. As a result, the simulated rating data sets contained missing values. Consequentially, the number of assigned ratings varied across raters, which created a ***rating-count*** variable which was recorded because this variable could affect the measurement error of the centrality indices and hypothesis tests based on those indices. For each combination of these independent variables, 1000 data sets were generated, resulting in 72,000 data sets (i.e., 1000 replications × 4 centrality-strength levels × 2 double-scoring rates × 3 rater-sample sizes × 3 ratee-sample sizes).

After the data sets were generated, I performed the following activities. First, parameters were estimated for each simulated data set for the Rasch rating scale model (Andrich, 1978) using the *Winsteps* software (Linacre, 2009b). Second, values for the six rater centrality indicators (standard deviation of assigned scores, mean-square fit indices, score-measure correlation, measure-residual correlation, expected-residual correlation, and rater slope) were computed for each rater (effect and non-effect). These indices along with the experimental parameters are the data source for the ensuing analyses. Except for the double-scoring rate, all other experimental parameters were treated as continuous variables because the chosen values of these continuous variables were based on empirical trails with a consideration on continuous distributions, which allows for accurately depicting trends across levels of these variables.

3.2 Study 1: The relationship between raw scores and several latent-trait indices for detecting rater centrality effect

The first dissertation study aims to address the first research question: which centrality-detection indices are most sensitive to the difference between effect raters and non-effect raters? In the subsequent analyses, I treated the difference of the values of each rater centrality index between rater types as the dependent variable in a separate analysis of variance method (ANOVA) with essential assumption checks. The remaining variables were treated as independent variables. For each centrality-effect index, the ANOVA analyses commenced with a saturated model under a full factorial design to capture all possible effects, then used variable elimination procedures to a reduced model containing only effect terms with meaningful effect sizes. I utilized the stepwise procedure for model reduction because the simulated data sets were large enough to be assumed generalizable to other samples from similar scoring settings. Effect

terms containing an interaction with the rater type variable are the most interesting because the effects may be indicative of the power of the index for differentiating effect and non-effect raters.

The ANOVA is an appropriate method to answer the research question for a couple of reasons. First, the way I simulated the effect rater is by shrinking the underlying ability of ratees who were assigned to a particular non-effect rater, namely a focused rater. Assigning ratings to the ratees with manipulated ability by the same rater was assumed to be equivalent to assigning effect ratings to the same ratees by an effect rater. These effect ratings were added into the simulated data file in order to create a rater-type variable. It is equivalent to having the same rater assign ratings twice, once exhibiting the centrality effect, and once not. Thus, after calibration, there were double measures of the centrality effect for a single focused rater, and the two observations of a particular centrality index on this focused rater were not independent to each other. Such a data setting allows us to use the focused rater as its own comparison or control case. That is, relevant factors that might contribute to the outcome of interest may be constant within individuals. Statistically, the ratings assigned by the focused rater and the pseudo effect rater are no longer independent to each other, I need to take this independence into account to get a proper estimates of variability for hypothesis testing.

Second, the ANOVA contains the main effect of relevant factors, the main effect of repeated measure, and the interaction of the relevant factors. In my case, when I treated the double-scoring rate, the centrality-strength level, and the number of raters and ratees as relevant factors, the difference between rater types as the outcome measure. The ANOVA model used

relevant factors and their interactions to explain the impact on difference between rater types measured by a centrality-detection index.

As a result, if a centrality-detection index is associated with effect terms that have meaningful effect size, I can conclude that this index is sensitive to the difference between rater types measured by this particular index. If an index is involved with more than one effect term, attention is given to the most complex one. I used η^2 as indicators of effect sizes, with the cut-off for practical significance being an η^2 greater than 0.06 (Cohen, 1988, pp. 285-288).

3.3 Study 2: The accuracy of centrality indices

This study aims to answer the second research question: how accurate (and inaccurate) in terms of Type I and statistical power, is each centrality-detection index in flagging raters, particularly in light of the interactions observed in the first study? I applied hypothesis testing procedures to each rater on each rater centrality index. That is, I specified a critical value for each index and declared each rater to be an "effect" or "non-effect" rater depending on whether the value of the rater centrality index was more extreme (effect) or less extreme (non-effect) than the specified critical values. To determine the accuracy and inaccuracy of these hypothesis test results for each rater centrality index, I compared the generating rater type (effect or non-effect) to the rater type classification based on the hypothesis test results. This comparison resulted in one of four outcomes for each rater: (a) correct classification as an effect rater (a true effect detection, depicted by statistical power), (b) correct classification as a non-effect rater (a true non-effect conclusion), (c) incorrect classification as an effect rater (a false positive, depicted by the Type I error rate), or (d) incorrect classification as a non-effect rater (a false negative,

depicted as the Type II error rate). Because the results of a previous study suggested that each index is differentially predictive of rater centrality across double-scoring rates and centrality strength, I report the findings of this study stratified on the levels of those variables.

It is important to note that, in the rater classification process, I relied on two types of critical values. The first type of critical value is based on a hypothesis testing framework. That is, for some of the rater centrality indices, I posited a parametric form for the null distribution of the rater centrality index and created critical values to set a proportion of the area of that null distribution as the rejection region for the hypothesis tests. These hypothesis testing critical values were determined for and applied to the standard deviation of assigned scores and the correlation-based indices (i.e., the point-measure correlation, the measure-residual correlation and the expected-residual correlation), because the value of these indices can be assumed to follow either an F distribution or a normal distribution. In both cases, the null hypothesis assumes no difference between a particular rater and the null distribution, which is based on non-effect raters, so raters for whom the null hypothesis is rejected are flagged as exhibiting the centrality effect.

In the hypothesis tests, values of correlation-based indices were first transformed using Fisher's Z calculated as

$$r' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

where r is correlation coefficient. We assume r' is approximately normally distributed around ρ', the transformed mean value of correlations for the sample of raters in an experimental condition with which a given rater is associated. From this it follows that

$$z = \frac{r' - \rho'}{S_e}$$

is a standard normal curve deviate. The standard error is calculated as

$$S_e = \sqrt{\frac{1}{total \big/ 1000 * (rater\_n - 1)} - 3}$$

where *total* is the number of ratings assigned by all raters in each condition combination of rater type, double-scoring rate, centrality strength, number of raters and ratees, and *rater_n* is the number of raters in which experimental condition a given rater is located. These calculated *z* statistics were compared to the critical values in order to conduct the rater's classification hypothesis test. I expected the index values for the effect raters to be negative or smaller than those for the non-effect raters. That is, I compared the observed *z* against the corresponding critical value at alpha level of 0.05, i.e, the critical value of -1.64 for one-tailed hypothesis test. A similar process was followed for the standard deviation of assigned scores (STD). First, each standard deviation was formulated as $F = \frac{S^2_{effect}}{S^2_{non-effect}}$. Second, critical values were determined based on the degrees of freedom for the numerator and the denominator at each group level. For effect raters, the degree of freedom is calculated as

$$df_{effect} = rating\_count - 1$$

where *rating_count* is the number of ratings assigned by a corresponding rater at each group level. For non-effect raters, the degree of freedom is calculated as

$$df_{non-effect} = \frac{total}{1000 * (rater\_n - 1)} - 1.$$

Third, each rater's F statistic was compared to this critical value, and the rater was declared to be an effect or non-effect rater. In this case, I expect the standard deviation of assigned scores for the effect raters to be smaller than those for the non-effect raters.

The second type of critical value I utilized in the study is not based on a hypothesis testing framework. Rather, as is often the case in applications of some of these indices, I adopted a rule-of-thumb critical value. This type of test applied to the standardized outfit (ZMSU) and infit (ZMSW) statistics, where, in practice, critical values for these indices that are commonly accepted in the literature range are +2.0 and -2.0. That is, a simulated rater was flagged if the observed value of standardized mean-square fell outside of this range.

For each simulated rater, both effect and non-effect raters, I determined the accuracy of the rater classification after applying the relevant critical value and then determined the proportion of incorrect classifications for non-effect raters (a Type I error or false positive) and the proportion of incorrect classifications for effect raters (a Type II error or false negative rate) as well as the inverse of these values, which depict the proportion of correct classifications for each group of raters.

3.4 Study 3: The impact of rating data collection design on rater classification accuracy by centrality indices

The third study addressed the third research question: how do the features of the data collection design (i.e., the independent variables including the level of centrality strength, double-scoring rate, and the number of raters and ratees) influence the rater classification by these centrality-detection indices? I first computed two dichotomous variables for each

centrality-detection index. One of those variables records the outcome of rater classification (1=correct, 0=incorrect) for non-effect raters. The other variable records the outcome of rater classification (1=correct, 0=incorrect) for effect raters. Then, I used logistic regressions to determine how the features of the data collection design influences these two outcome variables for each centrality-detection index. That is, I conducted a logistic regression separately for effect raters and for non-effect raters by specifying dichotomous rater classification accuracy (correct classification versus incorrect classification) as a function of several predictor variables (i.e., the double-scoring rate, the centrality strength, the number of raters, and the number of ratees) separately for each rater centrality index. For each index, the analyses commenced with a saturated model under a full factorial design to capture all possible effects and then used a stepwise procedure for model reduction.

To facilitate interpretation of the results, I first centered the independent variables (i.e., the experimental parameters except for the double-scoring rate) on their grand means, and then created cross-product interaction terms using the mean-centered experimental parameters. After parameters were estimated, I multiplied each estimate by its standard deviation (summarized in Table 3.4-1) and took the exponential of the products to get odds ratios. This is equivalent to standardizing each predictor terms prior to the logistic regressions. I used odds ratios as indicators of effect sizes, and applied Monahan's (2007) suggestion on cut-off values for large effect sizes being odds ratios greater than 1.89 (or less than 0.53, for negative relationships) and moderate effect sizes being an odds ratio greater than 1.53 (or less than 0.65, for negative relationships). The interpretations focus on either the main effects or the highest-order interaction effects with minimum of medium effect sizes.

Table 3.4-1: *Standard deviations of mean-centered experimental parameters and their interactions*

Mean-centered Variable	Std Dev
CS	0.28
RATERN	82.12
RATEEN	816.50
DSR*CS	0.20
DSR*RATERN	58.36
DSR*RATEEN	580.23
DSR*CS*RATERN	16.31
DSR*CS*RATEEN	162.18
DSR*CS*RATERN*RATEEN	13318.63
DSR*RATERN*RATEEN	47650.18
CS*RATERN*RATEEN	18741.91
CS*RATERN	22.95
CS*RATEEN	228.22
RATERN*RATEEN	67053.10

Note: CS = centrality strength, DSR = double-scoring rate, RATERN = number of raters, RATEEN = number of ratees. For each centered variable, N = 9,672,000 and Mean = 0.00.

4 MANUSCRIPT 1: THE RELATIONSHIP BETWEEN RAW SCORES AND SEVERAL LATENT-TRAIT INDICES FOR DETECTING RATER CENTRALITY EFFECT

4.1 Abstract

This simulation study compares several centrality-detection indices to determine which of them differentiate effect raters from non-effect raters using data designs similar to those in large scale performance assessment settings. I generated datasets by varying influential factors including the strength of centrality effect, the number of raters, the number of ratees, and the proportion of ratees assigned a second score. The results reveal that the measure-residual correlation, the expected-residual correlation, and the standardized deviation of assigned scores perform better than the point-measure correlation. The mean-square fit statistics, traditionally viewed as potential indicators of rater centrality, perform poorly in terms of differentiating central raters from normal raters. Along with the rater slope index, the mean-square fit statistics did not appear to be sensitive to the rater centrality effect.

4.2 Introduction

Performance assessments are used extensively in educational testing, particularly in content areas that are difficult to measure with multiple-choice items, and ratings assigned by a human are commonly used to judge the quality of the assessment products (Landy & Farr, 1980). Unfortunately, performance ratings are prone to various types of systematic and random error, potentially rendering the associated scores inaccurate as indicators of the student's true performance. That is, raters may exhibit *rater effects* that serve as sources of error in the

performance ratings. Therefore, identifying aberrant raters successfully is an important step in evaluating the psychometric quality of ratings in terms of validity and reliability. Previous research has identified several ways that rater effects are manifested in ratings, and researchers have developed a variety of criteria under different methodological frameworks (e.g., classical test theory, analysis of variance, regression-based analysis, generalizability theory, and Rasch measurement and item response theory) to evaluate the psychometric quality of performance ratings (Saal, et al., 1980). However, supporting simulation studies and operational applications of these methods have remained sparse.

Previous research concerning rater effects has identified several ways that raters introduce error into ratings, including severity/leniency, centrality/extremism, and inaccuracy (Saal, et al., 1980). This study focuses on detecting the centrality effect. A rater who exhibits the centrality effect overuses the middle categories of the rating scale while avoiding extreme categories. Hence, the centrality effect induces a reduction of the variation of assigned ratings. In addition, centrality results in accurate ratings in the central range of the ability continuum, but overestimates of ratee proficiency for non-proficient ratees and underestimates of ratee proficiency for highly proficient ratees. Clearly, the centrality effect will manifest itself in the standard deviation of the observed ratings. However, the standard deviation may be a poor choice as a rater effect index because it is inflated when random error exists in the ratings. As a result, it would be difficult to determine whether raters who produce a small standard deviation are engaging in centrality or are simply accurate raters. In addition, many operational analyses of rater effects employ latent trait modeling procedures (e.g., item response theory), and numerous

indices exist within those frameworks that might be better suited as indicators of the centrality effect.

One family of such indices includes the various mean-square fit indices that are commonly reported by commercial latent trait software, such as Winsteps (Engelhard, 1992, 1994). When a rater exhibits the centrality effect in ratings, we may expect the fit statistics to deviate away from 1.0 and the absolute value of standardized fit statistics to be larger than 2.0. We might expect the unweighted mean-square fit statistic to be more sensitive to centrality because it is more heavily influenced by large residuals for mismatched raters and ratees. We could expect these cases to produce larger residuals (i.e., a normal rater would produce ratings of high and low ability ratees that have large residuals). However, prior research has indicated that rater centrality may manifest itself inconsistently in these indices, and may, therefore, cause analysts to incorrectly conclude that accurate raters are engaging in central rating patterns (Wolfe, et al., 2000). Hence, in this article, I evaluated four additional indices to detect rater centrality: the score-measure correlation (also known as the point-measure correlation), the measure-residual correlation, the expected-residual correlation (Wolfe, 2004b, 2005), and a rater slope index (Wolfe, 1998a).

The score-measure correlation is also called the point-measure correlation, denoted as "PTMEA" in *Winsteps* output. In rating data, it correlates Rasch measures of ratees with ratings assigned by raters, and is computed by the formula

$$r_{score,measure} = \frac{1}{n-1}\sum_{n=1}^{N}\left(\frac{X_{nr}-\overline{X}_{r}}{S_{X_r}}\right)\left(\frac{\theta_n-\overline{\theta}}{S_{\theta}}\right),$$

where $\frac{X_{nr} - \overline{X}_r}{S_{X_r}}$ is the standard score of rating assigned to ratee *n* by rater *r*, $\overline{X}_r$ is the sample mean of ratings assigned by rater r, S_{X_r} is the sample standard deviation of ratings assigned by rater *r*, $\frac{\theta_n - \overline{\theta}}{S_\theta}$ is the standard score of Rasch measure of ratee *n*, and S_θ is the standard deviation of Rasch measure of all ratees. In typical item analyses using Rasch measurement, we use item correlations as an immediate check that the response level scoring is consistent with the scoring of other items (i.e., the rank ordering of ratees is the same for the item as for the estimated ratee measures). In the case of studies of rater centrality, we can evaluate the score-measure correlation to see if assigned ratings to ratees by a particular rater are consistent with the estimated ratee measures, which would be based on the ratings assigned by all raters. The value of a score-measure correlation from raters with centrality should be less than one obtained from raters without centrality. That is, a highly positive value of the score-measure correlation indicates a strong association between ratees' ability measure and ratings assigned by a rater, and raters with the centrality effect are expected to slightly weaker correlations due to the attenuation of the assigned score distribution. Additionally, the score-measure correlation (and the analogous point-biserial correlation) is likely to be influenced by other features of the data (e.g., the targeting of the rater on the person sample and the distribution of the personal sample) as well as other rater effects (e.g., rater inaccuracy). As a result, the score-measure correlation may have limited power for detecting the centrality effect.

The expected-residual correlation depicts the association of estimated ratings and residuals, $r_{\exp,res}$. Specifically, the expected-residual correlation is based on the notion that the

raw residual, $x_{nr} - E_{nr}$, produced by all ratings assigned by rater *r* who exhibits the centrality effect is positive for ratees of low ability (i.e., x_{nr} is greater than E_{nr}) and negative for ratees of high ability (i.e., x_{nr} is less than E_{nr}). As a result, when the centrality effect exists, a scatter plot of the expected scores (X axis) and residuals (Y axis) should exhibit a negative association. A value of the expected-residual correlation that approaches its upper limit of 1.00 indicates that the way a rater utilizes the rating scale is uniform across the ability continuum. Similarly, the measure-residual correlation correlates Rasch measures of ratees and residuals, $r_{measure,res}$, and this index behaves in a similar manner when the centrality effect exists

The rater slope index, also called rater discrimination, may be sensitive to rater centrality, In the Rasch measurement software *Winsteps*, the discrimination index is not an estimated parameter contained in the model, although the software will output a crude estimate of what that parameter would be, if it were to be estimated for the sake of allowing an analyst to evaluate the reasonableness of the common slopes assumption inherent in the Rasch model. In applications of the Rasch model, the slope is set to a constant value of 1.0. On the other hand, the empirical slope is estimated in applications of a 2-PL IRT model so that the slope differs for each item (or, in this case, rater). The reported values of rater slope that are output by *Winsteps* are a first approximation to the precise value of a_r obtained from the Newton-Raphson estimation equation (Wright & Masters, 1982, pp. 72-77):

$$\hat{a}_r = 1 + \left[\frac{\sum_{n=1}^{N} (X_{nr} - P_{nr})(\theta_n - b_r)}{\sum_{n=1}^{N} P_{nr}(1 - P_{nr})(\theta_n - b_r)^2} \right],$$

where X_{nr} is the rating assigned to ratee *n* by rater *r*, P_{nr} is the probability for ratee *n* rated by rater *r* to get the item correct, θ_n is the Rasch measure for ratee *n*, and b_r is the severity of ratee *r*. Because a rater exhibiting the centrality effect may assign accurate ratings in the central range of the ability continuum and will overestimate ratee proficiency for non-proficient ratees and underestimate ratee proficiency for highly proficient ratees, the centrality effect rater could have, on average, lower discrimination estimates than non-effect raters. Therefore, small values of the slope index could be a symptom of the centrality effect. The commonly acceptable lower acceptable limit for the slope index is 0.5, and raters with lower values would be flagged as candidates for the centrality effect.

The purpose of this simulation study is to compare these centrality-detection indices to determine which of them differentiate effect raters from non-effect raters using data designs similar to those in large scale performance assessment settings. In this paper, I generated datasets by varying influential factors including the number of raters, the number of ratees, and the proportion of ratees assigned a second score. The results reveal that the measure-residual correlation, the expected-residual correlation, and the standardized deviation of assigned scores perform better than the point-measure correlation and mean-square fit indices.

4.3 Methods

4.3.1 Simulation Design

I generated simulated ratings on a six-point scale using *ConQuest* 2.0 (Wu, et al., 2007) based on a unidimensional Rasch model that contains a single measurement facet (i.e., ratings assigned by raters to a single assessment item). I specified rating scale step threshold parameter

values to create an approximately normal distribution of ratings for each simulated rater. In order to avoid confounding rater centrality and rater severity/leniency effects, all rater severity parameters were set to zero for all raters during data generation. That is, data sets were generated according to a Rasch rating scale model in which the item "difficulty" parameter was replaced by a rater "severity" parameter which was set to a constant value of zero.

Starting with these "non-effect" data, a single simulated rater exhibiting a centrality effect was generated by shrinking the underlying ratee ability to one of four ***centrality-strength*** levels (0.10, 0.35, 0.60, and 0.85; i.e., the variance of the ability distribution is shrunk to a minimum of 10% and to a maximum of 85% of that of the non-effect rater) for a single rater, and these effect ratings were added into the simulated data file in order to create a ***rater-type*** variable (namely focused non-effect or normal rater and manipulated effect rater). The relationship between the centrality strength and the true ability distribution can be expressed as a variance ratio defined as,

$$\text{Variance ratio} = \frac{\sigma^2_{effect}}{\sigma^2_{true}},$$

where σ^2_{effect} is the variance of the ability distribution for the ratees whose responses are rated by effect raters and σ^2_{true} the variance of ability distribution for the ratees whose responses are rated by non-effect raters. Table 4.3-1 translates the centrality strength values into the equivalent empirical variance ratios. For example, a centrality-strength value of 0.1 causes the variance of the ability distribution for an effect rater to be 66% of the variance of the ability distribution

produced by normal raters. From a centrality-strength value of 0.85, we can consider the centrality with a rater to be trivial as the variance ratio is close to 1.

Table 4.3-1: *Variance ratio vs. centrality strength*

Centrality Strength	Variance Ratio
0.10	0.66
0.35	0.73
0.60	0.86
0.85	1.03

I generated data according to these procedures for three ***rater-sample sizes*** (50, 100, and 250) and three ***ratee-sample sizes*** (1000, 2000, and 3000). In all data sets, single raters were randomly assigned to ratees, and a second rating was randomly assigned to either 10% or 100% of the ratees (i.e., the ***double-scoring rate***), two common double-scoring designs observed in operational settings. Hence, the resulting data is not a full rater-by-ratee design (i.e., every rater does not rate every ratee); with random rater-response assignment, the simulated rating data sets contained missing values. Thus, the number of assigned ratings varied across raters, which created a ***rating-count*** variable which was recorded because this variable could affect the measurement error of the centrality indices and hypothesis tests based on those indices. As the simulated data is graphed in Figure 4.3-1, raters' average rating-count decreases as the rater sample size enlarges and increases as the ratee sample size diminishes. On average, each rater assigns more ratings when all responses receive a second score comparing to when 10% of responses receive a second score. For each combination of these independent variables, I generated 1000 data sets, resulting in 72,000 data sets (i.e., 1000 replications × 4 centrality-strength levels × 2 double-scoring rates × 3 rater-sample sizes × 3 ratee-sample sizes).

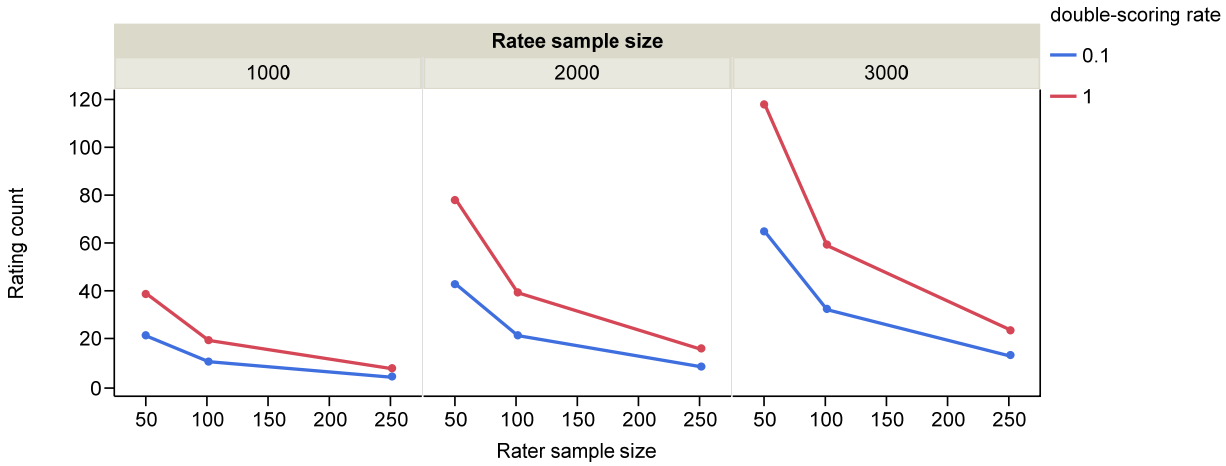

Figure 4.3-1: Raters' average rating count by the double-scoring rate and the number of raters and ratees

4.3.2 Analysis

After the data sets were generated, I performed the following activities. First, parameters were estimated for each simulated data set for the Rasch rating scale model (Andrich, 1978) using the *Winsteps* software (Linacre, 2009b). Second, each of the six rater centrality indicators (standard deviation of assigned scores, mean-square fit indices, score-measure correlation, measure-residual correlation, expected-residual correlation, and rater slope) were computed for each rater (effect and non-effect). These index values s along with the experimental parameters are the data source for the ensuing analyses. Except for the double-scoring rate, all other experimental parameters were treated as continuous variables because the chosen values of these continuous variables are based on empirical trails with a consideration on continuous distributions, which allows for accurately depicting trends across levels of these variables. In the subsequent analyses, I treated the difference of the values of each rater centrality index between

rater types as the dependent variable in a separate analysis of variance method[2] (ANOVA). The remaining variables were treated as the independent variables. For each centrality-effect index, the ANOVA analyses commenced with a saturated model under a full factorial design to capture all possible effects, and then relied on variable elimination procedures to a reduced model containing only effect terms with meaningfully large effect sizes. I utilized the stepwise procedure for model reduction because the simulated data sets were large enough to be assumed generalizable to other samples from similar scoring settings. I used η^2 as indicators of effect sizes, with the cut-off for practical significance being an η^2 greater than 0.06 (Cohen, 1988, pp. 285-288).

