# Distinguishing Several Rater Effects with the Rasch Model

National Council of Measurement in Education Annual Meeting, Chicago, IL.

Tian Song
Edward W. Wolfe

April 2015

ALWAYS LEARNING PEARSON

**Abstract**

Prior research about psychometric modeling of rater effects has focused on distinct effect in isolation even though multiple types of rater effects likely exist simultaneously in real data. This simulation study evaluates the performance of several rater effect indicators in data containing multiple rater effects. Results showed that rater location, standard deviation of rater thresholds, and correlation between the observed ratings and the estimated examinee abilities performed well in detecting rater severity, centrality, and inaccuracy respectively. All three indices produced low Type I and Type II error rates.

*Keywords:* rater effects, Rasch model, rater monitoring

# Distinguishing Several Rater Effects with the Rasch Model

A common concern about the scores assigned by human raters in educational settings is the degree to which those scores contain measurement error due to the subjectivity of raters' judgments. To address this concern, raters typically undergo a process that involves training and qualifying to score, and several rater monitoring procedures (e.g., double scoring, backreading) are employed in operational settings in order to ensure that assigned ratings as accurate and unbiased as possible. Regardless of the efforts taken to minimize inaccuracy and bias in ratings, raters may still exhibit patterns that are associated with measurement errors, which are commonly referred to as rater effects.

## Common Types of Rater Effects

Severity/Leniency is one of the most commonly studied rater effect in the literature which refers to a tendency toward systematic shifts of the average score assigned by a particular rater. Severe rater assign scores that are lower than true scores while lenient raters assign scores that are higher. Leniency/severity raises concerns for decision makers when scores are interpreted relative to a cut score, such as during college admissions or placement, determining graduation qualification, or awarding professional certifications. If leniency/severity exists in the scores, then some respondents will be incorrectly classified in decision making contexts such as these (Wolfe, 2014).

Centrality/Extremity is another common rater effect which occurs when raters tend to shift scores toward the middle categories (centrality) or tails of the rating scale (extremity). It is statistically evidenced by a decrease/increase in the standard deviation of the scores associated with a rater in the absence of an increase in the amount of random error. When centrality/extremity exists in the scores, respondents in the tails of the distribution may be

misclassified and/or decision makers may believe that examinees are less or more homogeneous than is actually the case.

Inaccuracy/accuracy is a third common rater effect which refers to trends toward increased or decreased randomness in the scores assigned by a particular rater. Inaccurate raters assign scores that deviate from true scores in a random, unpredictable manner. In a statistical analysis, we would expect there to be a very low percentage of perfect agreement for inaccurate raters, and a near-zero correlation between their scores and true scores. On the other hand, we would expect a high percentage of perfect agreement for accurate raters, and a near-one correlation between their scores and true scores. The important distinction between inaccuracy and rater leniency/severity or centrality/extremity is the level of randomness in the data. While the previous two rater effects produce a rank ordering of responses that is similar to that based on true scores, inaccuracy introduces undue randomness into that rank ordering. As a result, decision makers might attribute observed score differences to individual differences when they are, indeed, random errors.

**Detecting Rater Effects**

Latent trait modeling methods are often employed to detect rater effects. In the literature, there is considerable research that explores the application of Rasch and other latent trait models to the detection of rater effects, and several indices have been identified to detect those effects (Decarlo, Kim, & Johnson, 2011; Lunz & Stahl, 1993; Myford & Wolfe, 2003, 2004, 2009; Wolfe, 2004, 2005; Wolfe & Song, 2015). In this paper, we focus on one of the most commonly employed models, the Rasch partial credit model (Masters, 1982). In the context of monitoring rater effects, the parameter indexing and parameter estimate interpretation in this model needs to be slighted modified. Let $\pi_{nrk}$ denote the probability that rater *r* assigns a score in rating category

$k$ to examinee $n$ , and let $\pi_{nrk-1}$ denote the probability that rater $r$ assigns a score in rating category $k$-*1* to examinee $n$. The log-ratio of these events takes the following form

$$\log\left(\frac{\pi_{nrk}}{\pi_{nrk-1}}\right) = \theta_n - \lambda_r - \tau_{rk} \qquad (1)$$

Equivalently, the probability $\pi_{nrk}$ in the RPCM can be written as

$$\pi_{nrk} = \frac{\exp\left[\sum_{j=0}^{k}(\theta_n - \lambda_r - \tau_{rk})\right]}{\sum_{k=0}^{m}\exp\left[\sum_{j=0}^{k}(\theta_n - \lambda_r - \tau_{rk})\right]} \qquad (2)$$

In this model, $\theta_n$ is the examinee ability, $\lambda_r$ is the rater location, which is interpreted as the severity/leniency of a particular rater, and $\tau_{rk}$ is the relative difficulty of the two adjacent rating categories.

This model contains a parameter ( $\lambda_r$ ) that explicitly measures severity/leniency. Positive values of the $\lambda_r$ estimate indicate rater severity, and negative values indicate leniency. Rater centrality and extremity can be detected as the increase or decrease, respectively, in the standard deviation of raters' estimated category thresholds [SD( $\tau_{rk}$ )]. Central raters are expected to have a larger value of this index (i.e., a wider dispersion of thresholds), resulting in a concentration of expected rating in middle rating categories. Detection of rater inaccuracy is also straightforward. We use the correlation between the observed ratings and the estimated examinee abilities [ $R(X,\theta)$ ] as an indicator of inaccuracy. We would expect the correlation closer to zero for inaccurate raters and closer to one for accurate raters.