4.4 Results

In Table 4.4-1, we summarize the significant model effects along with their effect sizes according to the ANOVA results from reduced models[3]. Because of the large sample size of the simulated data set, all the models were statistically significant. These model effects accounted for up to 20% of variation in the values of centrality-detection indices. The model effects exhibiting a minimum of moderate effect size (i.e., η^2 larger than 0.06) indicate that the

[2] Upon checking the ANOVA assumptions, i.e., independence, normality and homoscedasticity, I found various levels of departure from the latter two assumptions. However, Lindman (1974) shows that the *F* statistic of the ANOVA test is quite robust against violations of these assumptions.

[3] The summaries reported here include only the results from the reduced models determined through stepwise ANOVA process and with all effect sizes of "zero" eliminated (i.e., model effects with eta-squared of zero are excluded from the final models).

associated centrality-detection indices were sensitive to difference between rater types and thus were indicative of the usefulness of the indices for differentiating effect and non-effect raters. To better illustrate these model effects, in the following graphs, I plotted separate lines for rater types instead of difference between them.

Table 4.4-1: *Hypothesis testing and effect size from the ANOVA results*

Centrality-detection Index	R^2	Model effect	F	Sig	η^2
Point-measure correlation	0.19	DSR	8288.229	<.0001	0.09
		DSR*CS	3856.667	<.0001	0.04
		CS	45.123	<.0001	0.00
Measure-residual correlation	0.05	CS	3456.066	<.0001	0.05
Expected-residual correlation	0.05	CS	3431.932	<.0001	0.05
STD	0.09	CS	7037.547	<.0001	0.09

Note: CS=centrality strength, DSR=double-scoring rate

For the point-measure correlation, two main effects, ***double-scoring rate*** and ***centrality-strength*** produced η^2 of 0.09 and 0.00), and their double-scoring rate-by-centrality-strength interaction had a small effect size (η^2 = 0.04) that approached my definition of a moderate effect size, so we chose to summarize the interaction effect. The η^2 value of 0.04 associated with the double-scoring rate-by-centrality-strength interaction indicated that 4% of the variation in the difference between the values of the point-measure correlations for effect and non-effect raters was explained by the interaction. Figure 4.4-1 displays this interaction graphically. When only 10% of the responses were double scored the values of the effect and non-effect rater point-measure correlations are fairly close to each other. That is, the point-measure correlation did not differentiate the effect rater from the non-effect rater. However, when all responses were double scored (shown in the right panel of Figure 4.1), as the magnitude of centrality increased, the

value of the point-measure correlation for effect rater (red line) decreased relative to the value of that index for non-effect rater (blue line).

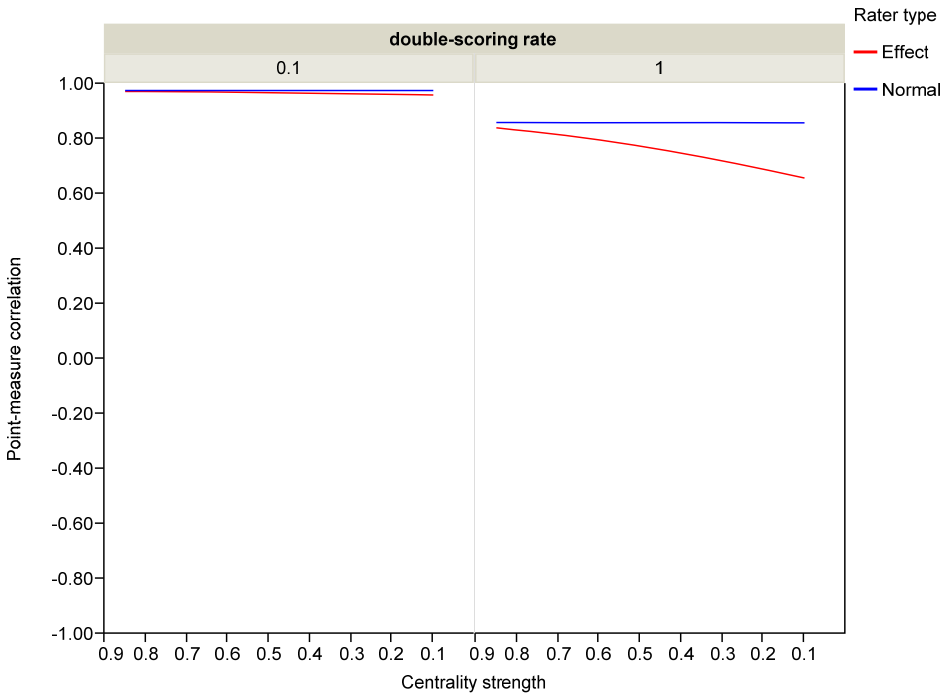

Figure 4.4-1. The double-scoring rate-by-centrality-strength interaction on the point-measure correlation

For the measure-residual correlation, the expected-residual correlation and the standard deviation of assigned scores, the ***centrality-strength*** main effect produced η^2 of 0.05, 0.05 and 0.09 respectively. As shown in Figures 4.4-2, 4.4-3, and 4.4-4, as the centrality effect became stronger, the value of these three rater centrality-detection indices decreased for the effect rater relative to the values of those indices for non-effect rater. In all cases, the gap between the values of the rater centrality index increases between the two rater types as the centrality strength increases, suggesting that these indices are indeed sensitive to the rater centrality effect and may be useful in differentiating effect and non-effect raters.

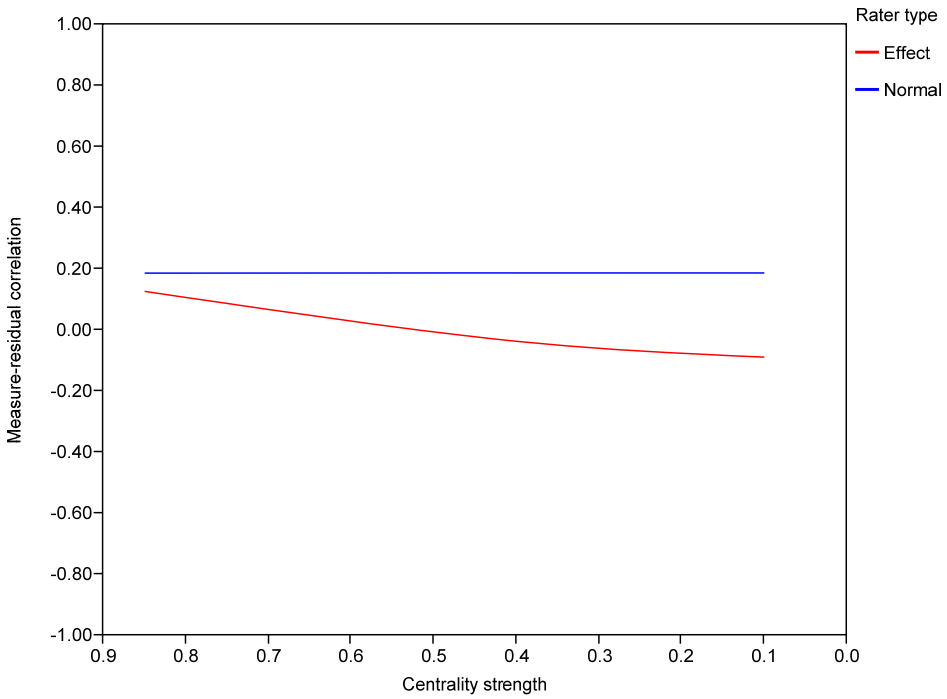

Figure 4.4-2. The centrality strength effect on the measure-residual correlation

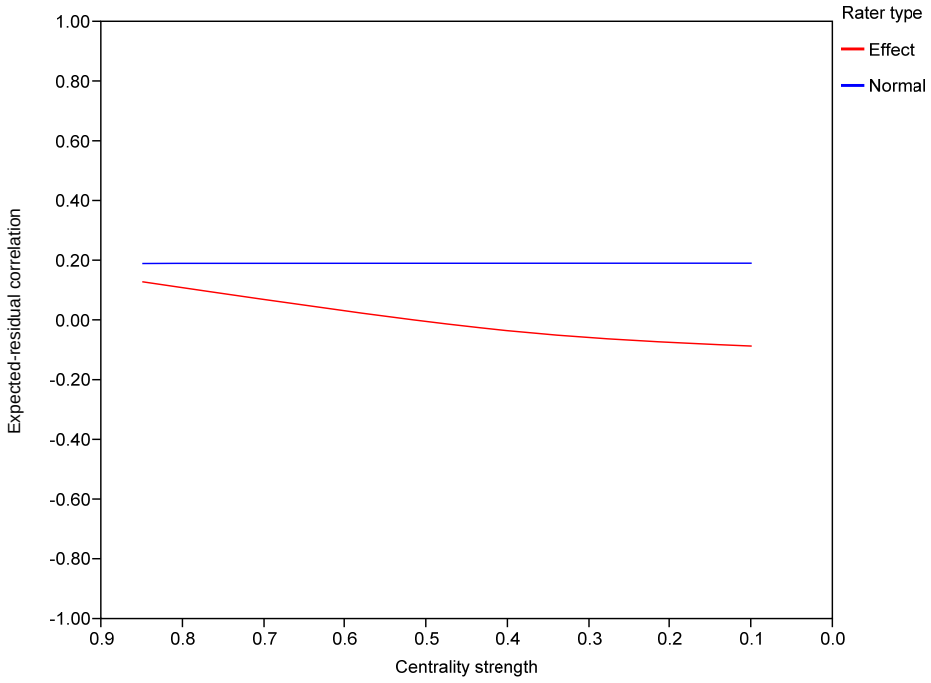

Figure 4.4-3. The centrality strength effect on the expected-residual correlation

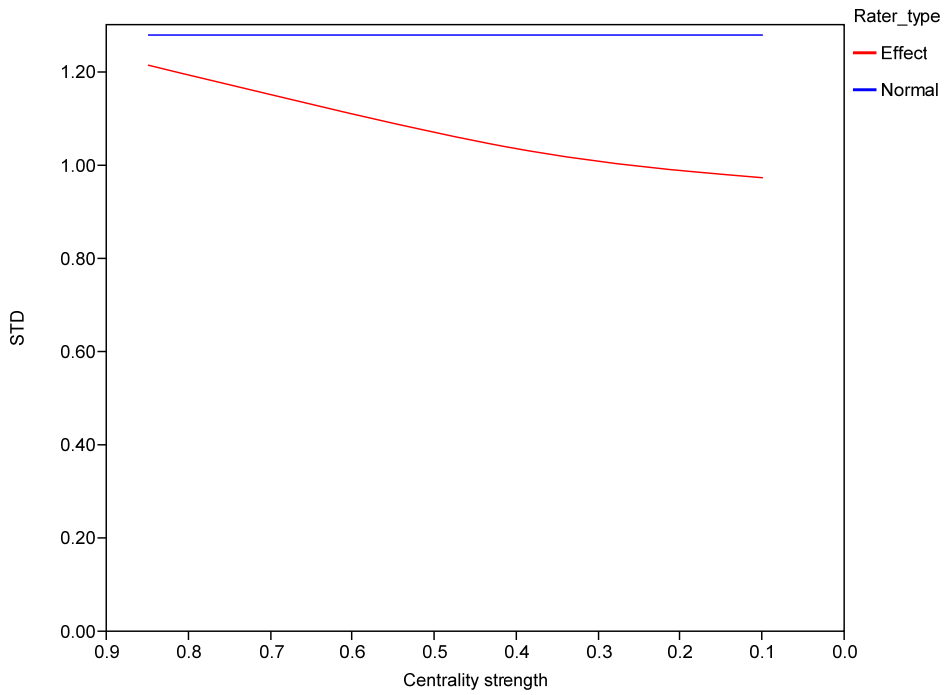

Figure 4.4-4. The centrality strength effect on the standard deviation (STD) of assigned scores

The mean-square fit (both standardized and non-standardized) and rater slope indices are not summarized in Table 4.4-1 because none of these indices produced a moderate effect size on any of the independent variables in the ANOVA. Specifically, the largest effect size for the standardized weighted mean-square fit statistic was $\eta^2 = 0.02$ for the double-scoring rate variable, and for the non-standardized was $\eta^2 = 0.01$ for the centrality strength variable. Similarly, for the standardized unweighted mean-square fit statistic was $\eta^2 = 0.03$ for the double-scoring rate variable, and for the non-standardized was $\eta^2 = 0.01$ for the centrality strength variable. For the rater slope index, the largest effect size was $\eta^2 = 0.01$ for the double-scoring rate variable. Generally, the small effect sizes for the mean-square fit statistics and the

rater slope index suggests that these two types of index were not sensitive to the difference between rater types and thus not suitable for identifying the rater centrality effect.

4.5 Conclusions

The results indicate that the indices that have been traditionally viewed as potential indicators of rater centrality perform quite poorly in terms of differentiating central raters from normal raters. Specifically, the centrality strength main effect for both the unweighted and weighted mean-square fit indices accounted for only about 1% of the observed variance in the difference between the values of these indices for simulated effect and non-effect raters. Although the standardized versions of these two fit indices performed slightly better, neither the standardized unweighted mean-square nor the standardized weighted mean-square fit indices achieved a moderate effect size in the analyses. In fact, the only index that performed worse than the fit indices at differentiating central and normal raters was the rater slope index, which has never been subjected to scrutiny as a potential rater centrality indicator.

On the other hand, two relatively new and related indices, the measure-residual correlation and the expected-residual correlation, both performed better at differentiating simulated raters who exhibit the centrality effect from those who do not. However, both of these indices produced only marginally larger effect sizes; ones that only approached the definition of moderately large effect size (i.e., η^2 was about 0.05 in both cases for the centrality strength main effect). In fact, neither of these indices performed as well as the assigned score standard deviation, the only raw score index I evaluated, which produced an effect size of 0.09 for the centrality strength main effect.

The results for the point-measure correlation as a rater centrality index are somewhat surprising. While I expected the point-measure correlation to be influenced by rater centrality due to the attenuation of the scores associated with any process that would reduce their variance, the fact that this index produced the largest effect size for differentiating effect and non-effect raters and that it was the only index that produced a meaningfully large two-way interaction effect were unexpected results. Due to the fact that rater centrality does not really alter the rank ordering of ratees, I expected the point-measure correlation to be minimally influenced by the introduction of rater centrality. What I did not expect was the fact that the process of reducing the variance of the assigned scores, which produces a significant number of "tied" scores when the continuous ability distribution is divided into polytomous discrete categories, would have such a strong attenuating impact on the correlation between ability estimates and assigned scores. It is also somewhat surprising that the ability of the point-measure correlation to differentiate effect and non-effect raters was strongly influenced by missing data, as evidenced by the moderate centrality strength-by-double-scoring rate interaction effect. Specifically, when the double-scoring rate was reduced from 100% of the simulated ratees to only 10% of the simulated ratees, the point-measure correlation did not differentiate effect from non-effect raters. Hence, its value as a rater centrality index may be limited by the sparseness of the data matrix.

I was also surprised by the fact that most of the experiment parameters (namely double-scoring rate, rater sample size, and ratee sample size) had almost no influence on the capacity of each of the rater centrality indices I investigated to differentiate effect and non-effect raters. Generally, the more severe the centrality effect was, the more sensitive to the difference between

rater types a centrality-detection index was with the exception of the two-way interaction that the point-measure correlation exhibited involving double-scoring rate.

Further study is warranted concerning each of these rater effects. Specifically, additional simulation work should be conducted to determine the Type I and Type II error rates associated with the dichotomous decisions to flag a rater as an effect or non-effect rater. In operational settings, a critical value would be chosen for each of these rater centrality indices, and raters would be categorized as either exhibiting or not exhibiting a large enough rater centrality effect to warrant intervention (e.g., retraining, rescoring, etc.). In addition to evaluating the Type I and Type II error rates, additional simulation work can focus on the degree to which these dichotomous categories that raters are placed into are accurate.

A limitation of this simulation study is the fact that I isolated rater centrality. That is, the only rater effect that I introduced into the simulated ratings was rater centrality. Real data are likely to contain a mix of rater effects, and each rater centrality index may be sensitive to multiple types of rater effects. Hence, additional simulation work should be carried out to determine how well these indices perform when multiple rater effects exist in the ratings and how well each index differentiates one rater effect from another. Similarly, due to the fact that the simulated data are extremely "clean," it is difficult to determine how well the simulation results will generalize to the "messy" data that are typical of operational contexts. That is, the data were modeled to fit the Rasch model, and other potential sources of model-data misfit were omitted from the data generation process (e.g., multidimensionality, rater drift over time, etc.). Hence, application of these rater centrality-detection indices to real data sets and to more complex simulated data may provide a useful test of the generalizability of the results.

Another limitation is that I did not control the number of ratees per rater and made the rating-count variable vary across raters with random assignment technique in the data generating process. Because this variable could affect the measurement error of the centrality indices and any associated hypothesis test based on those indices, it is hard to draw any conclusion on the appropriate number of ratees per rater for a particular centrality index to work well from this simulation study. Therefore, rather than letting it change with rater-response random assignment in the simulation process, the ideal case would be making the rating-count variable manipulatable.

I believe that it is safe to conclude, however, that mean-square fit indices and their standardized transformations are a poor choice as indicators of rater centrality. Although these indices have been cited as potential indicators of rater centrality by several authors, e.g., Engelhard (1994), Myford and Wolfe (2004), etc., the simulation results indicate that the mean-square fit indices do a poorer job of differentiating effect and non-effect raters than the measure-residual correlation and the expected-residual correlation. In fact, the raw score standard deviation and the point-measure correlation seem to be considerably more sensitive to the existence of the rater centrality effect than are the mean-square fit indices. However, the point-measure correlation seems to be much less sensitive when the data matrix is sparse. Further research should be conducted to determine the error rates and classification accuracy associated with various critical values that could be adopted for each of these indices before electing one of the indices as the “best candidate” for identifying rater centrality.

4.6 References

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

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

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Linacre, J. M. (2009b). WINSTEPS Rasch measurement computer program (Version 3.68.0). Chicago, IL: Winsteps.com.

Lindman, H. R. (1974). *Analysis of variance in complex experimental designs*. San Francisco: W. H. Freeman & Co. Hillsdall, NJ USA: Erlbaum.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

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.

Wolfe, E. W. (1998). *A two-parameter logistic reader model (2PLRM): Detecting reader harshness and centrality.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

Wolfe, E. W. (2005). Identifying Rater Effects in Performance Ratings. In S. Reddy (Ed.), *Performance Appraisals: A Critical View* (pp. 91-103). Hyderabad, India: ICFAI University Press.

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects with a multi-faceted Rasch rating scale model. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147-164). Stamford, CT: Ablex.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: generalised item response modeling software (Version 2.0). Camerwell, Victoria: ACER Press.

5 MANUSCRIPT 2: THE ACCURACY OF CENTRALITY INDICES

5.1 Abstract

This study compares the Type I error and the statistical power of several centrality-detection indices. I generated simulated datasets by varying the strength of centrality effect, the number of raters, the number of ratees, and the proportion of ratees assigned a second score. The results show that all of these indices provided reasonable protection against Type I errors when all responses were double scored, and that higher statistical power was achieved when responses were 100% double scored in comparison to only 10% being double scored. With a consideration on balancing both Type I error and statistical power, I recommend the measure-residual correlation and the expected-residual correlation for detecting the centrality effect. I suggest using the point-measure correlation only when responses are 100% double scored.

5.2 Introduction

Performance assessments are used extensively in educational testing, particularly in content areas that are difficult to measure with multiple-choice items, and ratings assigned by a human are commonly used to judge the quality of the assessment products (Landy & Farr, 1980). Unfortunately, performance ratings are prone to various types of systematic and random error, potentially rendering the associated scores inaccurate as indicators of the student's true performance. That is, raters may exhibit *rater effects* that serve as sources of error in the performance ratings. Therefore, identifying aberrant raters successfully is an important step in

evaluating the psychometric quality of ratings in terms of validity and reliability. Previous research has identified several ways that rater effects are manifested in ratings, and researchers have developed a variety of criteria under different methodological frameworks (e.g., classical test theory, analysis of variance, regression-based analysis, generalizability theory, and Rasch measurement and item response theory) to evaluate the psychometric quality of performance ratings (Saal, et al., 1980). However, supporting simulation studies and operational applications of these methods have remained sparse.

Previous research concerning rater effects has identified several ways that raters introduce error into ratings, including severity/leniency, centrality/extremism, and inaccuracy (Saal, et al., 1980). This study focuses on detecting the centrality effect. A rater who exhibits the centrality effect overuses the middle categories of the rating scale while avoiding extreme categories. Hence, the centrality effect induces a reduction of the variation of assigned ratings. In addition, centrality results in accurate ratings in the central range of the ability continuum, but overestimates of ratee proficiency for non-proficient ratees and underestimates of ratee proficiency for highly proficient ratees. Clearly, the centrality effect will manifest itself in the standard deviation of the observed ratings. However, the standard deviation may be a poor choice as a rater effect index because it is inflated when random error exists in the ratings. As a result, it would be difficult to determine whether raters who produce a small standard deviation are engaging in centrality or are simply accurate raters. In addition, many operational analyses of rater effects employ latent trait modeling procedures (e.g., item response theory), and numerous indices exist within those frameworks that might be better suited as indicators of the centrality effect.

One family of such indices includes the various mean-square fit indices that are commonly reported by commercial latent trait software, such as Winsteps (Engelhard, 1992, 1994). When a rater exhibits the centrality effect in ratings, we may expect the fit statistics to deviate away from 1.0 and the absolute value of standardized fit statistics to be larger than 2.0. We might expect the unweighted mean-square fit statistic to be more sensitive to centrality because it is more heavily influenced by large residuals for mismatched raters and ratees. We could expect these cases to produce larger residuals (i.e., a normal rater would produce ratings of high and low ability ratees that have large residuals). However, prior research has indicated that rater centrality may manifest itself inconsistently in these indices, and may, therefore, cause analysts to incorrectly conclude that accurate raters are engaging in central rating patterns (Wolfe, et al., 2000). Hence, in this article, I evaluated four additional indices to detect rater centrality: the score-measure correlation (also known as the point-measure correlation), the measure-residual correlation, the expected-residual correlation (Wolfe, 2004b, 2005), and a rater slope index (Wolfe, 1998a).

The score-measure correlation is also called the point-measure correlation, denoted as "PTMEA" in *Winsteps* output. In rating data, it correlates Rasch measures of ratees with ratings assigned by raters, and is computed by the formula

$$r_{score,measure} = \frac{1}{n-1}\sum_{n=1}^{N}\left(\frac{X_{nr}-\overline{X}_{r}}{S_{X_r}}\right)\left(\frac{\theta_n-\overline{\theta}}{S_{\theta}}\right),$$

where $\frac{X_{nr}-\overline{X}_{r}}{S_{X_r}}$ is the standard score of rating assigned to ratee *n* by rater *r*, $\overline{X}_{r}$ is the sample mean of ratings assigned by rater r, S_{X_r} is the sample standard deviation of ratings assigned by

rater *r*, $\frac{\theta_n - \bar{\theta}}{S_\theta}$ is the standard score of Rasch measure of ratee *n*, and S_θ is the standard deviation of Rasch measure of all ratees. In typical item analyses using Rasch measurement, we use item correlations as an immediate check that the response level scoring is consistent with the scoring of other items (i.e., the rank ordering of ratees is the same for the item as for the estimated ratee measures). In the case of studies of rater centrality, we can evaluate the score-measure correlation to see if assigned ratings to ratees by a particular rater are consistent with the estimated ratee measures, which would be based on the ratings assigned by all raters. The value of a score-measure correlation from raters with centrality should be less than one obtained from raters without centrality. That is, a highly positive value of the score-measure correlation indicates a strong association between ratees' ability measure and ratings assigned by a rater, and raters with the centrality effect are expected to slightly weaker correlations due to the attenuation of the assigned score distribution. Additionally, the score-measure correlation (and the analogous point-biserial correlation) is likely to be influenced by other features of the data (e.g., the targeting of the rater on the person sample and the distribution of the personal sample) as well as other rater effects (e.g., rater inaccuracy). As a result, the score-measure correlation may have limited power for detecting the centrality effect.

The expected-residual correlation depicts the association of estimated ratings and residuals, $r_{\exp,res}$. Specifically, the expected-residual correlation is based on the notion that the raw residual, $x_{nr} - E_{nr}$, produced by all ratings assigned by rater *r* who exhibits the centrality effect is positive for ratees of low ability (i.e., x_{nr} is greater than E_{nr}) and negative for ratees of high ability (i.e., x_{nr} is less than E_{nr}). As a result, when the centrality effect exists, a scatter plot

of the expected scores (X axis) and residuals (Y axis) should exhibit a negative association. A value of the expected-residual correlation that approaches its upper limit of 1.00 indicates that the way a rater utilizes the rating scale is uniform across the ability continuum. Similarly, the measure-residual correlation correlates Rasch measures of ratees and residuals, $r_{measure,res}$, and this index behaves in a similar manner when the centrality effect exists

The rater slope index, also called rater discrimination, may be sensitive to rater centrality, In the Rasch measurement software *Winsteps*, the discrimination index is not an estimated parameter contained in the model, although the software will output a crude estimate of what that parameter would be, if it were to be estimated for the sake of allowing an analyst to evaluate the reasonableness of the common slopes assumption inherent in the Rasch model. In applications of the Rasch model, the slope is set to a constant value of 1.0. On the other hand, the empirical slope is estimated in applications of a 2-PL IRT model so that the slope differs for each item (or, in this case, rater). The reported values of rater slope that are output by *Winsteps* are a first approximation to the precise value of a_r obtained from the Newton-Raphson estimation equation (Wright & Masters, 1982, pp. 72-77):

$$\hat{a}_r = 1 + \left[\frac{\sum_{n=1}^{N} (X_{nr} - P_{nr})(\theta_n - b_r)}{\sum_{n=1}^{N} P_{nr}(1 - P_{nr})(\theta_n - b_r)^2} \right],$$

where X_{nr} is the rating assigned to ratee *n* by rater *r*, P_{nr} is the probability for ratee *n* rated by rater *r* to get the item correct, θ_n is the Rasch measure for ratee *n*, and b_r is the severity of ratee *r*. Because a rater exhibiting the centrality effect may assign accurate ratings in the central range of

the ability continuum and will overestimate ratee proficiency for non-proficient ratees and underestimate ratee proficiency for highly proficient ratees, the centrality effect rater could have, on average, lower discrimination estimates than non-effect raters. Therefore, small values of the slope index could be a symptom of the centrality effect. The commonly acceptable lower acceptable limit for the slope index is 0.5, and raters with lower values would be flagged as candidates for the centrality effect.

The simulation study described in this manuscript is an extension of my earlier investigation (Chapter 4) on the relative strengths of these centrality-detection indices. In that study, I compared the relative strengths of several centrality-detection indices in differentiating effect rater from non-effect raters. The difference in the values of each rater centrality index between rater types were treated as the dependent variable an analysis of variance, and the rating context variables were treated as the independent variables. The results of that study revealed that the measure-residual correlation, the expected-residual correlation, and the standardized deviation of assigned scores are better able to differentiate rater types than is the point-measure correlation. The mean-square fit statistics, traditionally viewed as potential indicators of rater centrality, perform poorly in terms of differentiating central raters from normal raters. Along with the rater slope index, the mean-square fit statistics did not appear to be sensitive to the rater centrality effect.

The purpose this paper is to further explore the accuracy of these indices, that is, to compare the Type I error and the statistical power rates of these indices through group-level hypothesis testing. The indices that were sensitive to the centrality strength or that at least exhibited an interaction between the centrality strength and the double-scoring rate in the prior

study are included this paper. That is, the point-measure correlation, the measure-residual correlation, the expected-residual correlation and the standard deviation of assigned scores are the central focus of this manuscript. For completeness, the standardized mean-square fit statistics are also included in the comparisons.

5.3 Methods

5.3.1 Simulation Design

I generated simulated ratings on a six-point scale using *ConQuest* 2.0 (Wu, et al., 2007) based on a unidimensional Rasch model that contains a single measurement facet (i.e., ratings assigned by raters to a single assessment item). I specified rating scale step threshold parameter values to create an approximately normal distribution of ratings for each simulated rater. In order to avoid confounding rater centrality and rater severity/leniency effects, all rater severity parameters were set to zero for all raters during data generation. That is, data sets were generated according to a Rasch rating scale model in which the item "difficulty" parameter was replaced by a rater "severity" parameter which was set to a constant value of zero.

Starting with these "non-effect" data, a single simulated rater exhibiting a centrality effect was generated by shrinking the underlying ratee ability to one of four ***centrality-strength*** levels (0.10, 0.35, 0.60, and 0.85; i.e., the variance of the ability distribution is shrunk to a minimum of 10% and to a maximum of 85% of that of the non-effect rater) for a single rater, and these effect ratings were added into the simulated data file in order to create a ***rater-type*** variable (namely focused non-effect or normal rater and manipulated effect rater). The relationship

between the centrality strength and the true ability distribution can be expressed as a variance ratio defined as,

$$\text{Variance ratio} = \frac{\sigma^2_{effect}}{\sigma^2_{true}},$$

where σ^2_{effect} is the variance of the ability distribution for the ratees whose responses are rated by effect raters and σ^2_{true} the variance of ability distribution for the ratees whose responses are rated by non-effect raters. Table 5.3-1 translates the centrality strength values into the equivalent empirical variance ratios. For example, a centrality-strength value of 0.1 causes the variance of the ability distribution for an effect rater to be 66% of the variance of the ability distribution produced by normal raters. From a centrality-strength value of 0.85, we can consider the centrality with a rater to be trivial as the variance ratio is close to 1.

Table 5.3-1: *Variance ratio vs. centrality strength*

Centrality Strength	Variance Ratio
0.10	0.66
0.35	0.73
0.60	0.86
0.85	1.03

I generated data according to these procedures for three ***rater-sample sizes*** (50, 100, and 250) and three ***ratee-sample sizes*** (1000, 2000, and 3000). In all data sets, raters were randomly assigned to ratees, and a second rating was randomly assigned to either 10% or 100% of the ratees (i.e., the ***double-scoring rate***), two common double-scoring designs observed in operational settings. Hence, the resulting data is not a full rater-by-ratee design (i.e., every rater does not rate every ratee); with random rater-response assignment, the simulated rating data sets

contained missing values. Thus, the number of assigned ratings varied across raters, which created a ***rating-count*** variable which was recorded because this variable could affect the measurement error of the centrality indices and hypothesis tests based on those indices. As the simulated data is graphed in Figure 5.3-1, raters' average rating-count decreases as the rater sample size enlarges and increases as the ratee sample size diminishes. On average, each rater assigns more ratings when all responses receive a second score comparing to when 10% of responses receive a second score. For each combination of these independent variables, I generated 1000 data sets, resulting in 72,000 data sets (i.e., 1000 replications × 4 centrality-strength levels × 2 double-scoring rates × 3 rater-sample sizes × 3 ratee-sample sizes).

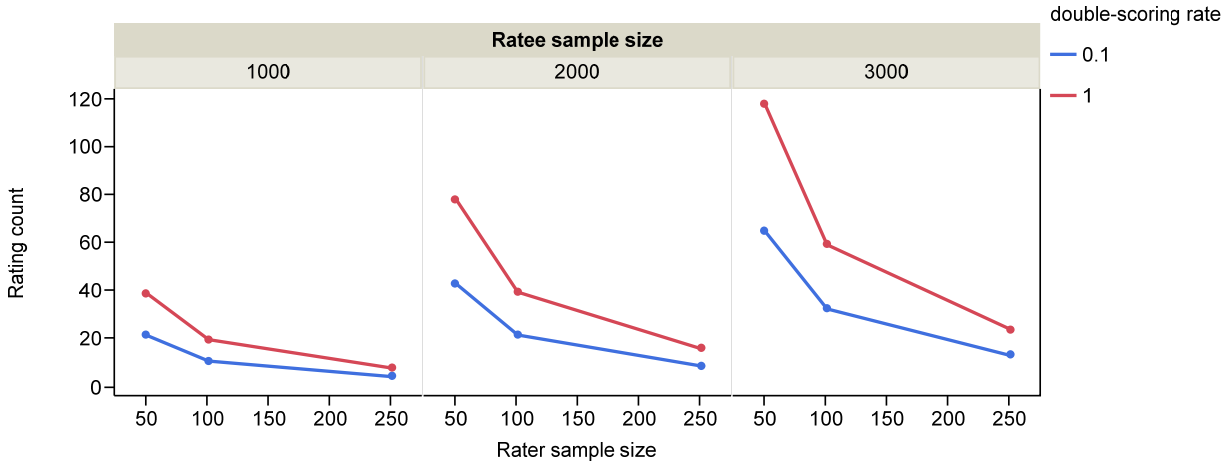

Figure 5.3-1: Raters' average rating count by the double-scoring rate and the number of raters and ratees

5.3.2 Analysis

After the data sets were generated, I performed the following activities. First, parameters were estimated for each simulated data set for the Rasch rating scale model (Andrich, 1978) using the *Winsteps* software (Linacre, 2009b). Second, each of the six rater centrality indicators

(standard deviation of assigned scores, standardized mean-square fit indices, score-measure correlation, measure-residual correlation, and expected-residual correlation) were computed for each rater (effect and non-effect). These index values, along with the experimental parameters, are the data source for the ensuing analyses. Third, I applied hypothesis testing procedures to each rater on each rater centrality index. That is, I specified a critical value for each index and declared each rater to be an "effect" or "non-effect" rater depending on whether the value of the rater centrality index was more extreme (effect) or less extreme (non-effect) than the specified critical values. Fourth, to determine the accuracy and inaccuracy of these hypothesis test results for each rater centrality index, I compared the generating rater type (effect or non-effect) to the rater type classification based on the hypothesis test results. This comparison resulted in one of four outcomes for each rater: (a) correct classification as an effect rater (a true effect detection, depicted by statistical power), (b) correct classification as a non-effect rater (a true non-effect conclusion), (c) incorrect classification as an effect rater (a false positive, depicted by the Type I error rate), or (d) incorrect classification as a non-effect rater (a false negative, depicted as the Type II error rate). Because the results of a previous study suggested that each index is differentially predictive of rater centrality across double-scoring rates and centrality strength, I report the findings of this study stratified on the levels of those variables.

It is important to note that, in the third step of this process, I relied on two types of critical values. The first type of critical value is based on a hypothesis testing framework. That is, for some of the rater centrality indices, I posited a parametric form for the null distribution of the rater centrality index and created critical values to set a proportion of the area of that null distribution as the rejection region for the hypothesis tests. These hypothesis testing critical

values were determined for and applied to the standard deviation of assigned scores and the correlation-based indices (i.e., the point-measure correlation, the measure-residual correlation and the expected-residual correlation), because the value of these indices can be assumed to follow either an F distribution or a normal distribution. In both cases, the null hypothesis assumes no difference between a particular rater and the null distribution, which is based on non-effect raters, so raters for whom the null hypothesis is rejected are flagged as exhibiting the centrality effect. Table 5.3-2 presents a summary of these hypothesis tests, followed by detailed description in the next couple of paragraphs.

Table 5.3-2: *Summary of hypothesis tests on the centrality-detection indices*

Indices	Hypothesis Test	Statistic	Critical Value
Point-measure correlation Measure-residual correlation Expected-residual correlation	$H_0: \rho_{effect} = \rho_{non-effect}$ $H_a: \rho_{effect} < \rho_{non-effect}$	$z = \frac{r' - \rho'}{S_e}$	-1.64 $\alpha = 0.05$
STD	$H_0: \sigma_{effect} = \sigma_{non-effect}$ $H_a: \sigma_{effect} < \sigma_{non-effect}$	$F = \frac{S^2_{effect}}{S^2_{non-effect}}$	$F_{.95, df_{effect}, df_{non-effect}}$
ZMSW ZMSU			± 2.0

Note: *r'*= converted correlation using Fisher's Z transformation (Fisher, 1921). ρ' = the transformed mean value of correlations for the sample of raters in an experimental condition with which a given rater is associated.

In the hypothesis tests, values of correlation-based indices were first transformed using Fisher's Z calculated as

$$r' = \frac{1}{2}\ln(\frac{1+r}{1-r})$$

where *r* is correlation coefficient. We assume *r'* is approximately normally distributed around ρ', the transformed mean value of correlations for the sample of raters in an experimental condition with which a given rater is associated. From this it follows that

$$z = \frac{r' - \rho'}{S_e}$$

is a standard normal curve deviate. The standard error is calculated as

$$S_e = \sqrt{\frac{1}{total / 1000 * (rater\_n - 1)} - 3}$$

where *total* is the number of ratings assigned by all raters in each condition combination of rater type, double-scoring rate, centrality strength, number of raters and ratees, and *rater_n* is the number of raters in which experimental condition a given rater is located. These calculated *z* statistics were compared to the critical values in order to conduct the rater's classification hypothesis test. I expected the index values for the effect raters to be negative or smaller than those for the non-effect raters. That is, I compared the observed *z* against the corresponding critical value at alpha level of 0.05, i.e., the critical value of -1.64 for one-tailed hypothesis test. A similar process was followed for the standard deviation of assigned scores (STD). First, each standard deviation was formulated as $F = \frac{S^2_{effect}}{S^2_{non-effect}}$. Second, critical values were determined based on the degrees of freedom for the numerator and the denominator at each group level. For effect raters, the degree of freedom is calculated as

$$df_{effect} = rating\_count - 1$$

where *rating_count* is the number of ratings assigned by a corresponding rater at each group level. For non-effect raters, the degree of freedom is calculated as

$$df_{non-effect} = \frac{total}{1000 * (rater\_n - 1)} - 1.$$

Third, each rater's F statistic was compared to this critical value, and the rater was declared to be an effect or non-effect rater. In this case, I expect the standard deviation of assigned scores for the effect raters to be smaller than those for the non-effect raters.

The second type of critical value I utilized in the study is not based on a hypothesis testing framework. Rather, as is often the case in applications of some of these indices, I adopted a rule-of-thumb critical value. This type of test applied to the standardized outfit (ZMSU) and infit (ZMSW) statistics, where, in practice, critical values for these indices that are commonly accepted in the literature range are +2.0 and -2.0. That is, a simulated rater was flagged if the observed value of standardized mean-square fell outside of this range.

For each simulated rater, both effect and non-effect raters, I determined the accuracy of the rater classification after applying the relevant critical value and then determined the proportion of incorrect classifications for non-effect raters (a Type I error or false positive) and the proportion of incorrect classifications for effect raters (a Type II error or false negative rate) as well as the inverse of these values, which depict the proportion of correct classifications for each group of raters.

5.4 Results

Table 5.4-1 summarizes the Type I error rates for the selected centrality-detection indices, stratified by selection rate and centrality effect strength. These figures reveal that all of these indices provided reasonable protection against Type I error, ranging from 0.01 to 0.04 when all responses received a second score. Type I error rates tended to be larger when only 10% of responses were double scored, and Type I error rates were considerably larger for the point-measure correlation (i.e., $\alpha = 0.22$), the measure-residual correlation (i.e., $\alpha = 0.09$), and the expected-residual correlation (i.e., $\alpha = 0.09$) under this condition. The raw score standard deviation and the two standardized mean-square fit indices provided the greatest protection

against Type I error under conditions involving this relatively large amount of missing data. It is noteworthy to point out that the Type I error rate did not vary across levels of centrality strength or the variation appeared to be close to a constant (also shown in the Figure 5.4-1). This indicates that the probability of incorrectly indentifying normal raters was not affected by the severity of the centrality effect.

Table 5.4-1: *Type I error rates for centrality-detection indices*

Double-scoring Rate	Centrality Strength	Point-measure Correlation	Measure-residual Correlation	Expected-residual Correlation	STD	ZMSW	ZMSU
0.1	0.10	0.22	0.09	0.09	0.01	0.06	0.05
	0.35	0.22	0.09	0.09	0.01	0.06	0.05
	0.60	0.22	0.09	0.09	0.01	0.06	0.05
	0.85	0.22	0.09	0.09	0.01	0.06	0.05
1.0	0.10	0.04	0.03	0.04	0.01	0.04	0.04
	0.35	0.04	0.03	0.04	0.01	0.04	0.04
	0.60	0.04	0.03	0.04	0.01	0.04	0.04
	0.85	0.04	0.03	0.04	0.01	0.04	0.04

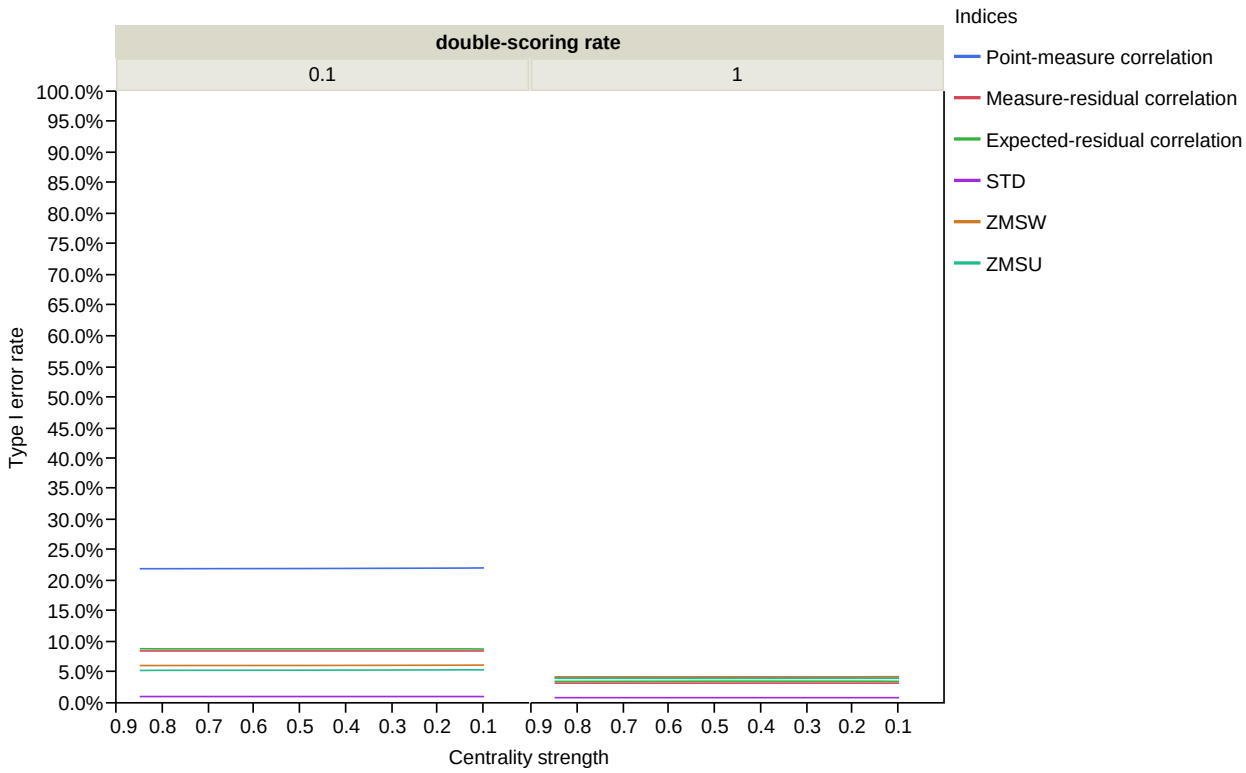

Figure 5.4-1. Type I error rate varying as a function of the double-scoring rate and strength of the centrality effect

Table 5.4-2 summarizes the statistical power rates for the studied centrality-detection indices. None of these indices provided good statistical power rate (i.e., 1 – Type II error rate > 0.80) unless the rater centrality effect shrank the variance of ratee latent distribution to at least half its true value. That is, the statistical power rate approached values of 0.80 and higher only when the centrality strength was set to 0.10, which is roughly equivalent to a variance ratio of ratee latent distribution of effect raters to normal raters at 0.66 (See Table 5.3-1). Higher statistical power rate was achieved when responses were 100% double scored in comparison to only 10% being double scored. The statistical power rates of the two mean-square fit statistics was especially low, indicating that they could only correctly identify aberrant raters for up to 23% of time. For

the other four indices, with 10% of responses being double scored, the statistical power was no higher than 50%. Also presented in Figure 5.4-2, the point-measure correlation seems to have the best statistical power of all, and the two mean-square fit statistics appeared to have the worst statistical power. The measure-residual and expected-residual correlations had fairly close values in the statistical power rates.

Table 5.4-2: *Statistical power rates for centrality-detection indices*

Double-scoring Rate	Centrality Strength	Point-measure Correlation	Measure-residual Correlation	Expected-residual Correlation	STD	ZMSW	ZMSU
0.10	0.10	0.43	0.37	0.38	0.31	0.19	0.05
	0.35	0.37	0.33	0.33	0.22	0.18	0.05
	0.60	0.29	0.23	0.23	0.09	0.18	0.06
	0.85	0.24	0.13	0.14	0.02	0.17	0.06
1.00	0.10	0.79	0.51	0.51	0.50	0.23	0.22
	0.35	0.62	0.42	0.43	0.38	0.08	0.08
	0.60	0.33	0.25	0.26	0.17	0.05	0.05
	0.85	0.10	0.08	0.09	0.03	0.05	0.04

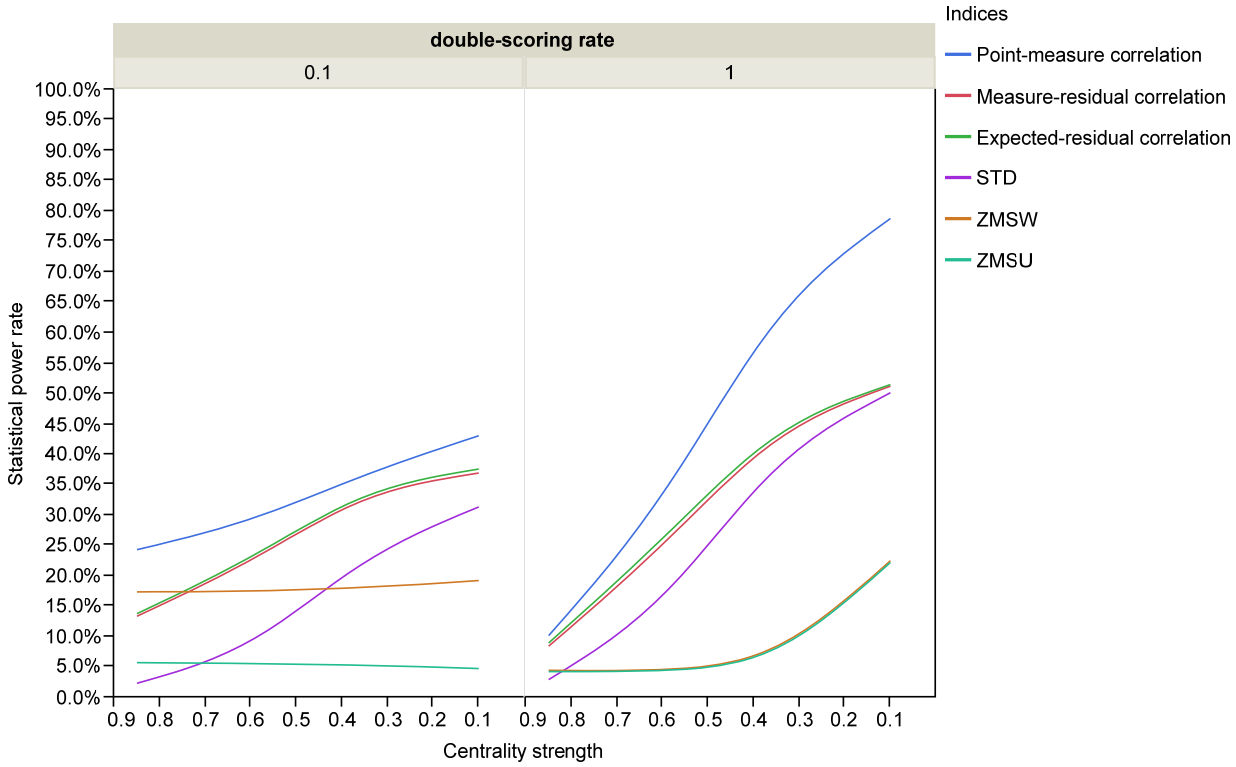

Figure 5.4-2. Statistical power rates varying as a function of the double-scoring rate and strength of the centrality effect

5.5 Conclusions

The results indicate that mean-square fit indices, indices that have been used to detect a variety of rater effects including rater centrality, provide reasonable protection against Type I errors but exhibit the lowest level of statistical power of any of the indices I examined. That is, they would be a very poor choice if the goal of the analyst was to detect rater centrality when it exists. Specifically, although the mean-squared fit indices maintained a Type I error rate close to 0.05, the statistical power rate of those indices never rose above 0.25 and remained less than 0.10 in most of the conditions that I examined.

Of the remaining indices, all of them provided reasonable protection against Type I errors when the double-scoring rate was 100%. However, when the double-scoring rate decreased to 10% of the simulated examinees, Type I error rates increased from about 0.04 to about 0.09 for the expected-residual correlation and the measure-residual correlation. The Type I error rate increased substantially for the point-measure correlation (from 0.04 to 0.22) with this decrease in double-scoring rate. The raw score standard deviation was the only index that did not exhibit an increase in Type I error rate when double-scoring rate was decreased from 100% to 10%, maintaining a rate of about 0.01.

Statistical power analysis of these indices, on the other hand, suggests that the raw score standard deviation may not be the best choice for flagging raters for the centrality effect. In all cases, the statistical power rates decreased when double-scoring rate decreased from 100% to 10%. Also in all cases, however, the raw score standard deviation exhibited lower statistical power than any of the remaining indices with the exception of the mean-square fit indices. Additionally, the raw score standard deviation exhibited only marginally acceptable levels of statistical power (i.e., greater than 0.20) when the rater centrality strength was relatively strong (i.e., the ability variance was shrunk to about 35% of its true value). Under the condition of 100% double-scoring, the point-measure correlation exhibited superior statistical power, achieving levels greater than 0.33 for relatively low levels of rater centrality strength and achieving levels greater than 0.62 for higher levels of rater centrality strength. On the other hand, the point-measure correlation performed only slightly better than the measure-residual correlation and the residual-expected correlation when 10% of the responses were double-scored.

This fact, coupled with the high Type I error rates suggests that the point-measure correlation may not be a good choice when the double-scoring rate is low.

I recommend that analysts use either the measure-residual correlation or the expected-residual correlation for detecting the centrality effect when attempting to balance both Type I error and statistical power. I suggest using the point-measure correlation only when responses are 100% double scored. Furthermore, in some practical applications of these centrality-detection indices, Type I errors are more important than statistical power. Specifically, in a situation where incorrectly identifying normal raters as effect raters is more of concern than successfully identifying effect raters (e.g., when training raters is expensive or when the pool of potential raters is very small), one may want to avoid using the point-measure correlation, since in such a case care is usually focused on minimizing the occurrence of statistical errors.

A limitation of this study is the fact that the critical values picked for the hypothesis testing were simulation-based rather than from the "reality". Due to the fact that the simulated data are extremely "clean," it is difficult to determine how well the simulation results will generalize to the "messy" data that are typical of operational contexts. That is, the data were modeled to fit the Rasch model, and other potential sources of model-data misfit were omitted from the data generation process (e.g., multidimensionality, rater drift over time, etc.). Hence, application of these rater centrality-detection indices to real data sets and to more complex simulated data may provide a support of the generalizability of the results.

Further study is warranted concerning each of these rater effects. Specifically, in addition to evaluating the degree to which these dichotomous categories that raters are placed into are

accurate, additional analyses can focus on the influential factors from the rating data collection design on the classification accuracy.

5.6 References

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

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

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

Fisher, R. A. (1921). On the probable rror of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3-32.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Linacre, J. M. (2009b). WINSTEPS Rasch measurement computer program (Version 3.68.0). Chicago, IL: Winsteps.com.

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.

Wolfe, E. W. (1998). *A two-parameter logistic reader model (2PLRM): Detecting reader harshness and centrality.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

Wolfe, E. W. (2005). Identifying Rater Effects in Performance Ratings. In S. Reddy (Ed.), *Performance Appraisals: A Critical View* (pp. 91-103). Hyderabad, India: ICFAI University Press.

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects with a multi-faceted Rasch rating scale model. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147-164). Stamford, CT: Ablex.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: generalised item response modeling software (Version 2.0). Camerwell, Victoria: ACER Press.

6 MANUSCRIPT 3: THE IMPACT OF DATA COLLECTION DESIGN ON RATER CLASSIFICATION BY CENTRALITY INDICES

6.1 Abstract

This study evaluates how data collection design influences the accuracy of rater classifications indicated by centrality-detection indices. In this simulation study, I generated datasets by varying the strength of centrality effect, the number of raters, the number of ratees, and the proportion of ratees assigned a second score. The results indicate that these experimental parameters had different impact on the accuracy of rater classification and that improving the classification accuracy for non-effect raters may come at a cost of reducing the classification accuracy for effect raters. Some simple guidelines for the expected impact of classification accuracy when a higher-order interaction exists summarized from the analyses offer a glimpse of the "pros" and "cons" in adjusting the magnitude of the parameters when we evaluate the impact of the four experimental parameters on the outcomes of rater classification.

6.2 Introduction

Performance assessments are used extensively in educational testing, particularly in content areas that are difficult to measure with multiple-choice items, and ratings assigned by a human are commonly used to judge the quality of the assessment products (Landy & Farr, 1980). Unfortunately, performance ratings are prone to various types of systematic and random error, potentially rendering the associated scores inaccurate as indicators of the student's true performance. That is, raters may exhibit *rater effects* that serve as sources of error in the

performance ratings. Therefore, identifying aberrant raters successfully is an important step in evaluating the psychometric quality of ratings in terms of validity and reliability. Previous research has identified several ways that rater effects are manifested in ratings, and researchers have developed a variety of criteria under different methodological frameworks (e.g., classical test theory, analysis of variance, regression-based analysis, generalizability theory, and Rasch measurement and item response theory) to evaluate the psychometric quality of performance ratings (Saal, et al., 1980). However, supporting simulation studies and operational applications of these methods have remained sparse.

Previous research concerning rater effects has identified several ways that raters introduce error into ratings, including severity/leniency, centrality/extremism, and inaccuracy (Saal, et al., 1980). This study focuses on detecting the centrality effect. A rater who exhibits the centrality effect overuses the middle categories of the rating scale while avoiding extreme categories. Hence, the centrality effect induces a reduction of the variation of assigned ratings. In addition, centrality results in accurate ratings in the central range of the ability continuum, but overestimates of ratee proficiency for non-proficient ratees and underestimates of ratee proficiency for highly proficient ratees. Clearly, the centrality effect will manifest itself in the standard deviation of the observed ratings. However, the standard deviation may be a poor choice as a rater effect index because it is inflated when random error exists in the ratings. As a result, it would be difficult to determine whether raters who produce a small standard deviation are engaging in centrality or are simply accurate raters. In addition, many operational analyses of rater effects employ latent trait modeling procedures (e.g., item response theory), and numerous

indices exist within those frameworks that might be better suited as indicators of the centrality effect.

One family of such indices includes the various mean-square fit indices that are commonly reported by commercial latent trait software, such as Winsteps (Engelhard, 1992, 1994). When a rater exhibits the centrality effect in ratings, we may expect the fit statistics to deviate away from 1.0 and the absolute value of standardized fit statistics to be larger than 2.0. We might expect the unweighted mean-square fit statistic to be more sensitive to centrality because it is more heavily influenced by large residuals for mismatched raters and ratees. We could expect these cases to produce larger residuals (i.e., a normal rater would produce ratings of high and low ability ratees that have large residuals). However, prior research has indicated that rater centrality may manifest itself inconsistently in these indices, and may, therefore, cause analysts to incorrectly conclude that accurate raters are engaging in central rating patterns (Wolfe, et al., 2000). Hence, in this article, I evaluated four additional indices to detect rater centrality: the score-measure correlation (also known as the point-measure correlation), the measure-residual correlation, the expected-residual correlation (Wolfe, 2004b, 2005), and a rater slope index (Wolfe, 1998a).

The score-measure correlation is also called the point-measure correlation, denoted as "PTMEA" in *Winsteps* output. In rating data, it correlates Rasch measures of ratees with ratings assigned by raters, and is computed by the formula

$$r_{score,measure} = \frac{1}{n-1}\sum_{n=1}^{N}\left(\frac{X_{nr}-\overline{X}_{r}}{S_{X_r}}\right)\left(\frac{\theta_n-\overline{\theta}}{S_{\theta}}\right),$$

where $\frac{X_{nr} - \overline{X}_r}{S_{X_r}}$ is the standard score of rating assigned to ratee *n* by rater *r*, $\overline{X}_r$ is the sample mean of ratings assigned by rater r, S_{X_r} is the sample standard deviation of ratings assigned by rater *r*, $\frac{\theta_n - \overline{\theta}}{S_\theta}$ is the standard score of Rasch measure of ratee *n*, and S_θ is the standard deviation of Rasch measure of all ratees. In typical item analyses using Rasch measurement, we use item correlations as an immediate check that the response level scoring is consistent with the scoring of other items (i.e., the rank ordering of ratees is the same for the item as for the estimated ratee measures). In the case of studies of rater centrality, we can evaluate the score-measure correlation to see if assigned ratings to ratees by a particular rater are consistent with the estimated ratee measures, which would be based on the ratings assigned by all raters. The value of a score-measure correlation from raters with centrality should be less than one obtained from raters without centrality. That is, a highly positive value of the score-measure correlation indicates a strong association between ratees' ability measure and ratings assigned by a rater, and raters with the centrality effect are expected to slightly weaker correlations due to the attenuation of the assigned score distribution. Additionally, the score-measure correlation (and the analogous point-biserial correlation) is likely to be influenced by other features of the data (e.g., the targeting of the rater on the person sample and the distribution of the personal sample) as well as other rater effects (e.g., rater inaccuracy). As a result, the score-measure correlation may have limited power for detecting the centrality effect.

The expected-residual correlation depicts the association of estimated ratings and residuals, $r_{\exp,res}$. Specifically, the expected-residual correlation is based on the notion that the

raw residual, $x_{nr} - E_{nr}$, produced by all ratings assigned by rater *r* who exhibits the centrality effect is positive for ratees of low ability (i.e., x_{nr} is greater than E_{nr}) and negative for ratees of high ability (i.e., x_{nr} is less than E_{nr}). As a result, when the centrality effect exists, a scatter plot of the expected scores (X axis) and residuals (Y axis) should exhibit a negative association. A value of the expected-residual correlation that approaches its upper limit of 1.00 indicates that the way a rater utilizes the rating scale is uniform across the ability continuum. Similarly, the measure-residual correlation correlates Rasch measures of ratees and residuals, $r_{measure,res}$, and this index behaves in a similar manner when the centrality effect exists

The rater slope index, also called rater discrimination, may be sensitive to rater centrality, In the Rasch measurement software *Winsteps*, the discrimination index is not an estimated parameter contained in the model, although the software will output a crude estimate of what that parameter would be, if it were to be estimated for the sake of allowing an analyst to evaluate the reasonableness of the common slopes assumption inherent in the Rasch model. In applications of the Rasch model, the slope is set to a constant value of 1.0. On the other hand, the empirical slope is estimated in applications of a 2-PL IRT model so that the slope differs for each item (or, in this case, rater). The reported values of rater slope that are output by *Winsteps* are a first approximation to the precise value of a_r obtained from the Newton-Raphson estimation equation (Wright & Masters, 1982, pp. 72-77):

$$\hat{a}_r = 1 + \left[\frac{\sum_{n=1}^{N} (X_{nr} - P_{nr})(\theta_n - b_r)}{\sum_{n=1}^{N} P_{nr}(1 - P_{nr})(\theta_n - b_r)^2} \right],$$

where X_{nr} is the rating assigned to ratee *n* by rater *r*, P_{nr} is the probability for ratee *n* rated by rater *r* to get the item correct, θ_n is the Rasch measure for ratee *n*, and b_r is the severity of ratee *r*. Because a rater exhibiting the centrality effect may assign accurate ratings in the central range of the ability continuum and will overestimate ratee proficiency for non-proficient ratees and underestimate ratee proficiency for highly proficient ratees, the centrality effect rater could have, on average, lower discrimination estimates than non-effect raters. Therefore, small values of the slope index could be a symptom of the centrality effect. The commonly acceptable lower acceptable limit for the slope index is 0.5, and raters with lower values would be flagged as candidates for the centrality effect.

This simulation study is an extension of previous investigations (Chapter 4 and 5) of the relative strengths and the accuracy of these centrality-detection indices. In the first study, I compared the relative strengths of several centrality-detection indices in differentiating effect rater from non-effect raters. The difference in the values of each rater centrality index between rater types were treated as the dependent variable in an analysis of variance, and the remaining variables were treated as the independent variables. The results revealed that the measure-residual correlation, the expected-residual correlation, and the standardized deviation of assigned scores exhibited stronger discrimination between effect and non-effect raters than did the point-measure correlation. The mean-square fit statistics, traditionally viewed as potential indicators of rater centrality, performed poorly in terms of differentiating central raters from normal raters. Along with the rater slope index, the mean-square fit statistics did not appear to be sensitive to the rater centrality effect. In the second study, I explored the accuracy of theses indices, specifically their Type I and Type II errors and statistical power. Hypothesis testing was applied

to each rater's centrality indices and the results of this hypothesis test were compared to the rater's true centrality status, differentiating the results as a function of centrality strength and the double-scoring rate. The results provide reports of accuracy and inaccuracy of these indices through Type I and II error and statistical power rates. The results show that all of these indices provided reasonable protection against Type I errors when all responses were double scored, and that higher statistical power rate was achieved when responses were 100% double scored in comparison to only 10% being double scored. With a consideration on balancing both Type I error and statistical power, I recommended the measure-residual correlation and the expected-residual correlation for detecting the centrality effect. I did not suggest using the point-measure correlation when double-scoring rate is low.

The purpose this paper is to further explore the impact of rating data collection design on each index's rater-classification accuracy. That is, this manuscript explores the question "how do the features of the data collection design including double-scoring rate, centrality strength, and the number of raters and ratees, influence the accuracy of rater classification by the centrality-detection indices to be considered." Specifically, how the probability of a rater being correctly classified as effect rater (i.e., true positive rate) and the probability of a rater being correctly classified as non-effect rater (i.e., true negative rate) are influenced by those experimental parameters? The indices that were sensitive to the centrality strength or that at least exhibited an interaction between the centrality strength and the double-scoring rate in the prior study are included this paper. That is, the point-measure correlation, the measure-residual correlation, the expected-residual correlation and the standard deviation of assigned scores are the central focus

of this manuscript. For completeness, the standardized mean-square fit statistics are also included in the comparisons.

6.3 Methods

6.3.1 Simulation Design

I generated simulated ratings on a six-point scale using *ConQuest* 2.0 (Wu, et al., 2007) based on a unidimensional Rasch model that contains a single measurement facet (i.e., ratings assigned by raters to a single assessment item). I specified rating scale step threshold parameter values to create an approximately normal distribution of ratings for each simulated rater. In order to avoid confounding rater centrality and rater severity/leniency effects, all rater severity parameters were set to zero for all raters during data generation. That is, data sets were generated according to a Rasch rating scale model in which the item "difficulty" parameter was replaced by a rater "severity" parameter which was set to a constant value of zero.

Starting with these "non-effect" data, a single simulated rater exhibiting a centrality effect was generated by shrinking the underlying ratee ability to one of four ***centrality-strength*** levels (0.10, 0.35, 0.60, and 0.85; i.e., the variance of the ability distribution is shrunk to a minimum of 10% and to a maximum of 85% of that of the non-effect rater) for a single rater, and these effect ratings were added into the simulated data file in order to create a ***rater-type*** variable (namely focused non-effect or normal rater and manipulated effect rater). The relationship between the centrality strength and the true ability distribution can be expressed as a variance ratio defined as,

$$\text{Variance ratio} = \frac{\sigma^2_{effect}}{\sigma^2_{true}},$$

where σ^2_{effect} is the variance of the ability distribution for the ratees whose responses are rated by effect raters and σ^2_{true} the variance of ability distribution for the ratees whose responses are rated by non-effect raters. Table 6.3-1 translates the centrality strength values into the equivalent empirical variance ratios. For example, a centrality-strength value of 0.1 causes the variance of the ability distribution for an effect rater to be 66% of the variance of the ability distribution produced by normal raters. From a centrality-strength value of 0.85, we can consider the centrality with a rater to be trivial as the variance ratio is close to 1.

Table 6.3-1: *Variance ratio vs. centrality strength*

Centrality Strength	Variance Ratio
0.10	0.66
0.35	0.73
0.60	0.86
0.85	1.03

I generated data according to these procedures for three ***rater-sample sizes*** (50, 100, and 250) and three ***ratee-sample sizes*** (1000, 2000, and 3000). In all data sets, raters were randomly assigned to ratees, and a second rating was randomly assigned to either 10% or 100% of the ratees (i.e., the ***double-scoring rate***), two common double-scoring designs observed in operational settings. Hence, the resulting data is not a full rater-by-ratee design (i.e., every rater does not rate every ratee); with random rater-response assignment, the simulated rating data sets contained missing values. Thus, the number of assigned ratings varied across raters, which created a ***rating-count*** variable which was recorded because this variable could affect the measurement error of the centrality indices and hypothesis tests based on those indices. As the

simulated data is graphed in Figure 6.3-1, raters' average rating-count decreases as the rater sample size enlarges and increases as the ratee sample size diminishes. On average, each rater assigns more ratings when all responses receive a second score comparing to when 10% of responses receive a second score. For each combination of these independent variables, I generated 1000 data sets, resulting in 72,000 data sets (i.e., 1000 replications × 4 centrality-strength levels × 2 double-scoring rates × 3 rater-sample sizes × 3 ratee-sample sizes).

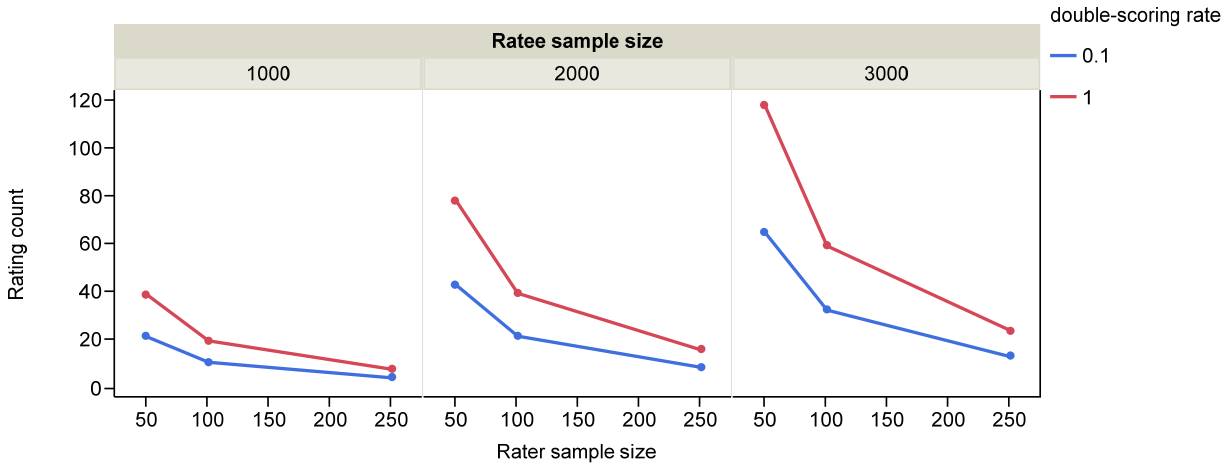

Figure 6.3-1: Raters' average rating count by the double-scoring rate and the number of raters and ratees

6.3.2 Analysis

After the data sets were generated, I performed the following activities. First, parameters were estimated for each simulated data set for the Rasch rating scale model (Andrich, 1978) using the *Winsteps* software (Linacre, 2009b). Second, each of the six rater centrality indicators (standard deviation of assigned scores, mean-square fit indices, score-measure correlation, measure-residual correlation, and expected-residual correlation) were computed for each rater (effect and non-effect). These index values, along with the experimental parameters, are the data

source for the ensuing analyses. Fourth, I computed two dichotomous variables for each centrality-detection index. One of those variables records the outcome of rater classification (1=correct, 0=incorrect) for non-effect raters. The other variable records the outcome of rater classification (1=correct, 0=incorrect) for effect raters. Lastly, I used logistic regressions to determine how the features of the data collection design influences these two outcome variables for each centrality-detection index. That is, I conducted a logistic regression separately for effect raters and for non-effect raters by specifying dichotomous rater classification accuracy (correct classification versus incorrect classification) as a function of several predictor variables (i.e., the double-scoring rate, the centrality strength, the number of raters, and the number of ratees) separately for each rater centrality index. For each index, the analyses commenced with a saturated model under a full factorial design to capture all possible effects and then used a stepwise procedure for model reduction.

To facilitate interpretation of the results, I first centered the independent variables (i.e., the experimental parameters except for the double-scoring rate) on their grand means, and then created cross-product interaction terms using the mean-centered experimental parameters. After parameters were estimated, I multiplied each estimate by its standard deviation (summarized in Table 6.3-2) and took the exponential of the products to get odds ratios. This is equivalent to standardizing each predictor terms prior to the logistic regressions. I used odds ratios as indicators of effect sizes, and applied Monahan's (2007) suggestion on cut-off values for large effect sizes being odds ratios greater than 1.89 (or less than 0.53, for negative relationships) and moderate effect sizes being an odds ratio greater than 1.53 (or less than 0.65, for negative

relationships). The interpretations focus on either the main effects or the highest-order interaction effects with minimum of medium effect sizes.

Table 6.3-2: *Standard deviations of mean-centered experimental parameters and their interactions*

Mean-centered Variable	Std Dev
CS	0.28
RATERN	82.12
RATEEN	816.50
DSR*CS	0.20
DSR*RATERN	58.36
DSR*RATEEN	580.23
DSR*CS*RATERN	16.31
DSR*CS*RATEEN	162.18
DSR*CS*RATERN*RATEEN	13318.63
DSR*RATERN*RATEEN	47650.18
CS*RATERN*RATEEN	18741.91
CS*RATERN	22.95
CS*RATEEN	228.22
RATERN*RATEEN	67053.10

Note: CS = centrality strength, DSR = double-scoring rate, RATERN = number of raters, RATEEN = number of ratees. For each centered variable, N = 9,672,000 and Mean = 0.00.

6.4 Results

Table 6.4-1 presents the results from the logistic regression analyses on the true negative outcomes. For the point-measure correlation (Table 6.4-1.a), the main effect "RATEEN" (number of examinees) and the interaction effect "DSR*RATERN" (double-scoring rate-by-number of raters) had medium effect sizes (i.e., odds ratios greater than 1.53 for positive relationship or less than 0.65 for negative relationships). For the measure-residual correlation (Table 6.4-1.b), the interaction effect "DSR*RATEEN" (double-scoring rate-by-number of examinees) had a large effect size (i.e., odds ratios greater than 1.89 or less than 0.53 for

negative relationships). For the expected-residual correlation (Table 6.4-1.c), the interaction effects “DSR*RATEEN” (double-scoring rate-by-number of examinees) and “RATERN*RATEEN” (number of raters-by-number of examinees) had medium effects sizes. For the infit mean-square statistic (Table 6.4-1.e), the interaction effect “DSR*RATERN*RATEEN” (double-scoring rate-by-number of raters-by-number of examinees) had a medium effect size. For the outfit mean-square statistic (Table 6.4-1.f), the interaction effects “DSR*RATERN” and “DSR*RATEEN” had large effect sizes.

Table 6.4-1: R*esults from logistic regression analyses on the true negative outcomes*

Effect	Estimate	Pr > ChiSq	Odds Ratio	Direction of Relationship
a. Point-measure Correlation				
Main Effect				
Intercept	2.613	<.0001		
DSR 0.1 vs. 1	0.491	<.0001	1.634	
CS	0.005	0.6083	1.001	
RATERN	0.008	<.0001	1.926	
RATEEN	-0.001	<.0001	**0.633**	-
Interaction Effect				
DSR*CS	0.027	0.0407	1.005	
DSR*RATERN	-0.008	<.0001	**0.621**	-
DSR*RATEEN	0.001	<.0001	1.450	
DSR*CS*RATERN	0.000	0.0106	0.994	
DSR*RATERN*RATEEN	0.000	<.0001	1.249	
CS*RATERN	0.000	0.4925	0.998	
RATERN*RATEEN	0.000	<.0001	0.788	
b. Measure-residual Correlation				
Main Effect				
Intercept	2.564	<.0001		
DSR 0.1 vs. 1	0.708	<.0001	2.030	
RATERN	0.004	<.0001	1.351	
RATEEN	-0.001	<.0001	0.643	
Interaction Effect				

DSR*RATERN	-0.006	<.0001	0.713	
DSR*RATEEN	0.001	<.0001	**1.542**	+
DSR*RATERN*RATEEN	0.000	<.0001	1.409	
RATERN*RATEEN	0.000	<.0001	0.655	

c. Expected-residual Correlation

Main Effect

Intercept	2.569	<.0001		
DSR 0.1 vs. 1	0.626	<.0001	1.870	
RATERN	0.004	<.0001	1.410	
RATEEN	-0.001	<.0001	0.603	

Interaction Effect

DSR*RATERN	-0.006	<.0001	0.705	
DSR*RATEEN	0.001	<.0001	**1.590**	+
DSR*RATERN*RATEEN	0.000	<.0001	1.445	
RATERN*RATEEN	0.000	<.0001	**0.630**	-

d. STD

Main Effect

Intercept	4.365	<.0001		
DSR 0.1 vs. 1	0.147	<.0001	1.158	
RATERN	-0.002	<.0001	0.852	
RATEEN	0.000	0.0002	1.018	

Interaction Effect

DSR*RATERN	-0.002	<.0001	0.906	
DSR*RATEEN	0.000	<.0001	1.170	
DSR*RATERN*RATEEN	0.000	<.0001	1.149	
RATERN*RATEEN	0.000	<.0001	0.830	

e. ZMSW (standardized weighted mean-square statistic or infit)

Main Effect

Intercept	4.527	<.0001		
DSR 0.1 vs. 1	-1.493	<.0001	0.225	
RATERN	0.020	<.0001	5.126	
RATEEN	-0.002	<.0001	0.195	

Interaction Effect

DSR*RATERN	-0.021	<.0001	0.294	
DSR*RATEEN	0.002	<.0001	3.362	
DSR*RATERN*RATEEN	0.000	<.0001	**1.689**	+

RATERN*RATEEN		0.000	<.0001	0.511	

f. ZMSU (standardized unweighted mean-square statistic or outfit)

Main Effect					
Intercept		4.396	<.0001		
DSR	0.1 vs. 1	-1.308	<.0001	0.270	
CS		0.006	0.3002	1.002	
RATERN		0.017	<.0001	4.106	
RATEEN		-0.001	<.0001	0.349	
Interaction Effect					
DSR*RATERN		-0.018	<.0001	**0.342**	-
DSR*RATEEN		0.001	<.0001	**2.253**	+
DSR*RATERN*RATEEN		0.000	<.0001	1.262	
CS*RATEEN		0.000	0.1754	1.002	
RATERN*RATEEN		0.000	<.0001	0.760	

Note: CS = centrality strength, DSR = double-scoring rate, RATERN = number of raters, RATEEN = number of rates, + = positive relationship, - = negative relationship

Table 6.4-2 presents the results from the logistic regression analyses on the true positive outcomes. For the point-measure correlation (Table 6.4-2.a), the main effects “RATERN” (number of raters) and “RATEEN” and the interaction effect “DSR*CS” (double-scoring rate-by-centrality strength) had large effect sizes. For the measure-residual correlation (Table 6.4-2.b), the interaction effect “DSR*CS*RATERN” (double-scoring rate-by-number of raters) had a large effect size. For the expected-residual correlation (Table 6.4-2.c), the main effects “RATERN” and “RATEEN” had medium effects sizes. For the standard deviation of assigned scores (Table 6.4-2.d), all the four main effects had either large or medium effect sizes. For the infit mean-square statistic (Table 6.4-2.e), the interaction effect “DSR*RATERN*RATEEN” had a medium effect size. For the outfit mean-square statistic (Table 6.4-2.f), the interaction effect “DSR*RATERN*RATEEN” had a medium effect size.

Table 6.4-2: *Results from logistic regression analyses on the true positive outcomes*

Effect	Estimate	Pr > ChiSq	Odds Ratio	Direction of Relationship
a. Point-measure Correlation				
Main Effect				
Intercept	-2.416	<.0001		
DSR 0.1 vs. 1	1.823	<.0001	6.188	
CS	0.125	0.5515	1.036	
RATERN	-0.014	<.0001	**0.325**	-
RATEEN	0.001	<.0001	**2.357**	+
Interaction Effect				
DSR*CS	-4.932	<.0001	**0.375**	-
DSR*RATERN	0.003	<.0001	1.204	
DSR*RATEEN	0.000	<.0001	0.807	
DSR*CS*RATERN	0.008	<.0001	1.144	
DSR*CS*RATEEN	-0.001	0.0043	0.890	
DSR*CS*RATERN* RATEEN	0.000	0.0036	1.088	
DSR*RATERN*RATEEN	0.000	<.0001	0.697	
CS*RATERN*RATEEN	0.000	0.0435	0.929	
CS*RATERN	0.012	<.0001	1.317	
CS*RATEEN	-0.001	0.0001	0.814	
RATERN*RATEEN	0.000	<.0001	1.397	
b. Measure-residual Correlation				
Main Effect				
Intercept	-1.384	<.0001		
DSR 0.1 vs. 1	-0.019	0.5649	0.981	
CS	-1.367	<.0001	1.000	
RATERN	-0.004	<.0001	0.999	
RATEEN	0.001	<.0001	1.045	
Interaction Effect				
DSR*CS	-1.221	<.0001	0.000	
DSR*RATERN	-0.004	<.0001	0.999	
DSR*RATEEN	0.000	0.0775	0.995	
DSR*CS*RATERN	0.008	<.0001	**87.162**	+
DSR*CS*RATEEN	-0.001	<.0001	0.989	
DSR*RATERN*RATEEN	0.000	<.0001	0.778	
CS*RATERN*RATEEN	0.000	<.0001	1.048	
CS*RATERN	0.004	<.0001	1.085	

CS*RATEEN		0.000	0.0447	0.958	
RATERN*RATEEN		0.000	<.0001	1.248	
c. Expected-residual Correlation					
Main Effect					
Intercept		-1.415	<.0001		
DSR	0.1 vs. 1	0.049	0.15	1.050	
CS		-1.341	<.0001	0.687	
RATERN		-0.005	<.0001	**0.652**	-
RATEEN		0.001	<.0001	**1.695**	+
Interaction Effect					
DSR*CS		0.001	<.0001	0.790	
DSR*RATERN		-1.189	<.0001	0.822	
DSR*RATEEN		-0.003	<.0001	0.901	
DSR*CS*RATERN		0.000	<.0001	1.124	
DSR*CS*RATEEN		0.007	<.0001	0.901	
DSR*RATERN*RATEEN		-0.001	<.0001	0.750	
CS*RATERN*RATEEN		0.000	<.0001	1.049	
CS*RATERN		0.000	<.0001	1.096	
CS*RATEEN		0.004	<.0001	0.953	
RATERN*RATEEN		0.000	0.0279	1.321	
d. STD					
Main Effect					
Intercept		-2.608	<.0001		
DSR	0.1 vs. 1	0.688	<.0001	**1.990**	+
CS		-2.546	<.0001	**0.491**	-
RATERN		-0.009	<.0001	**0.491**	-
RATEEN		0.001	<.0001	**1.568**	+
Interaction Effect					
DSR*CS		-1.017	<.0001	0.817	
DSR*RATERN		-0.001	0.0042	0.933	
DSR*RATEEN		0.000	0.8609	0.995	
DSR*CS*RATERN		0.004	0.0111	1.062	
DSR*CS*RATEEN		-0.001	<.0001	0.904	
DSR*RATERN*RATEEN		0.000	<.0001	0.902	
CS*RATERN*RATEEN		0.000	<.0001	1.074	
CS*RATERN		0.010	<.0001	1.264	
CS*RATEEN		0.000	<.0001	0.898	

RATERN*RATEEN		0.000	0.5048	0.983	

e. ZMSW (standardized weighted mean-square statistic or infit)

Main Effect					
Intercept		-4.419	<.0001		
DSR	0.1 vs. 1	1.860	<.0001	6.421	
CS		-0.130	0.5075	0.964	
RATERN		-0.019	<.0001	0.203	
RATEEN		0.002	<.0001	5.465	
Interaction Effect					
DSR*CS		-1.967	<.0001	0.677	
DSR*RATERN		0.018	<.0001	2.909	
DSR*RATEEN		-0.002	<.0001	0.301	
DSR*CS*RATERN		0.013	<.0001	1.242	
DSR*CS*RATEEN		-0.001	<.0001	0.877	
DSR*RATERN*RATEEN		0.000	<.0001	**0.621**	-
CS*RATERN		-0.001	0.5148	0.977	
CS*RATEEN		0.000	0.7832	0.993	
RATERN*RATEEN		0.000	<.0001	2.236	

f. ZMSU (standardized unweighted mean-square statistic or outfit)

Main Effect					
Intercept		-4.688	<.0001		
DSR	0.1 vs. 1	2.108	<.0001	8.231	
CS		-0.026	0.9208	0.993	
RATERN		-0.019	<.0001	0.203	
RATEEN		0.002	<.0001	3.723	
Interaction Effect					
DSR*CS		-2.160	<.0001	0.651	
DSR*RATERN		0.019	<.0001	2.961	
DSR*RATEEN		-0.002	<.0001	0.388	
DSR*CS*RATERN		0.017	<.0001	1.317	
DSR*CS*RATEEN		-0.001	<.0001	0.794	
DSR*RATERN*RATEEN		0.000	<.0001	**0.638**	-
CS*RATERN		-0.005	0.0191	0.893	
CS*RATEEN		0.001	<.0001	1.127	
RATERN*RATEEN		0.000	<.0001	1.591	

Note: CS = centrality strength, DSR = double-scoring rate, RATERN = number of raters, RATEEN = number of rates, + = positive relationship, - = negative relationship

The odds ratios for the model effects with medium or large effect sizes suggest that these corresponding effects have a practically important impact on the classification accuracy of these centrality-detection indices. Especially, those interaction effects indicate varying influence of one variable on different levels of other interacting variables. Figures 6.4-1 to 6.4-16, a series of marginal means plot of the effects on the probabilities for the two types of outcomes, illustrate the average influence from these model effects.

For the true negative outcomes, only the rater sample size as a main effect had a practical important impact on the classification accuracy for non-effect raters, while other three experimental parameters influenced the classification accuracy as interactions with one or more of other parameters. In Figure 6.4-1, for the point-measure correlation, as the ratee sample size increased from 1000 to 3000, the classification accuracy for non-effect raters dropped about 2.5%. Specifically, for a non-effect rater, one standard unit increase in the ratee sample size results in a 36.7% decrease in the odds of a non-effect rater being correctly classified.

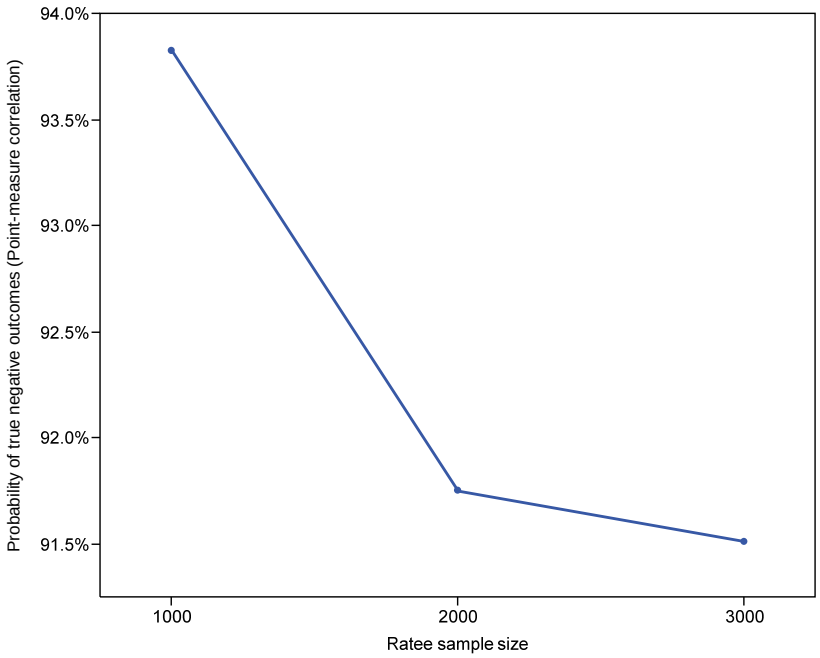

Figure 6.4-1. Probabilities of the true negative outcomes for the point-measure correlation as a function of number of ratees

The interaction of the double-scoring rate and the rater sample size had a negative impact on the true negative outcomes for the point-measure correlation (Figure 6.4-2) and the ZMSU (Figure 6.4-3). While the probability held about constant with 100% of responses double scored, the classification accuracy elevated about 10% for both indices as the rater increased its sample size five times with 10% of responses double scored.

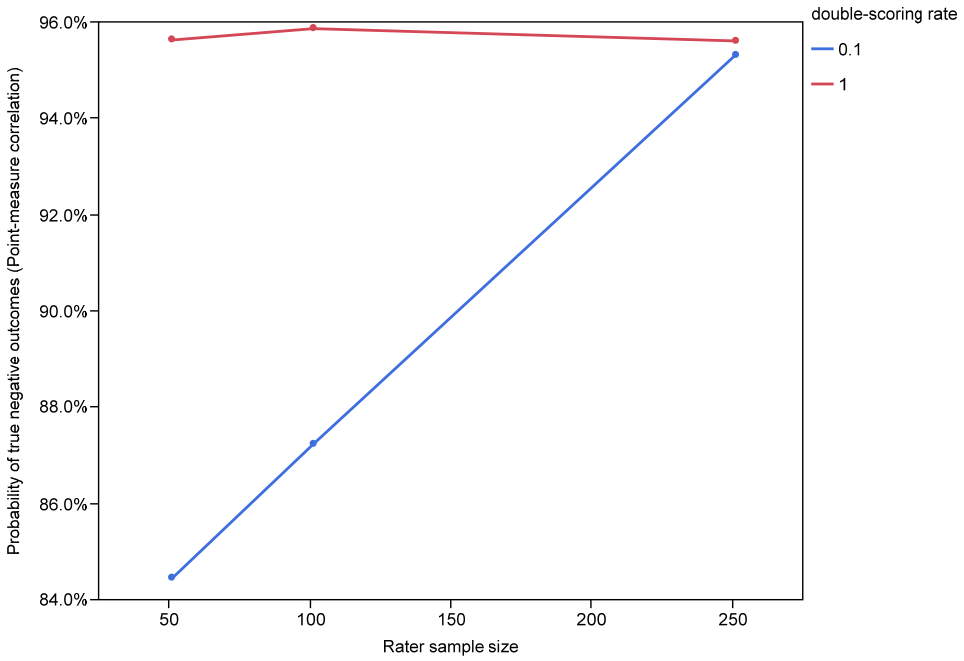

Figure 6.4-2. Probabilities of the true negative outcomes for the point-measure correlation as a function of number of rater and double-scoring rate

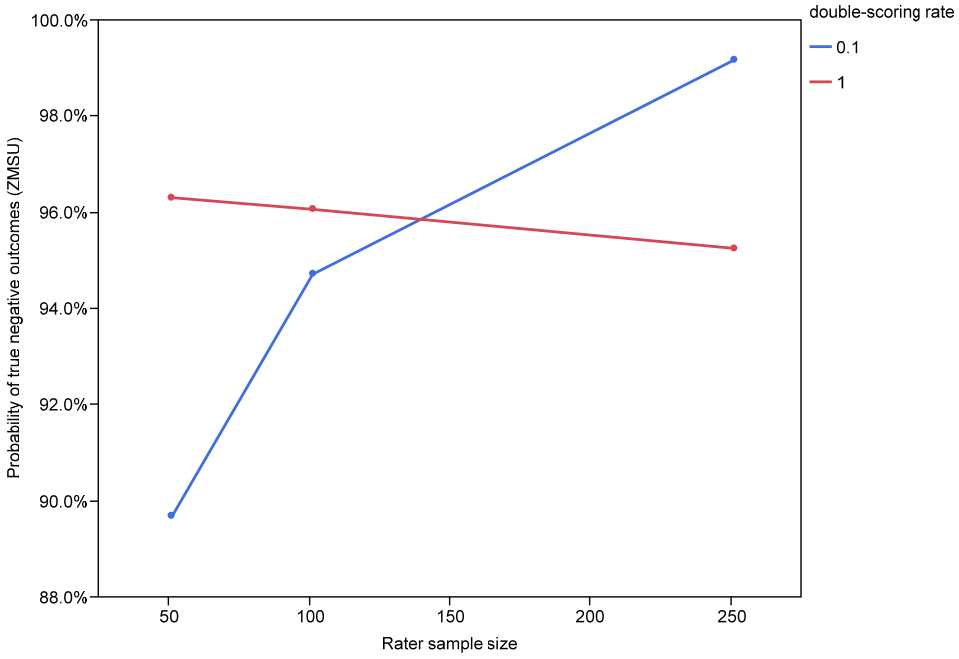

Figure 6.4-3. Probabilities of the true negative outcomes for the ZMSU as a function of double-scoring rate and number of raters

The interaction of the double-scoring rate and the ratee sample size had a positive impact for the measure-residual correlation (Figure 6.4-4), the expected-residual correlation (Figure 6.4-5) and the ZMSU (Figure 6.4-6). While the probability increased gradually within a range of 1% with 100% of responses double scored, the classification accuracy decreased about 2% for the two correlation indices and about 7% for the ZMSW with 10% of responses double scored.

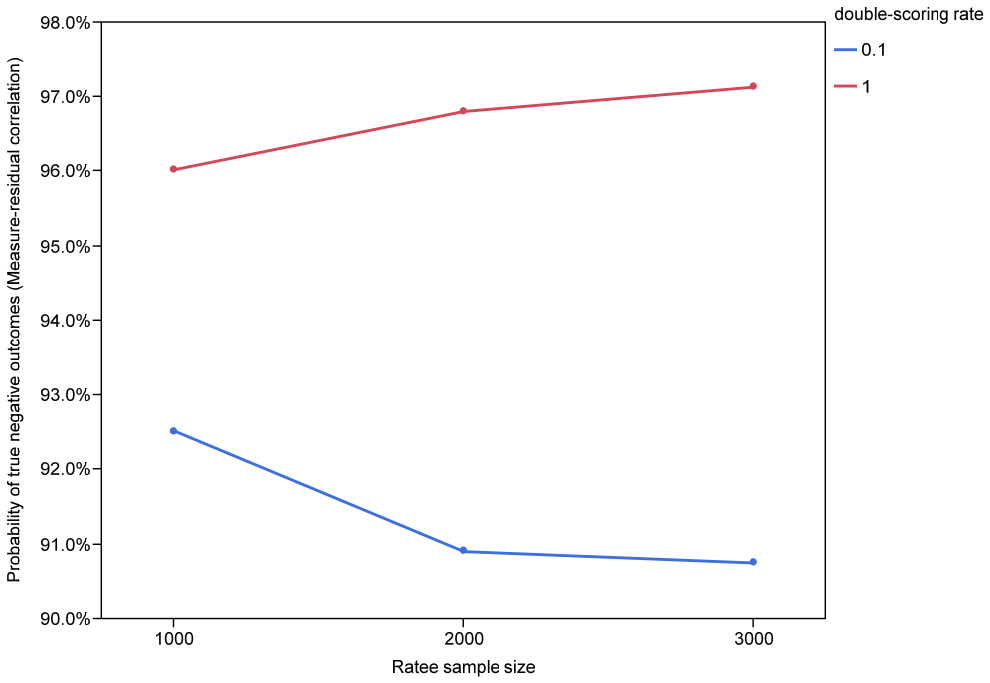

Figure 6.4-4. Probabilities of the true negative outcomes for the measure-residual correlation as a function of double-scoring rate and number of ratees

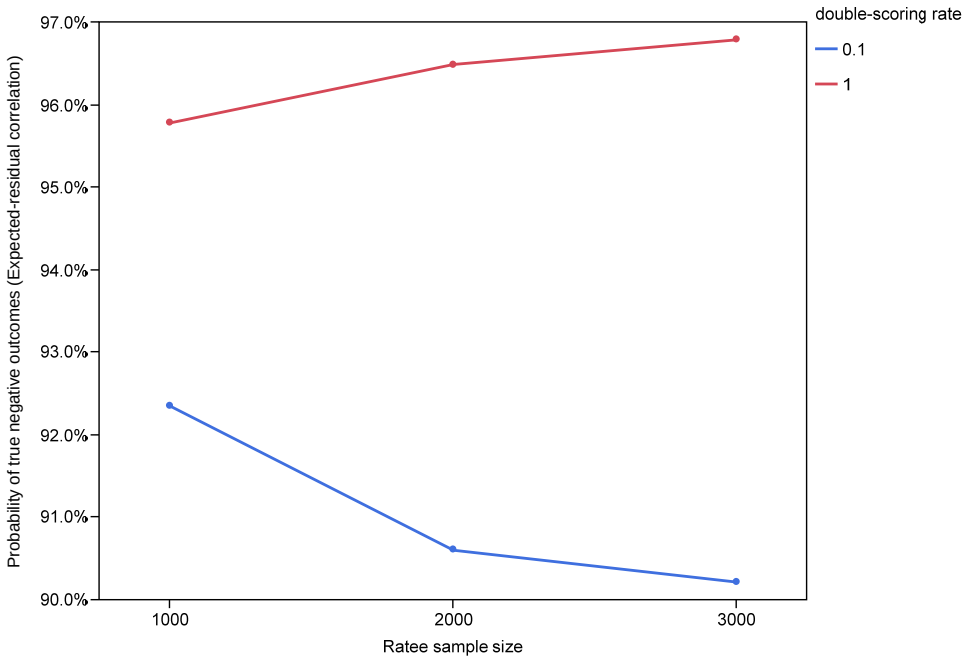

Figure 6.4-5. Probabilities of the true negative outcomes for the expected-residual correlation as a function of double-scoring rate and number of ratees

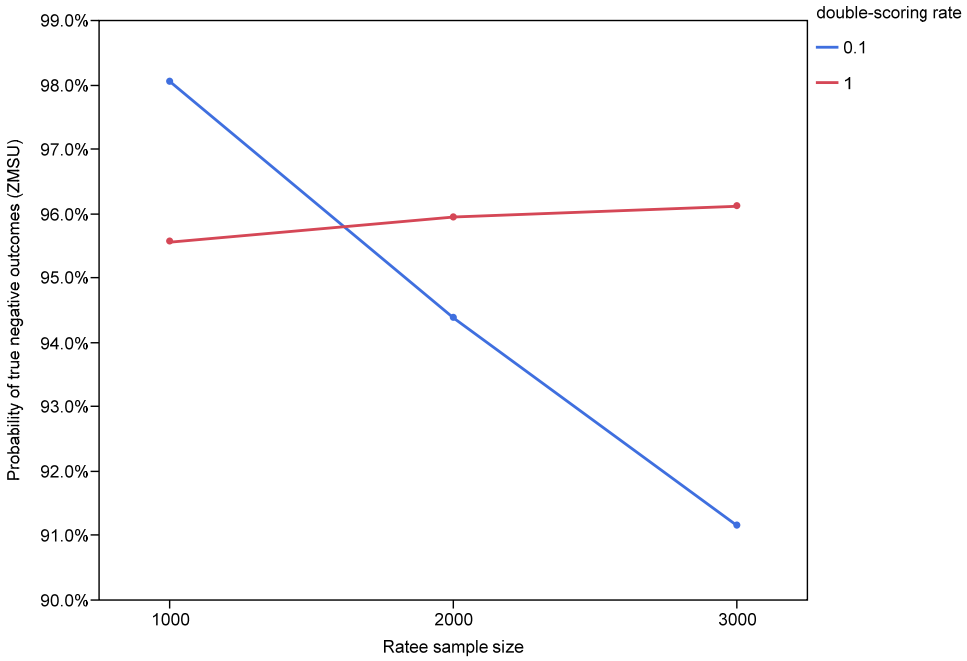

Figure 6.4-6. Probabilities of the true negative outcomes for the ZMSU as a function of double-scoring rate and number of ratees

The interaction of the rater and the ratee sample sizes exhibited a negative impact for the expected-residual correlation (Figure 6.4-7). The interaction of the double-scoring rate and the rater and the ratee sample sizes exhibited a positive impact for the ZMSW (Figure 6.4-8). With 100% of responses double scored, the probability seemed stable with little fluctuation across changes of the rater and ratee sample sizes. With 10% of responses double scored, the probability exhibited a slightly increasing trend along the increase of the rater sample size and a decreasing trend along the increase of the ratee sample size.

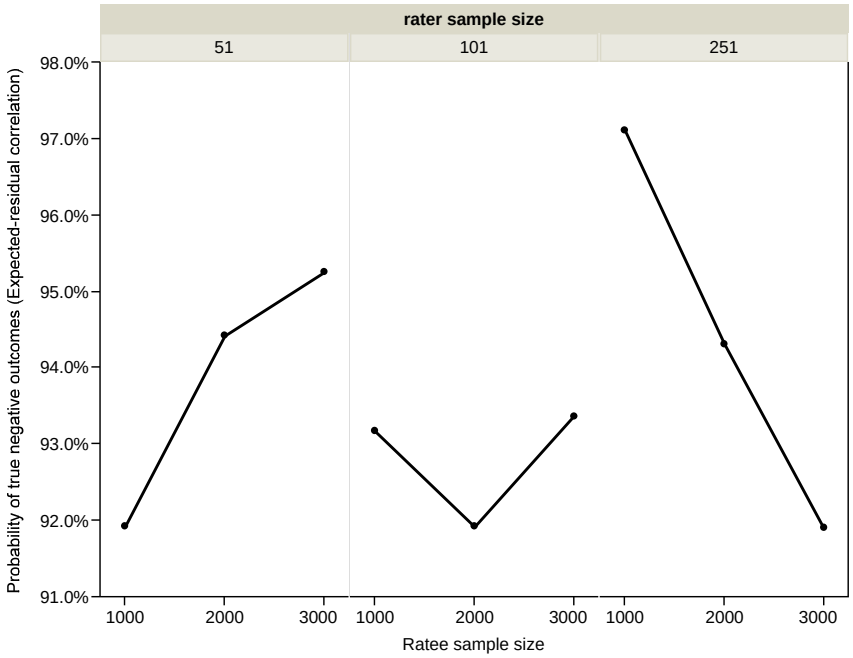

Figure 6.4-7. Probabilities of the true negative outcomes for the expected-residual correlation as a function of number of raters and ratees

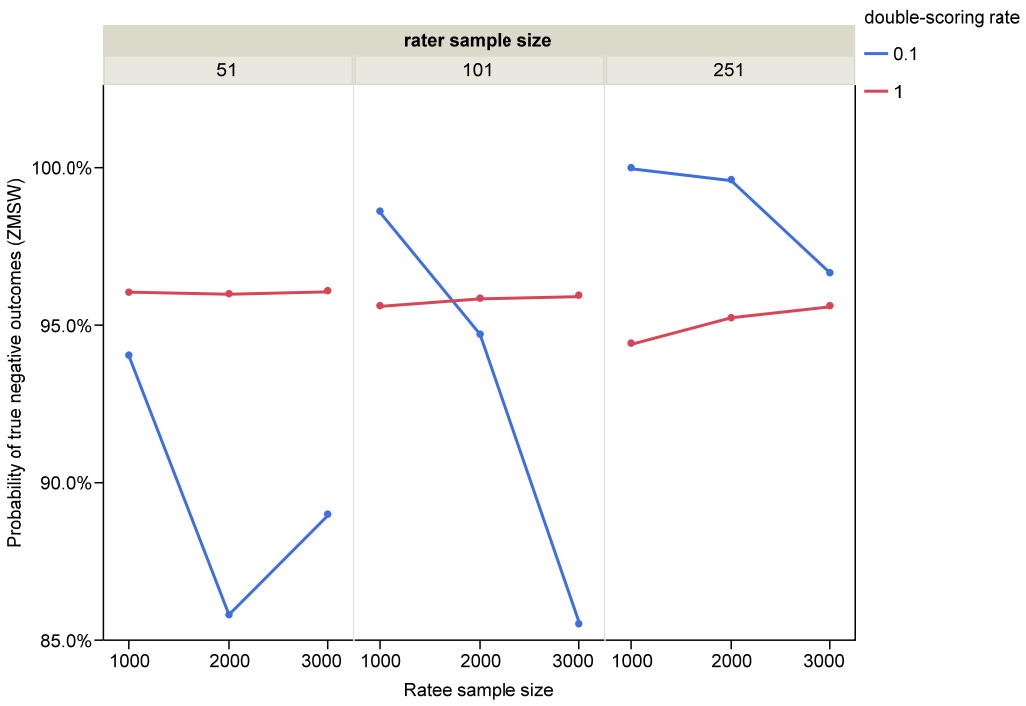

Figure 6.4-8. Probabilities of the true negative outcomes for the ZMSW as a function of double-scoring rate, number of raters and ratees

For the true positive outcomes, a noteworthy difference from the results of logistic regression for the true negative outcomes is the impact of the centrality strength either as a main effect or as an interacting variable with other experimental parameters. Each of the four experimental parameters as a main effect had a practical important impact for the STD (Figure 6.4-9 and 6.4-10). Specifically, with 10% of responses double scored, for an effect rater, the odds of being correctly classified is 99% more likely than that with 100% of responses double scored. One standard unit increase in the centrality strength results in a 50.9% increase in the odds of being correctly classified. One standard unit increase in the rater sample size results in 50.9% decrease in the odds of being correctly classified. One standard unit increase in the ratee sample size results in 56.8% increase in the odds of being correctly classified.

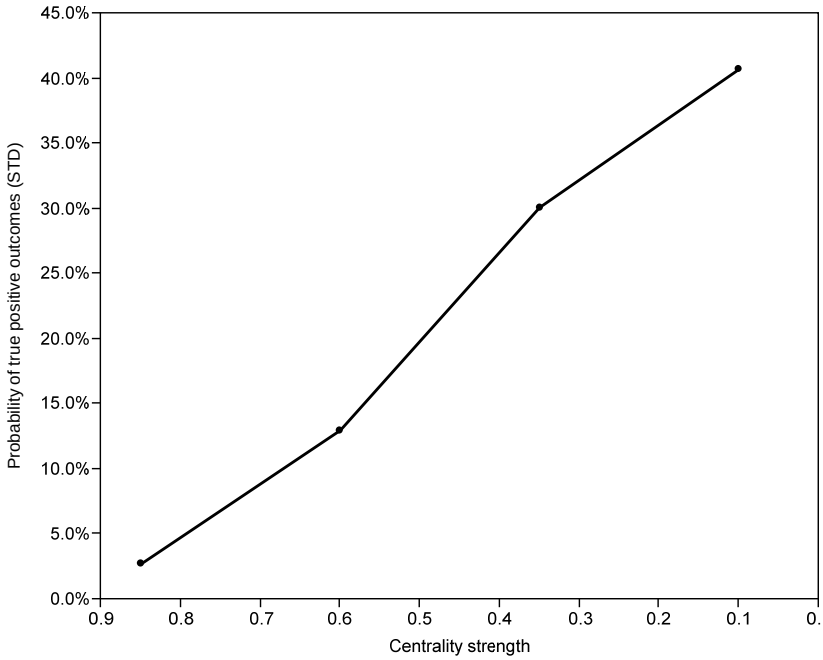

Figure 6.4-9. Probabilities of the true positive outcomes for the STD as a function of centrality strength

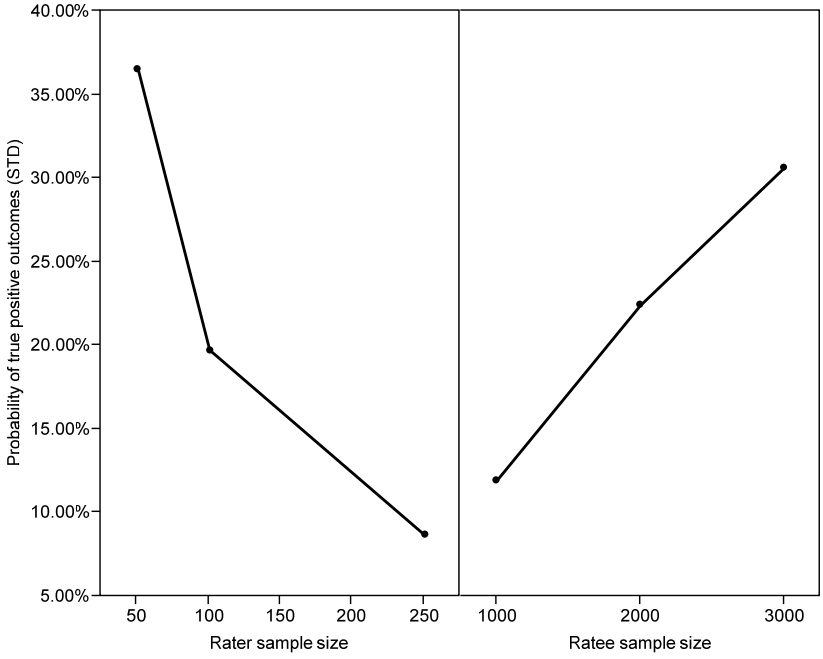

Figure 6.4-10. Probabilities of the true positive outcomes for the STD as functions of number of raters and of number of ratees

The rater and ratee sample sizes, as main effects, also had a substantial impact for the point-measure correlation (Figure 6.4-11) and the expected-residual correlation (Figure 6.4-12). Specifically, for an effect rater, one standard unit increase in the rater sample size results in 67.5% decrease in the odds of being correctly classified by the point-measure correlation and 34.8% decrease in the odds of being correctly classified by the expected-residual correlation. One standard unit increase in the ratee sample size results in 135.7% increase in the odds of being correctly classified by the point-measure correlation and 69.5% increase in the odds of being correctly classified by the expected-residual correlation.

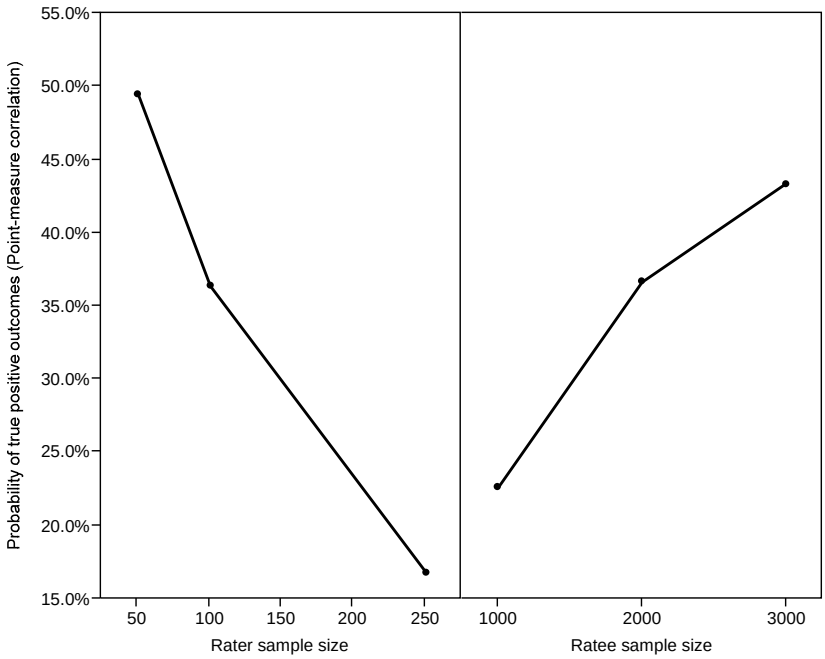

Figure 6.4-11. Probabilities of the true positive outcomes for the point-measure correlation as functions of number of raters and of number of ratees

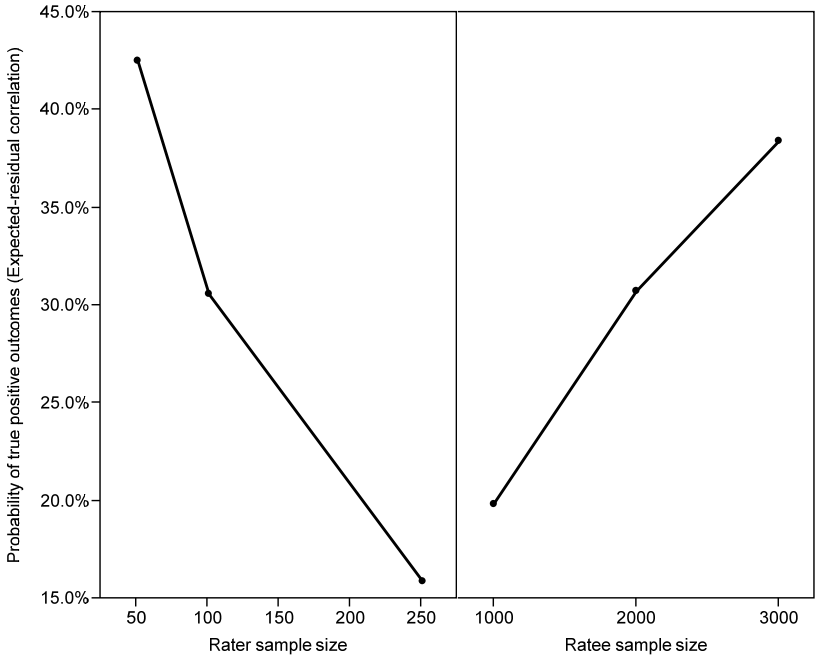

Figure 6.4-12. Probabilities of the true positive outcomes for the expected-residual correlation as functions of number of raters and of number of ratees

The interaction of the double-scoring rate and the centrality strength had a negative impact on the true positive outcomes for the point-measure correlation (Figure 6.4-13). With 100% of responses double scored, as the centrality strength intensified, the classification accuracy for effect raters elevated about 70%. Yet with 10% of responses double scored, the classification accuracy improved within a range of 15%.

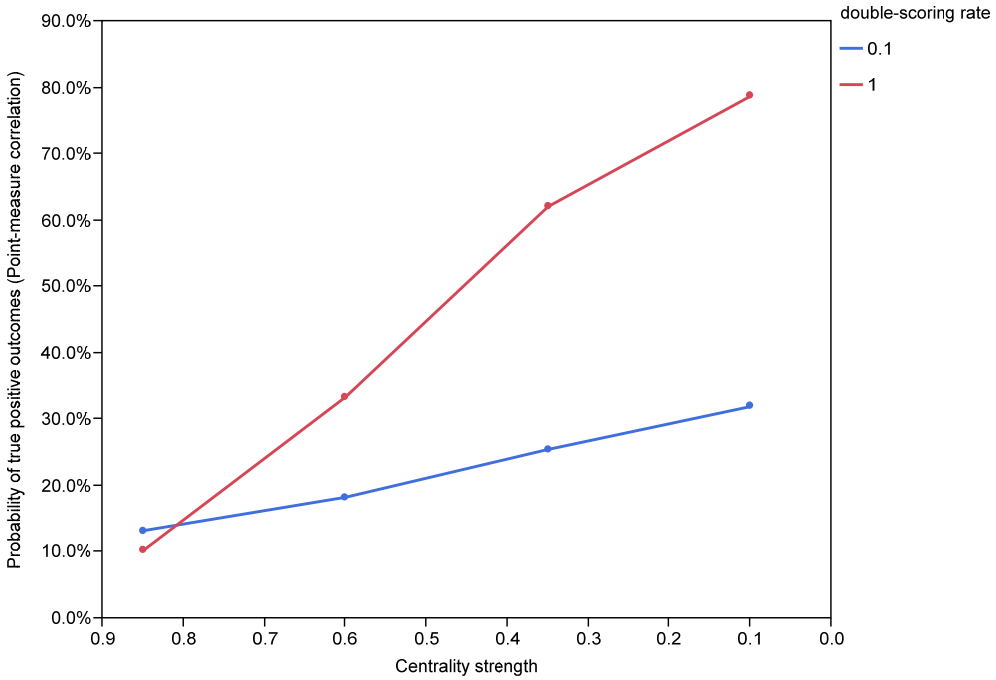

Figure 6.4-13. Probabilities of the true positive outcomes for the point-measure correlation as a function of double-scoring rate and centrality strength

The interaction of the double-scoring rate, the centrality strength and the rater sample size had a positive impact for the measure-residual correlation (Figure 6.4-14). The classification accuracy for effect raters increased as the centrality strength intensified but decreased as the rater sample size enlarged, and the discrepancy between the probabilities at the two double-scoring rates generally diminished as the rater sample size enlarged.

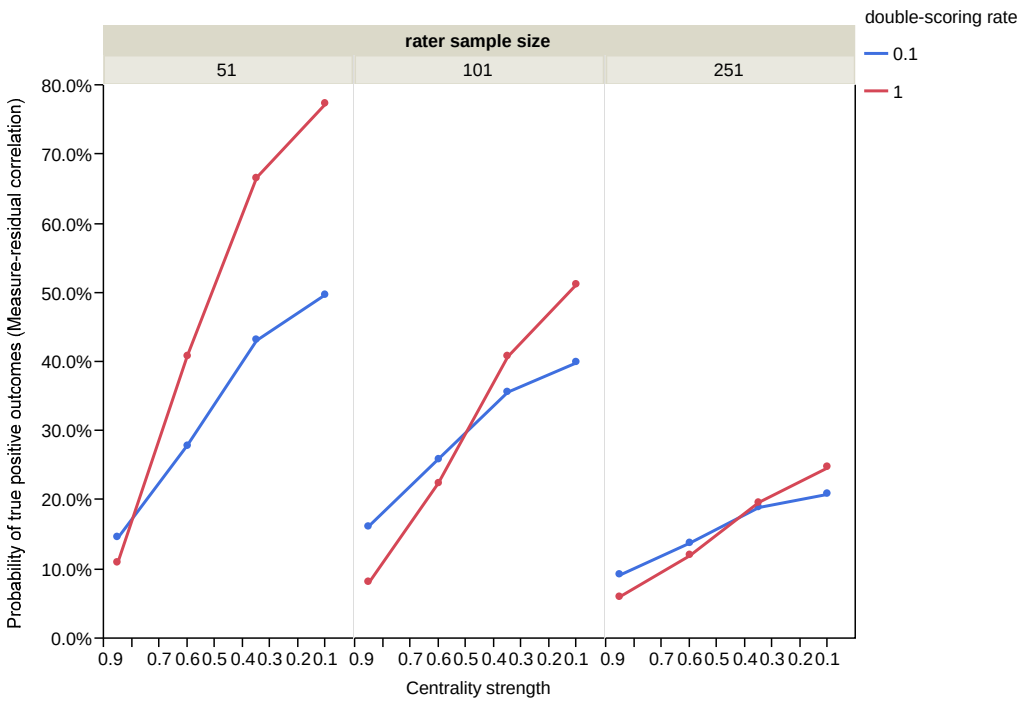

Figure 6.4-14. Probabilities of the true positive outcomes for the measure-residual correlation as a function of double-scoring rate, centrality strength and number of raters

The interaction of the double-scoring rate and the rater and the ratee sample sizes exhibited a negative impact for the ZMSW (Figure 6.4-15) and ZMSU (Figure 6.4-16). The classification accuracy for effect raters decreased as the rater sample size increased but increased as the ratee sample size enlarged, and the discrepancy between the probabilities at the two double-scoring rates generally enlarged along the increment of the rater and ratee sample sizes.

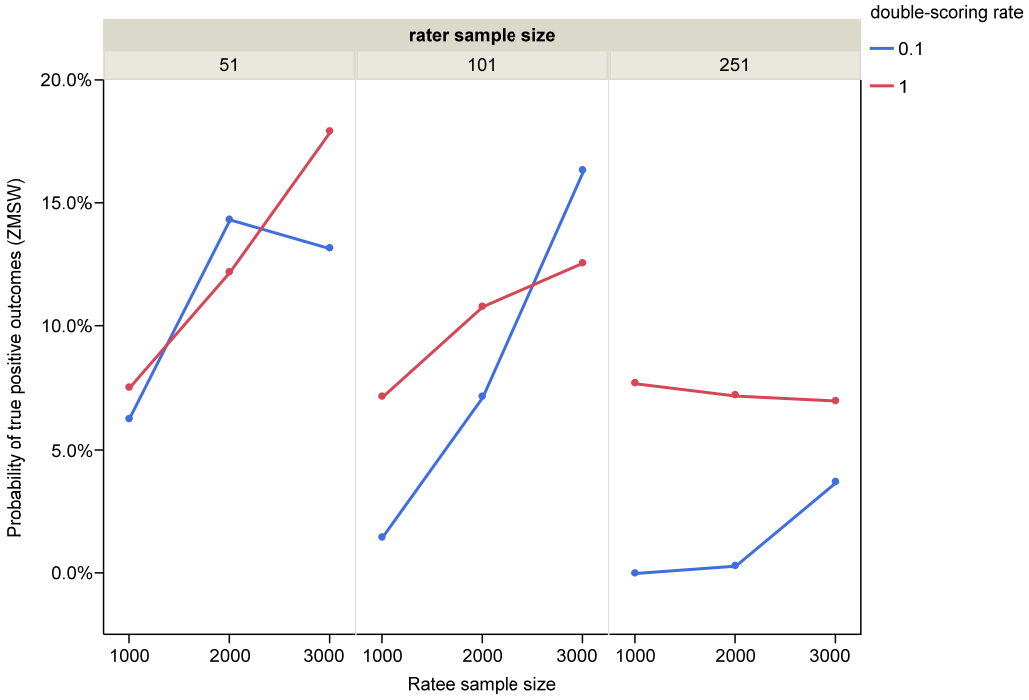

Figure 6.4-15. Probabilities of the true positive outcomes for the ZMSW as a function of double-scoring rate, number of raters and ratees

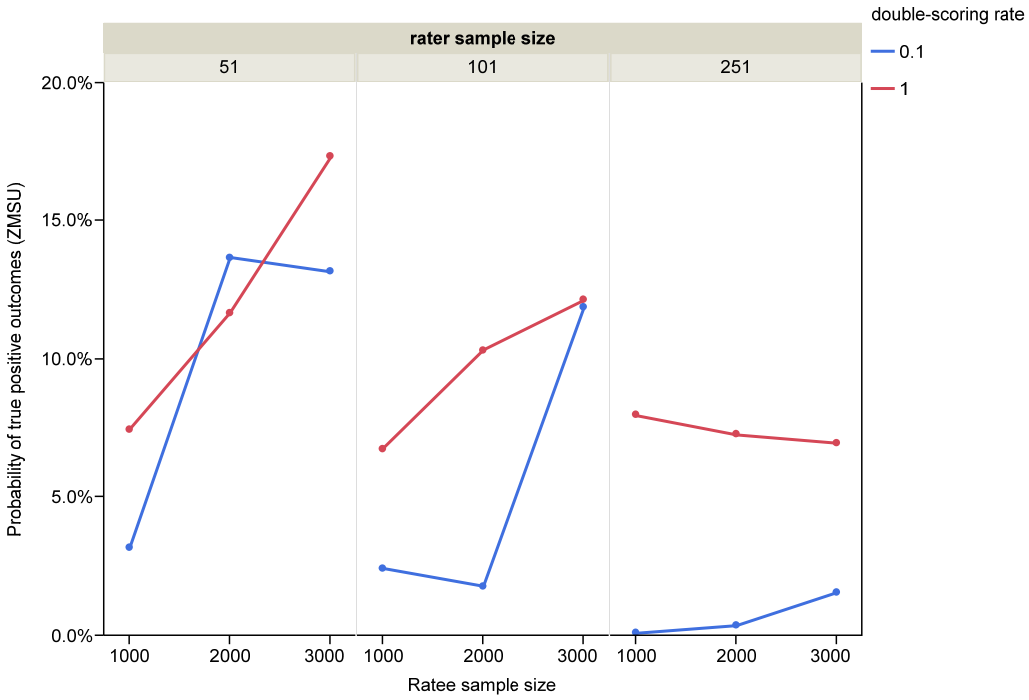

Figure 6.4-16. Probabilities of the true positive outcomes for the ZMSU as a function of double-scoring rate, number of raters and ratees

6.5 Conclusions

The results indicate that these experimental parameters and their interactions had a different impact on the classification accuracy for non-effect raters versus the classification accuracy for effect raters. These centrality-detection indices are expected to be sensitive to the change of centrality strength for effect raters but not for non-effect raters, and this is what the results portray. Among the non-effect raters, the centrality strength appeared to have no important influence on the outcomes of rater classification for all the indices. That is, correctly classifying non-effect raters is independent of the existence of the rater centrality effect, which is what one would hope for. On the other hand, with the effect raters, the centrality strength had a substantial impact with moderate or large effect sizes on the outcomes of rater classification for some of the indices (i.e., the point-measure-residual correlation, the measure-residual correlation and the standard deviation of assigned scores). That is, for the involved indices, correctly classifying effect raters is closely related to the strength of the rater centrality effect. Again, this is what the data analyst would hope for, and this finding is consistent with my previous studies that these indices were found to be sensitive to rater types, and that their statistical power rates varied across the levels of centrality strength.

The two levels of double-scoring rate introduce different amount of missing values into the simulated data sets. The logistic regression analyses showed that the double-scoring rate had a strong influence on both the true negative and positive outcomes either by acting itself (i.e., as a main effect) or by interacting with the other experimental parameters. Consistent with my previous studies, if we compare at the same group levels, in most cases, these indices performed

better when 100% of responses received a second score than when only 10% of responses received a second score.

The rater sample size exhibited a negative impact as a main effect on the true positive outcomes for the point-measure correlation, expected-residual correlation and the standard deviation of assigned scores. For these three indices to better correctly classify effect raters, the group of raters should be restricted to a smaller size, meaning that it is easier to detect effect raters when they make up a larger proportion of the rater pool. This finding suggests that smaller rater sizes are needed for accurate detection of the centrality effect. Meanwhile, the ratee sample size asserted a positive impact as a main effect on the true positive outcomes for the same three indices. For these three indices to better correctly identify non-effect raters, more ratees should be included in a scoring group. This finding suggests that larger ratee sizes are needed for accurate detection of the centrality effect. Again, this is what one would expect—in order to obtain more accurate information about raters, they should rate more ratees. If we translate the relationship of the number of raters and ratees into a rater-ratee ratio (i.e., rater sample size divided by ratee sample size), we can simply say that lower rater-ratee ratio would result in more accurate detection of the centrality effect or higher statistical power, that higher rater-ratee ratio would lead to more accurate identification for non-effect raters, and that achieving an accurate detection rate on both outcomes, i.e., high accuracy in true positive outcome and true negative outcome, seems impossible. In operational scoring settings, if raters are paid by hour, a more common case in testing industry, we can imagine two scenarios to imitate the dilemma discussed above. One scenario is that 10 raters are hired for a single-item scoring project containing 1000 ratees and they spend 100 hours to finish the whole project. Another is that 100 raters are hired

for the same scoring project and they spend 10 hours to finish the whole project. The total costs are the same for the two scenarios. However, the former scenario with a rater-ratee ratio of 1/100 would achieve better statistical power than the latter with a rater-ratee ratio of 1/10 and consequentially enhance the rating quality. On the other hand, the latter scenario would be more conservative in terms of picking up "bad" raters and consequentially lower the cost in rater retraining or replacing disqualified raters if necessary. It turns out that to use these indices to detect the centrality rater effect, it is important to be aware that improving the classification accuracy comes at a cost of reducing the classification accuracy for non-effect raters, and that improving the classification accuracy for non-effect raters comes at a cost of reducing the classification accuracy for effect raters.

Combing the directions of relationship presented in Table 6.4-1 and 6.4-2 and the perplexing pattern discussed above, I suggest some simple guidelines (Table 6.5-1) for the direction of the expected impact on classification accuracy from either main effects or higher order interactions if they are statistically significant and practically important. Although the guidelines do not clearly solve the dilemma of knowing what to expect in more complex applied settings, they do offer a glimpse of the "pros" and "cons" in adjusting the magnitude of the parameters when we evaluate the impact of the four experimental parameters on the accuracy of rater classification.

Table 6.5-1: *Simple guidelines for the expected impact of classification accuracy*

Effect	Direction of impact on the true negative outcomes	Direction of impact on the true positive outcomes
Double-scoring rate 0.1 vs. 1.0	-	+
Centrality strength	N/A	-
Rater sample size	+	-

Ratee sample size	-	+

For example, if the double-scoring rate and the centrality strength are shown to conjunctly influence the accuracy of rater classification in a scoring project, we would expect the interaction of these two effects to have a negative overall impact on the true positive outcomes of rater classification (i.e., positive multiplied by negative results in negative). The stronger the centrality effect is, the higher the proportion of correct in effect-rater classifications will we achieve (e.g., the point-measure correlation). As another example, if the double-scoring rate and the rater and ratee sample sizes are shown to conjunctly influence the accuracy of rater classification, we would expect the highest-order interaction of these three effects, if with substantial importance, to have a positive overall impact on (or positive relationship with) the true negative outcomes and to have a negative overall impact on (or negative relationship with) the true positive outcomes of rater classification. To enhance the proportion of correct classification for non-effect rater (i.e., the true negative outcomes), the group size of raters needs to be enlarged and the group size of ratees needs to be decreased (e.g., ZMSW). To enhance the classification accuracy for effect-rater classification (i.e., the true positive outcomes), the group size of raters needs to be decreased and the group size of ratees needs to be enlarged (e.g., ZMSW and ZMSU). In essence, as I pointed out earlier, improving the classification accuracy for non-effect raters may come at a cost of reducing the classification accuracy for effect raters and vice versa.

A limitation of this study is the fact that the impact evaluated in this simulation study is much less complex than the “reality”. Confounded with the four experimental parameters, numerous factors such as rater psychological traits (e.g., motivation, anxiety, motivational intensity, achievement, self-efficacy and mood. etc.) and other types of rater effect (e.g.,

severity/leniency, halo, and range restriction) can very likely affect the rater classification accuracy. These unsimulated factors inevitably place constraints on the generalizability of the results. Hence, application of these rater centrality-detection indices to real data sets and to more complex simulated data may provide additional support. Regardless, the bottom line is that this study opens up new opportunities in minimizing systematic errors in rating data by adjusting rater classification accuracy through influential factors.

6.6 References

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

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

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

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Linacre, J. M. (2009b). WINSTEPS Rasch measurement computer program (Version 3.68.0). Chicago, IL: Winsteps.com.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, amd standarization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92-109.

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.

Wolfe, E. W. (1998). *A two-parameter logistic reader model (2PLRM): Detecting reader harshness and centrality.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

Wolfe, E. W. (2005). Identifying Rater Effects in Performance Ratings. In S. Reddy (Ed.), *Performance Appraisals: A Critical View* (pp. 91-103). Hyderabad, India: ICFAI University Press.

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects with a multi-faceted Rasch rating scale model. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147-164). Stamford, CT: Ablex.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: generalised item response modeling software (Version 2.0). Camerwell, Victoria: ACER Press.

7 DISCUSSION

The results from the three studies summarized in this dissertation indicate that (1) different rater centrality indices have different levels of sensitivity when differentiating effect raters from normal raters; (2) all of the indices provided reasonable protection against Type I error when all responses were double scored but provide low statistical power unless the rater centrality effect is quite strong; and (3) the rater classification accuracy is influenced by rating context variables, and improving the classification accuracy for non-effect raters may come at a cost of reducing the classification accuracy for effect raters. In general, these three studies showed that the traditional practice of focusing on mean-squared fit indices as indicators of rater centrality is not warranted and that better options exist when selecting a rater centrality index. However, the choice of which index to adopt depends on the desired Type I error rate and various characteristics of the rating context.

The point-measure correlation may have limited usefulness for rater centrality detection because it seemed to be strongly influenced by missing data, such as that created by the double-scoring practices adopted in many operational contexts. The point-measure correlation only worked well in differentiating rater types on the centrality effect when 100% of the ratees receive second scores, and the capacity of the point-measure correlation for separating rater types was diminished when only 10% of ratees receive second scores. Its Type I error rates tended to be larger when 10% of responses were double scored, and its Type I error rates were quite inflated under this condition. Higher statistical power was achieved when responses were 100% double scored in comparison to only 10% being double scored. I suggest using the point-measure correlation only when missing data is minimal and when the connections between raters is

relatively “thick”. Hence, in practical applications, it may be useful to utilize the point-measure correlation only when Type I errors are of less importance. In other words, in a situation where incorrectly identifying normal raters as effect raters is more of concern than successfully identifying all effect raters, we may want to avoid using the point-measure correlation, since in such a case care is usually focused on minimizing the occurrence of statistical errors. Moreover, the accuracy of rater classification using this index, for non-effect raters, was influenced by the number of ratees and the interaction of the double-scoring rate and the number of raters. For effect raters, the accuracy was largely influenced by the number of raters and ratees and the interaction of the double-scoring rate and the centrality strength.

The measure-residual and the expected-residual correlations, two very similar indices, appear to be promising as indicators of the centrality effect. Because these indices have not been carefully studied in previous research, the research reported in this dissertation provides an important contribution to our understanding of how analysts can best detect centrality effects. These two indices usefully differentiated effect raters from non-effect raters and were sensitive to the magnitude of the strength of the effect. They also provided better identification of the centrality effect, comparing to other indices. The Type I error rates for these two correlation-based indices tended to be relatively large when only 10% of responses were double scored. However, they also achieved higher statistical power than did most of the remaining indices when responses were 100% double scored. From the perspectives of statistical errors, I would recommend using either the measure-residual correlation or the expected-residual correlation for detecting the centrality effect with a consideration on balancing both Type I error and statistical power. Besides, the accuracy of rater classification using these two indices was influenced by the

four experimental parameters in a similar way with some variations possibly due to the difference in index calculations and the randomness in data simulation. For the classification of accuracy for non-effect raters, both indices were influenced by the interaction of the double-scoring rate and the number of ratees. For the classification of accuracy for effect raters, the inaction of the double-scoring rate, the centrality strength and the number of raters largely influenced the measure-residual correlation but not the expected-residual correlation.

The standard deviation of assigned scores differentiated rater types as well as the measure-residual and the expected-residual correlations. This raw score was among the most sensitive in terms of differentiating effect raters from non-effect raters. It provided the greatest protection against Type I error (perhaps too much so) and acceptable statistical power when the centrality effect was strong. Moreover, this index seemed not to be strongly impacted by rating context variables in terms of rater classification, making it useful across a variety of rating designs. However, the standard deviation can be easily inflated when random error exists in the ratings, so further research is needed to determine the potential limitations of this, and other, rater centrality indices when it comes to differentiating different kinds of rater effects. Also, for the classification of accuracy for non-effect raters, this index seems to be robust to the variations in the four experimental parameters. For the classification of accuracy for effect raters, however, it was substantially affected by each of the parameters.

As mentioned previously, the mean-square fit statistics are widely used to detect rater effects. However, when it comes to identify the centrality effect, these indices do not seem to provide useful information when it comes to identifying rater centrality. Another important contribution of this dissertation is the fact that the studies summarized here point out an

important weakness in current psychometric practice relating to the detection of rater effects. Along with the raw score standard deviation, the two standardized mean-square fit indices provided excellent protection against Type I error but poor statistical power. Their capacity for rater classification is also restricted by double-scoring rate, rater sample size, and ratee sample size, which makes this type of indices unreliable when it comes to detecting the centrality effect. This type of indices is also shown to be influenced by the double-scoring rate and the number of raters or ratees but not the centrality strength. Therefore, I suggest abandoning the current practice of relying on this index in operational settings. Similarly, the rater slope index, an index that has not been evaluated as an indicator of rater centrality to date, was shown to be of limited usefulness in differentiating rater types on the effect.

A limitation of this dissertation is the fact that the four-parameter experimental design in this simulation study is much less complex than the "reality". Confounded with the four experimental parameters, numerous factors such as rater psychological traits (e.g., motivation, anxiety, motivational intensity, achievement, self-efficacy and mood. etc.) and other types of rater effect (e.g., severity/leniency, halo, and range restriction) can very likely affect the rater classification accuracy. The simulated data are extremely "clean," and therefore it is difficult to determine how well the simulation results will generalize to the "messy" data that are typical of operational contexts. Thus, these unmodeled factors inevitably place constraints on the generalizability of the results.

Another limitation is that I did not control the number of ratees per rater and made the rating-count variable vary across raters with random assignment technique in the data generating process. Because this variable could affect the measurement error of the centrality indices and

any associated hypothesis test based on those indices, it is hard to draw any conclusion on the appropriate number of ratees per rater for a particular centrality index to work well from this simulation study. Therefore, rather than letting it change with rater-response random assignment in the simulation process, the ideal case would be making the rating-count variable manipulatable.

A contextual limitation of this dissertation is that the indices studied in this dissertation are probably only relevant and important in high-stakes applications of highly trained raters. One focus of the Race to the Top effort has been to promote teacher scoring as a professional development activity. It is likely that teachers would approach the task of scoring student work in a very different manner than do "professional" scorers such as the ones that work at Pearson and other testing companies (refer to Gambell and Hunter's work (2004)). As a result, they may utilize different approaches to scoring, may respond to feedback differently, and may require different types of intervention in order to increase the quality of their scores to a level that would permit high-stakes decisions to be made based upon the scores that they assign.

Future research in this line can expand to several potentially important areas. (1) Apply these rater centrality-detection indices to real data sets and to more complex simulated data. Generating and estimating the data based on more complex models like the two-parameter and the three-parameter logistic models (2PL and 3PL) that involve item discrimination and guessing parameters, or introducing multidimensionality into the data would be of great interest. Furthermore, the underlying Rasch I focused on in this dissertation is a single-facet model (i.e., one item-by-multiple raters design). The multi-faceted model would be required (and likely a hierarchical version of that model) if one was interested in detecting rater effects in contexts in

which multiple items are scored or multiple features of a single response are scored (i.e., trait scoring). (2) Embedding recently emerging research topics on measuring rater effect drift over time (Congdon & McQueen, 2000; Wolfe & Moulder, 2001) would be another valuable contribution to the field, as rating patterns may not be stable over time due to many factors such as circumstance changes, psychological changes, etc. (3) Compare rater centrality-detection indices with rater inaccuracy-detection indices. A rater who exhibits the inaccuracy effect makes seemingly random errors which reduce the usefulness of the assigned ratings as indicators of the ratee's true proficiency. Rater inaccuracy may occur when the rater has insufficient content background, insufficient training in applying the scoring rubric, or immutable biases or content-based beliefs that cause the rater to assign scores that are not consistent with the scoring rubric. When a rater assigns inaccurate ratings, any ratee scored by this rater could be misplaced on the underlying ability scale, and it is likely that the rank ordering of ratees scored by the effect rater disagree with accurate raters, which makes its manifestation similar to the rater centrality effect. As a result, the inaccuracy effect may manifest itself in residual-based statistics (e.g., mean-square fit statistics) as well as correlation-type indices (e.g., the point-measure correlation, the measure-residual correlation, and the expected-residual correlation). Evaluating the capacity of these indices that I studied in this dissertation for differentiating rater effects is an important next step in research in this effort. (4) This dissertation research focuses on detecting rater effects, and the results of the study can be directly applied to the monitoring and remediation of raters in operational settings. However, the study does not provide information to help those working in applied settings to understand why those effects exist. Hence, another extension of the work I have done would be to determine how the manifestation of rater effects coincides with raters' training experience, background, and rater cognition.

REFERENCES

(Not included in Chapter 4, 5 or 6)

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System, 32*(3), 407-425.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC 20036: American Educational Research Association.

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

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

Bernardin, H. J. (1978). Effects of Rater Training on Leniency and Halo Errors in Student Ratings of Instructors. *Journal of Applied Psychology, 63*(3), 301-308.

Bernardin, H. J., & Pence, E. C. (1980). Effects of Rater Training: Creating New Response Sets and Decreasing Accuracy. *Journal of Applied Psychology, 65*(1), 60-66.

Bernardin, H. J., & Villanova, P. (2005). Research Streams in Rater Self-Efficacy. *Group & Organization Management, 30*(1), 61-88.

Bernardin, H. J., & Walter, C. S. (1977). Effects of Rater Training and Diary-Keeping on Psychometric Error in Ratings. *Journal of Applied Psychology, 62*(1), 64-69.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*, 238-252.

Borman, W. C. (1979). Format and Training Effects on Rating Accuracy and Rater Errors. *Journal of Applied Psychology, 64*(4), 410-421.

Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology, 60*, 561-565.

Cascio, W. F. (1982). *Applied psychology in personnel management* (2 ed.). Reston, VA: Reston Publishing.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Congdon, P. J., & McQueen, J. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement, 37*(2), 163-178.

Cooper, W. H. (1981). Conceptual similarity as a source of illusory halo in job performance ratings. *Journal of Applied Psychology, 66*(3), 302-307.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper and Row.

DeCotiis, T. A. (1977). An Analysis of the External Validity and Applied Relevance of Three Rating Formats. *Organizational Behavior & Human Performance, 19*(2), 247-266.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

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

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

Fisicaro, S. A., & Lance, C. E. (1990). Implications of Three Causal Models for the Measurement of Halo Error. *Applied Psychological Measurement, 14*(4), 419.

Fried, Y., Levi, A. S., Ben-David, H. A., Tiegs, R. B., & Avital, N. (2000). Rater positive and negative mood predispositions as predictors of performance ratings of ratees in simulated and real organizational settings: Evidence from US and Israeli samples. *Journal of Occupational & Organizational Psychology, 73*(3), 373-378.

Gambell, T., & Hunter, D. (2004). Teacher scoring of large-scale assessment: professional development or debilitation? *Journal of Curriculum Studies, 36*(6), 697-724.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the Accuracy of Performance Evaluations: Comparison of Three Methods of Performance Appraiser Training. *Journal of Applied Psychology, 73*(1), 68-73.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403-424.

Huot, B. A. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Cresskill, NJ: Hampton Press.

Kim, S. C., & Wilson, M. (2009). A Comparative Analysis of the Ratings in Performance Assessment Using Generalizability Theory and The Many-Facet Rasch Model. *Journal of Applied Measurement, 10*(4), 40-423.

Korman, A. K. (1971). *Industrial and organizational psychology*. Englewood Cliffs, NJ: Prentic-Hall.

LaHuis, D. M., & Avis, J. M. (2007). Using Multilevel Random Coefficient Modeling to Investigate Rater Effects in Performance Ratings. *Organizational Research Methods, 10*(1), 97-107.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance*. New York: Academic Press.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: theory into practice* (Vol. 3, pp. 85-112). Norwood, NJ: Alex Publishing Corporation.

Linacre, J. M. (2009a). Facets Rasch measurement computer program (Version 3.66.0). Chicago, IL: Winsteps.com.

Linacre, J. M. (2009b). WINSTEPS Rasch measurement computer program (Version 3.68.0). Chicago, IL: Winsteps.com.

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. [Article]. *Language Testing, 15*(2), 158-180.

May, G. L., & Gueldenzoph, L. E. (2006). The effect of social style on peer evaluation ratings in project teams. *Journal of Business Communication, 43*(1), 4-20.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, amd standarization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92-109.

Mount, M. K., & Systma, M. R. (1997). Rater-ratee race effects in developmental performace ratings of managers. *Personnel Psychology, 50*(1), 51-69.

Murphy, K. R., & Anhalt, R. L. (1992). Is halo error a property of the raters, ratees, or the specific behaviors observed? *Journal of Applied Psychology, 72*(4), 494-500.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*(4), 619-624.

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

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. W. B. A. Huot (Ed.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

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.

Scherer, R. F., Owen, C. L., & Brodzinski, J. D. (1991). Rater and ratee sex effects on performance evaluations in a field setting. *Management Communication Quarterly, 5*(2), 174.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Smith, E. V. (2004). An Application of Generalizability Theory and Many-Facet Rasch Measurement Using a Complex Problem-Solving Skills Assessment *Educational & Psychological Measurement, 64*(4), 617-639.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A Comparison of Generalizability Theory and Many-Facet Rasch Measurement in an Analysis of College Sophomore Writing. *Assessing Writing, 9*(3), 239-261.

Tziner, A., Murphy, K. R., Cleveland, J. N., Beaudin, G., & Marchand, S. (1998). Impact of rater beliefs regarding performance appraisal and its organizational context on appraisal quality. *Journal of Business & Psychology, 12*(4), 457-468.

Van Iddekinge, C. H., Putka, D. J., Raymark, P. H., & Eidson, C. E., Jr. (2005). Modeling Error Variance in Job Specification Ratings: The Influence of Rater, Job, and Organization-Level Factors. *Journal of Applied Psychology, 90*(2), 323-334.

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

Wen, S.-S. (1979). Racial Halo on Evaluative Rating: General or Differential? *Contemporary Educational Psychology, 4*(1), 15-19.

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

Williamson, M. M. (1988). A model for investigating the functions of written language in difference disciplines. In D. A. Jolliffe (Ed.), *Advances in writing resarch* (Vol. 2, pp. 89-132). Norwood, MJ: Ablex.

Wolfe, E. W. (1998a). *A two-parameter logistic reader model (2PLRM): Detecting reader harshness and centrality.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wolfe, E. W. (1998b). *A two-parameter logistic reader model (2PLRM): Detecting reader harshness and centrality.* Paper presented at the the annual meeting of the American Educational Research Association, San Diego, CA.

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

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

Wolfe, E. W. (2005). Identifying Rater Effects in Performance Ratings. In S. Reddy (Ed.), *Performance Appraisals: A Critical View* (pp. 91-103). Hyderabad, India: ICFAI University Press.

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects with a multi-faceted Rasch rating scale model. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147-164). Stamford, CT: Ablex.

Wolfe, E. W., & Feltovich, B. (1994). *Learning to rate essays: A study of scorer cognition*. Paper presented at the American Educational Research Association.

Wolfe, E. W., & Moulder, B. C. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*(3), 256-280.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: generalised item response modeling software (Version 2.0). Camerwell, Victoria: ACER Press.

APPENDIX: SAS code for sample data generation

Part I: Main code

```
*************************************************************************************;
* Title:Script for study on centrality rater effect and inaccuracy
* Author:       Jessica Yue
* Version:      6.29
* Date:         06/29/2010
* Note:The data generating design takes missing data into account. Each ratee has one score
from one randomly selected rater and 10% of the ratees will have second score assigned by
another randomly selected rater. It can also be expanded to "(n-1)+nth@x%" design, each ratee
has n-1 scores from n-1 randomly selected raters, x% of the ratees will have another score
assigned by a randomly selected rater.
*************************************************************************************;

*** FILENAME MPRINT 'pathname and name of file';
*FILENAME MPRINT 'MACRO DUMP.SAS';
*OPTIONS MPRINT MFILE;
*****************************************************************;
option nodate linesize=100 noxwait xsync mlogic;

*Deletes all SAS files in the SAS data library;
proc datasets kill;
run;

*Suppress the detailed log;
filename junk dummy;
proc printto log=junk; run;

*Turn on the detailed log;
/*proc printto; run; */

%include 'thresholds.sas';
%include 'array.sas';

*Simulation parameter specifications;

%let study=1;           *1=study of centrality effect;
%let score_dist=2;      *1=uniform score distribution, 2=normal score distribution, 3=skewed
score distribution;
%let rating_scale_type=6;     *6=6pt-scale;
%let choose=2;                *minimum sores per ratee, choose*2<=nrater;
%let iteration=1000;
```

```
%array (select_rate, 0.1 1.0, scope=GLOBAL);
%array (strength, .1 .35 .60 .85, scope=GLOBAL);
%array (correl, .1 .35 .60 .85, scope=GLOBAL);
%array (nrater, 50 100 250, scope=GLOBAL);
%array (nperson, 1000 2000 3000, scope=GLOBAL);

%let delimiter=_;                    *for final report file name;
%let ext=.csv;

%macro sort(data, by_var);
 proc sort data=&data.;
  by &by_var.;
 run;
 quit;
%mend sort;

%macro sim();
 %if &rating_scale_type=4 %then %do;
          %if &score_dist=1 %then %call_sim(-0.67,0,-99,-99);
          %if &score_dist=2 %then %call_sim(-2.28,0,-99,-99);
          %if &score_dist=3 %then %call_sim(-2.33,-0.64,-99,-99);
 %end;
 %else %if &rating_scale_type=6 %then %do;
          %if &score_dist=1 %then %call_sim(-0.97,-0.43,0,0.43);
          %if &score_dist=2 %then %call_sim(-2,-1,0,1);
          %if &score_dist=3 %then %call_sim(-1.88,-1.08,-0.13,0.84);
 %end;
%mend sim;


*****************************************************************;
* GENERATE EFFECT AND NON-EFFECT RATERS     WITH MISSING DATA
*****************************************************************;
%macro
gen_data(select_rate=,seed=,correl=0,strength=,nrater=,nperson=,maxscore=,normmu=0,normst
d=1);
  data _null_;
    file 'data_generate_cmd.cqc';
    put
    "set warnings=no, seed=&seed.;" /
    "generate !nitems=&nrater., npersons=&nperson., maxscore=&maxscore., itemdist=item.txt,"
/
    "abilitydist=NORMAL(&normmu.:&normstd.) >> score.txt,item_gen.txt,ability.txt;" /
    "quit;";
```

```
run;

x console data_generate_cmd.cqc;

*creat another response matrix with the last rater to be the effect rater;
*the last rater is manipulated to be the effect one by adding a parameter;
*to the existing ability distribution;
data _null_;
  infile 'ability.txt';
  input rater ability;

  if &study.=1 then do;
    ability=ability*&strength;
    end;
  else do;
    do q=1 to &nperson.;
          error=rannor(&seed.);
          output;
          end;
        ability=&correl.*ability+((1-&correl.**2)**.5)*error;
        end;

        file 'ability_1.txt';
        put rater ability;
run;

*Add centrality effect by generating a new dataset(score_#_1.txt) with the;
*same item distribution and the manipulated ability distribution;
data _null_;
        file "data_generate_cmd.cqc";
        put
        "set warnings=no, seed=&seed.;" /
         "generate !nitems=&nrater., npersons=&nperson., maxscore=&maxscore.,
itemdist=item.txt," /
        "abilitydist=ability_1.txt >> score_1.txt;" /
        "quit;";
run;

x console data_generate_cmd.cqc;

data score;
  infile 'score.txt';
        input @1 (rater1-rater&nrater.) (1.);
run;
```

```
*Pick the last item in the manipulated response matrix(score_#_1.txt) and ;
*insert(or replace the last rater with this one) in the original response;
*matrix(score_#.txt);
%let nrater_total = %EVAL(&nrater.+1); *for insertion only. Delete this line if for replacement;
data score_1;
        infile 'score_1.txt';
  input rater&nrater_total. &nrater.;
run;

*convert the wide response file to a long file format;
data temp0(keep=serial item value1 second);
  merge score score_1;
        second=2-rantbl(2*&seed.,&select_rate.,1-&select_rate.);*0=non-selected,1=selected;
        array position{*}u rater1-rater&nrater_total.;
        do sub=1 to &nrater_total.;
          serial=_n_;
          item=sub;
          value1=position{sub};
          output;
        end;
run;

data temp0;
  set temp0;
        random=ranuni(&seed.);
        output;
run;

proc sort data=temp0;
  by serial DESCENDING random ;
run;

*Convert the long response file back to a wide file format;
data temp1;
        set temp0;
        value2=value1;
        by serial;
        if first.serial then do;
          count=0;
          retain count;
        end;
        count+1;
        if count > &choose. then value1=.;
        if count > &choose. then value2=.;
        if count=&choose. and second=0 then value2=.;
```

```
        output;
 run;

 proc sort data=temp1;
        by serial item;
 run;

 data _null_;
        set temp1;
        by serial;
        retain rater1-rater&nrater_total.;
        array position{*} rater1-rater&nrater_total.;
        if first.serial then
          do sub=1 to &nrater_total.;
                position{sub}=.; *initializing to missing;
          end;
          position(item)=value2;*looping across values in the variable item;
          if last.serial then do;   *outputs only the last obs in a subject;
          file 'response.txt';
          put @1 (rater1-rater&nrater_total.)(1.);
        end;
 run;
 %end;

 quit;
%mend gen_data;


*****************************************************************;
* WINSTEPS ESTIMATION
*****************************************************************;
%macro proc_data(nrater_total);
 *Create a Winsteps control file to estimate theta;
 %if &rating_scale_type = 6 %then %do;
   data _null_;
    file 'est_cmd.txt';
    put
   '&inst' / 'item1=1' /
   'name1=1' / 'inumb=y' / 'codes=012345' /
   'hlines=n' / '&end';
   run;
 %end;
 %else %do;
   data _null_;
    file 'est_cmd.txt';
```

```
    put
    '&inst' / 'item1=1' /
    'name1=1' / 'inumb=y' / 'codes=0123' /
    'hlines=n' / '&end';
   run;
 %end;

 *Call Winsteps to estimate residual, item, and person;
 x c:\winsteps\winsteps est_cmd.txt wincmd.out batch=yes data=response.txt ni=&nrater_total.
xfile=residual.txt pfile=person_est.txt ifile=item_est.txt;

 quit;
%mend proc_data;

*************************************************************************;
* RESULT SUMMARY
*************************************************************************;
%macro analysis(iter=,select_rate=,value=,nrater_total=,nperson=);
 %if &study=1 %then %do;
  %let value_lable=strength;
 %end;
 %else %do;
  %let value_lable=correl;
 %end;

 *************************************************************************;
 * Get fit statistics, observed, expected ability and
 * expected scores from Winsteps output file
 *************************************************************************;
  data Residual;
   infile 'residual.txt';
   input Person Item OBS ORD EXPECT VARIAN ZSCORE RESIDL;
   drop ORD VARIAN ZSCORE ;
   if OBS=-1 then OBS=.;
   if (OBS^=EXPECT) & (RESIDL=0) then RESIDL=.;
  run;

  *Restructure WINSTEPS output files;
  *Transpose the long file to a wide file;
  proc transpose data=Residual out=wide1 prefix=Exprat;
   by PERSON;
   id ITEM;
   var EXPECT;
  proc transpose data=Residual out=wide2 prefix=Obs;
   by PERSON;
```

```
 id ITEM;
 var OBS;
proc transpose data=Residual out=wide3 prefix=Res;
 by PERSON;
 id ITEM;
 var RESIDL;

data ability;
 infile 'person_est.txt';
 input person measure;
 drop person;

data residual_wide;
 merge wide1(drop=_name_) wide2(drop=_name_) wide3(drop=_name_) ability;
run;

*****************************************************************;
* Calculate exp_rating-obs and exp_rating-res correlations
*****************************************************************;
%do r=1 %to &nrater_total.;

 proc corr data=Residual_wide nosimple noprint outp=correlations;
  var Obs&r. Res&r.;
  with Exprat&r. ;
 run;

  *The following tedious code is to collect the necessary pieces and put them into a desirable
format;
  data correlations_w;
   set correlations;
   if _TYPE_^="CORR" then delete;
   drop _NAME_;
  run;

  proc transpose data=correlations_w out=correlations_l; run;

  data correlations_l;
   set correlations_l(keep=COL1);
   if _n_=1 then type="exprat_obs";
   if _n_=2 then type="exprat_res";
  run;

  proc transpose data=correlations_l out=correlations_l;id type; run;

  data correlations_l;
```

```
 set correlations_l(keep = exprat_obs exprat_res);
 iteration = &iter.;
 rater = &r.;
       select_rate = &select_rate.;
 score_dist = &score_dist.;
 &value_lable. = &value.;
 rater_n = &nrater_total.;
 ratee_n = &nperson.;
run;
 *The above tedious code is to collect the necessary pieces and put them into a desirable format;

 proc append base=corr_rpt1 data=correlations_l;run;
%end;

 ****************************************************************;
 * Calculate measure-residual correlations
 ****************************************************************;
 %do r=1 %to &nrater_total.;
 proc corr data=Residual_wide nosimple noprint outp=correlations;
  var Res&r.;
  with measure;
 run;

  *The following tedious code is to collect the necessary pieces and put them into a desirable
format;
  data correlations_w;
   set correlations;
   if _TYPE_^='CORR' then delete;
   drop _NAME_;
  run;

  proc transpose data=correlations_w out=correlations_l; run;

  data correlations_l;
   set correlations_l(keep=COL1);
   if _n_=1 then type='measure_resid';
  run;

  proc transpose data=correlations_l out=correlations_l;id type; run;

  data correlations_l;
   set correlations_l(keep=measure_resid);
   iteration=&iter.;
   rater=&r.;
       select_rate = &select_rate.;
```

```
  score_dist = &score_dist.;
  &value_lable. = &value.;
  rater_n = &nrater_total.;
  ratee_n = &nperson.;
 run;
 *The above code is to collect the necessary pieces and put them into a desirable format;

 proc append base=corr_rpt2 data=correlations_l;run;
%end;

%sort(corr_rpt1,iteration select_rate &value_lable ratee_n rater_n rater);
%sort(corr_rpt2,iteration select_rate &value_lable ratee_n rater_n rater);

data corr_rpt;
 merge corr_rpt1 corr_rpt2;
 by iteration select_rate &value_lable. ratee_n rater_n rater;
run;

*****************************************************************;
* Calculate rating mean, std and the ratio of the normal rater
* score variance and the effect rater variance
*****************************************************************;
*Items are treated as raters;
data rater;
 infile 'item_est.txt';
 input rater MEASURE STTS COUNT SCORE ERROR INMSQ INZSTD OUTMS
OUTZSTD DISPL PTME WEIGHT OBSMA EXPMA DISCR;
 keep rater COUNT INMSQ INZSTD OUTMS OUTZSTD PTME DISCR;
run;

data response;
 infile 'response.txt';
 input @1(rater1-rater&nrater_total.) (1.);
run;

proc means MEAN STD data=response noprint;
 var rater1-rater&nrater_total.;
 output out=descriptive_w;
run;

data descriptive_w;
 set descriptive_w;
 if _STAT_ ^='MEAN' and _STAT_^='STD' then delete;
 drop _TYPE_ _FREQ_ _STAT_;
run;
```

```
data temp2(keep=rater%eval(&nrater_total.-1) rater&nrater_total.);
 set descriptive_w ;
        if _n_=1 then delete;
run;

data temp3;
 set temp2;
        variance_ratio=(rater&nrater_total./rater%eval(&nrater_total.-1))**2;
 rater_n = &nrater_total.;
 ratee_n = &nperson.;
 score_dist = &score_dist.;
        select_rate = &select_rate.;
        value = &value.;
 study = &study.;
        iteration = &iter.;
        drop rater&nrater_total. rater%eval(&nrater_total.-1);
run;

proc append base=var_ratio_rpt data=temp3;

proc transpose data=descriptive_w out=descriptive_l; run;

data descriptive_l;
 set descriptive_l(rename=(COL1=MEAN COL2=STD));
 rater = _n_;
 iteration=&iter.;
 keep rater MEAN STD iteration;
run;

data rater;
 merge descriptive_l rater;
 by rater;
 select_rate = &select_rate.;
 score_dist = &score_dist.;
 &value_lable. = &value.;
 rater_n = &nrater_total.;
 ratee_n = &nperson.;
run;

proc append base=rater_rpt data=rater;

************************************************************;
* Merge correlations and fit statistics;
************************************************************;
```

```
 %sort(rater_rpt,iteration select_rate &value_lable ratee_n rater_n rater);
 %sort(corr_rpt,iteration select_rate &value_lable ratee_n rater_n rater);

 data final_rpt;
  merge corr_rpt rater_rpt;
  by iteration select_rate &value_lable. ratee_n rater_n rater;
 run;

 quit;
%mend analysis;

*****************************************************************;
* ERROR CHECK;
*****************************************************************;
%macro error_check(nrater_total);
*Check correlations of orininal and estimated parameters for theta;
  data theta;
   infile 'ability.txt';
   input person theta;

  data theta_est;
   infile 'person_est.txt';
   input person theta_est;
  run;

  data theta_all;
   merge theta theta_est;
   by person;
  run;

  proc corr data=theta_all nosimple;
   var theta theta_est;
  run;

  *Check the % of threashholds for last two raters;
  %let pos = %SYSEVALF(&nrater_total.-1);
  data last2raters;
   infile 'response.txt';
   input @&pos. last_normal_rater 1. @&nrater_total. effect_rater 1.;
  run;

  proc freq data=last2raters;
   tables last_normal_rater effect_rater;
  run;
```

```
 quit;
%mend error_check;

%macro call_sim(t1,t2,t3,t4);
 %if &rating_scale_type.=4 %then %let maxscore=3;
 %else %if &rating_scale_type.=6 %then %let maxscore=5;

 %if &study.=1 %then %do;

   %do n=1 %to &select_rate_size.;
          %let select_rate = &&select_rate&n.;

          %do p=1 %to &strength_size.;
          %let strength = &&strength&p.;

      %do j=1 %to &nrater_size.;
           %let nrater = &&nrater&j.;
           %global nrater_total;
           %let nrater_total = %EVAL(&nrater.+1);
           %thresholds(&t1,&t2,&t3,&t4,&nrater.,%eval(100+&j.))

           %do k=1 %to &nperson_size.;
           %let nperson = &&nperson&k.;

            *correl is set to 0 for study 1;
            %do m=1 %to &iteration.;

            %gen_data(select_rate=&select_rate.,seed=%eval(1000000*&n.+100000*&p.+10000
*&j.+1000*&k.+100*&m.),correl=0,strength=&strength.,nrater=&nrater.,nperson=&nperson.,m
axscore=&maxscore.)
            %proc_data(&nrater_total.)

             %analysis(iter=&m.,select_rate=&select_rate.,value=&strength.,nrater_total=&nrater
_total.,nperson=&nperson.)
             *%error_check(&nrater_total.);
          %end;

           %let rpt_name=study1_rpt_;
           %let
myfile1=&rpt_name.&score_dist.&delimiter.&select_rate.&delimiter.&strength.&delimiter.&nr
ater.&delimiter.&nperson.&ext.;
          proc export data=final_rpt outfile="&myfile1." dbms=csv replace; putnames=yes; run;
          proc datasets; delete rater_rpt corr_rpt1 corr_rpt2 corr_rpt; run;
           %end;
          %end;
```

```
        %end;
  %end;

  %let myfile2=study1_ratio_rpt&ext.;
  proc export data=var_ratio_rpt outfile="&myfile2." dbms=csv replace; putnames=yes; run;
  proc datasets; delete var_ratio_rpt; run;
 %end;
 %else %do;
  %do n=1 %to &select_rate_size.;
        %let select_rate = &&select_rate&n.;

    %do p=1 %to &correl_size.;
         %let correl = &&correl&p.;

     %do j=1 %to &nrater_size.;
         %let nrater = &&nrater&j.;
         %global nrater_total;
         %let nrater_total = %EVAL(&nrater.+1);
         %thresholds(&t1,&t2,&t3,&t4,&nrater.,%eval(100+&j.))

         %do k=1 %to &nperson_size.;
          %let nperson = &&nperson&k.;

          *strength is set to 0 for study 2;
          *Missing data design A;
          %do m=1 %to &iteration.;

        %gen_data(select_rate=&select_rate.,seed=%eval(1000000*&n.+100000*&p.+10000*
&j.+1000*&k.+100*&m.),correl=&correl.,strength=0,nrater=&nrater.,nperson=&nperson.,maxs
core=&maxscore.)
        %proc_data(&nrater_total.)

        %analysis(iter=&m.,select_rate=&select_rate.,value=&correl.,nrater_total=&nrater_total
.,nperson=&nperson.)
         *%error_check(&nrater_total.) ;
        %end;

        %let rpt_name1=study2_rpt_;
        %let
myfile1=&rpt_name1.&score_dist.&delimiter.&select_rate.&delimiter.&correl.&delimiter.&nrat
er.&delimiter.&nperson.&ext.;
        proc export data=final_rpt outfile="&myfile1." dbms=csv replace; putnames=yes; run;
        proc datasets; delete rater_rpt corr_rpt1 corr_rpt2 corr_rpt; run;
        %end;
         %end;
```

```
 %end;
%end;

%let myfile2=study2_ratio_rpt&ext.;
proc export data=var_ratio_rpt outfile="&myfile2." dbms=csv replace; putnames=yes; run;
proc datasets; delete var_ratio_rpt; run;
 %end;

 *x 'del "*.txt"';
 quit;
%mend call_sim;

%sim()
```

Part II: SAS macro for "threshold.sas"

```
/*CREATES AN ITEM DIFFICULTY FILE TO READ INTO CONQUEST*/
%macro thresholds(t1,t2,t3,t4,items,item_seed);
 data item;
  threshold=0;
  do item=1 to &items.;
   *difficulty=(2*ranuni(&item_seed.))-1;
   difficulty=0;
   output;
  end;
 run;


 %if &t3 = -99 and &t4 = -99 %then %do;
  data threshold;
   t1=&t1.;
   t2=&t2.;
   do item=1 to &items.;
    output;
   end;
  run;

  proc transpose data=threshold out=threshold;
   by item;
   var t1-t2;
  run;
 %end;
 %else %do;
  data threshold;
```

```
 t1=&t1.;
 t2=&t2.;
 t3=&t3.;
 t4=&t4.;
 do item=1 to &items.;
  output;
 end;
run;

proc transpose data=threshold out=threshold;
 by item;
 var t1-t4;
run;
%end;

data threshold;
 set threshold;
 if _NAME_='t1' then threshold=1;
 if _NAME_='t2' then threshold=2;
 %if &t3 ^= -99 and &t4 ^= -99 %then %do;
 if _NAME_='t3' then threshold=3;
 if _NAME_='t4' then threshold=4;
 %end;
 difficulty=col1;
 drop col1 _NAME_;
run;

proc append base=item data=threshold;
run;

data parameters;
 file 'item.txt';
 set item;
 put item threshold difficulty;
run;

%mend thresholds;
```

Part III: SAS macro for "array.sas"

```
%macro array (

   array
 , items
```

```
, sep = %str( )
, scope = RESOLVE
, lets =
, locals =

);

%* Richard A. DeVenezia
%* 031119 - Revised as [array], added scope, lets and locals
%* 931109 - Initial coding as [makelist]
%*
%* create a macro array from a delimited list of items
%*
%* value is
%* ------------------------------------------------------------
%* array  - macro array name, prefix of numbered macro variables that
%*          will contain the item values
%* items  - original list of items, separated by sep
%* sep    - separator between items of items (incoming)
%* scope  - scope of macro array variables.
%*          GLOBAL, INHERIT, RESOLVE
%*          GLOBAL - avoid if possible
%*          INHERIT - the invoker _must_ ensure the variables
%*          <&array>_size <&array>_1 ... <&array>_n exist
%*          prior to invoking %array.
%*          Why? because a macro is not allowed to create a macro variable
%*          in a local scope above its own.
%*         (It may however, access any macro variables in scope above itself)
%*          If this macro implicitly 'creates' a macro variable, it will
%*          be destroyed when the macro ends, and thus will not be available
%*          to invoker.
%*         RESOLVE - macro var named in lets will receive a quoted macro
%*         statement which is a series of %lets. The invoker is responsible
%*         for unquoting the statement to get macro vars in its scope.
%* lets   - name of macro var existing in invokers scope.
%*         upon return, the invoker should unquote the value
%*         to cause the macro array variables to be assigned.
%* locals - name of macro var existing in invokers scope.
%*         upon return, the invoker should resolve this variable in a
%*         %local statement to ensure the variables in the lets variable
%*         will not accidently overwrite an existing macro variable in scope
%*          higher than invoker.
%*;

%if (&array. =) %then %do;
```

```
  %put ERROR: array name is missing;
  %goto EndMacro;
%end;

%let scope = %upcase(&scope);

%if 0 = %index (|GLOBAL|INHERIT|RESOLVE|, |&SCOPE.|) %then %do;
  %put ERROR: scope = &scope is unknown;
  %goto EndMacro;
%end;

%if (&scope = RESOLVE) and (&lets = ) %then %do;
  %put ERROR: scope=&scope requires an lets=<macro-var>;
  %goto EndMacro;
%end;

%let lets = %upcase(&lets);

%if (&scope = RESOLVE) and (&lets = LETS) %then %do;
  %put ERROR: lets= can not be LETS, try lets=_let;
  %goto EndMacro;
%end;

%let locals = %upcase(&locals);
%if (&locals=LOCALS) %then %do;
  %put ERROR: locals= can not be LOCALS, try locals=_local;
  %goto EndMacro;
%end;

%local item N local;

%let N=1;

%if  (&scope = GLOBAL) %then
  %global &array._size;

%if (&scope = RESOLVE) %then
  %let &lets = ;

%if (&locals ^= ) %then
  %let &locals = &array._size;

%let item = %scan(&items,&N,%quote(&sep));

%do %while (&item ^= );
```

```
    %if &scope = GLOBAL %then
      %global &array.&N;

    %if (&scope = GLOBAL) or (&scope = INHERIT) %then
      %let &array.&N = &item;
    %else
      %let &lets = %nrquote(&&&lets)%nrstr(%let )&array.&N=&item%str(;);

    %if (&locals ^= ) %then
      %let &locals = &&&locals &array&N;

    %let N = %eval(&N + 1);
    %let item = %scan(&items,&N,%quote(&sep));

  %end;

  %let N = %eval(&N - 1);

  %if (&scope = GLOBAL) or (&scope = INHERIT) %then
    %let &array._size = &N;
  %else
    %let &lets = %nrquote(&&&lets)%nrstr(%let )&array._size=&N%str(;);

%EndMacro:

%mend;
```