Previous research has shown that those indices are effective in detecting rater effects (Wolfe, 2004, 2005, 2014; Wolfe and Song, 2015), but they have only been studied in simulated data that contain a single type of rater effect in a particular data set. In operational settings, it is likely that multiple rater effects exist in a pool of raters simultaneously, so that prior research is somewhat limited in its applicability to operational testing scenarios. It may be that the simultaneous existence of multiple rater effects in a data set influences the accuracy with which the rater effect indicators are able to detect the existence of the intended rater effect. The purpose of this paper is to demonstrate the usefulness of these indices for differentiating several types of rater effects simultaneously in data that are simulated to realistically represent the existence of rater effects.

## Method

### Simulation Design

We simulated scores from a single-prompt/single-holistic-score data collection design—a design that is commonly employed for monitoring raters in an operational scoring project. In this design, raters assign a single holistic score to each student response to a single prompt. In our simulation, we generated fully-crossed data in which all raters assigned scores to all examinees. Specifically, we simulated data for 2,000 examinees, each rated by 100 raters with ratings on a scale ranging from 0 to 4. Our simulation focused on four design variables (See Table 1):

1. Rater effect groups. There are four groups: normal raters, severe raters, central raters, and inaccurate raters.
2. Quality of rater pool. We defined this as the correlation between normal raters' raw scores and true scores. Two levels of raw score-true score correlations are considered: .90 and .75, which indicates high or low quality of rater pool.

3. Magnitude of rater effects. Two levels of effect strength are considered. For severe raters, the ratings are altered so that the data exhibited a tendency toward negative deviations from the true score by about one half or one score point. For central raters, the dispersion of ratings is shrunk to be about 75% or 50% of that of the true scores. For inaccurate raters, a large amount of random error was added so that the correlation between raw scores and true scores for those raters to be .65 or .50.
4. Percentage of raters exhibiting effects. 6% or 18% of the raters are simulated as raters with effects (evenly split among three types of rater effects), representing small or large rater effect prevalence. In our case, all conditions have 100 raters, so there are 6 or 18 raters that exhibit rater effects.

Table 1

*Simulation Design Variables*

| **Variable** | **Measured as** | **Levels** |
|---|---|---|
| Rater Group | Rater effect designation | [Normal, Severe, Central, Inaccurate] |
| Quality of Rater Pool | Correlation between normal raters' raw scores and true scores | [.9, .75] |
| Magnitude of Rater Effects | Severity: Raw score points below mean of normal raters | [.5 , 1] |
| | Centrality: $SD_{central}$ / $SD_{normal}$ | [.75, .5] |
| | Inaccuracy: Correlation between inaccurate raters' raw scores and true scores | [.65, .5] |
| Prevalence of Rater Effects | Percentage of raters exhibiting effect | [6%, 18%] |

Our data generation processes resulted in an experimental design containing 2 (quality of rater pool) × 2 (magnitude of rater effects) × 2 (percentage of raters with effects) = 8 cells. We replicated each of these cells 100 times.

**Analysis**

For each simulated data file, we estimated parameters for the RPCM via Winsteps (Linacre, 2012). Using the estimated parameters, we computed the rater effect indices for each rater. Specifically, we estimated rater location parameters as indicators of rater severity/leniency, computed the standard deviation of the rater thresholds as an indicator of rater centrality/extremity, and computed the correlation between observed ratings and ability estimates as an indicator of rater accuracy/inaccuracy.

For each index, we also identified a critical value so that each rater would be flagged as normal rater, severe rater, central rater, or inaccurate rater. Specifically, for each simulated data file, we computed the $5^{th}$ or $95^{th}$ percentile of each index based on normal raters' index values, assuming an overall Type I error rate of .05. Depending on the direction of the difference between normal raters and problem raters, we used the $5^{th}$ or $95^{th}$ percentile as the critical value for flagging. We then averaged these values across iterations. Finally, the mean critical values were applied to indices for each rater within each simulated data file, and each rater was categorized into one of the four rater groups or multiple groups. To evaluate the effectiveness of each index in detecting rater effects, we computed Type I error rate (i.e., a normal raters is erroneously flagged as exhibiting a rater effect) and Type II error rate (i.e., an effect rater is not identified).

## Results

Descriptive statistics for the simulated ratings are shown in Table 2. Within each simulated data file, we calculated mean, standard deviation, and the correlation between the simulated ratings and true scores, and then averaged these values across iterations within a condition. These statistics indicate that our simulation process produced scores that are consistent with our expectations concerning the various rater effects. The mean of the simulated ratings is about equal to the mean of true score ratings for all rater groups but the severe raters who have a mean about one-half or one score point lower, depending on the magnitude of the simulated effect. The standard deviations of ratings for simulated raters are also about equal to those of the true score ratings. Central raters, the lone exception, have a dispersion of ratings to be about 75% or 50% of that of the true scores. Correlation between simulated ratings and true ratings are about .90 or .75, again depending on the simulation condition, but inaccurate raters exhibit significantly lower correlations of about .65 or .47. Overall, these statistics demonstrate that the simulation process produced three rater effects that are easily distinguishable in the raw scores.

Table 3 summarizes the average values of the rater effect indices. We first calculated the mean of each index within each simulated data file, and then averaged these values across iterations within a condition. Overall, these statistics indicate that the rater effect indices are good indicators of their corresponding rater effects.

First, the values of rater location estimates, $\lambda_r$ , are high for severe raters while the values are close to zero for other raters. Trends in the values of these parameter estimates are depicted in Figure 1. For severe raters, the values increase as the rater quality increases, especially when the severity strength is large. Similarly, the values become greater as the severity strength

becomes stronger. For example, in a rater pool with high quality and 6% of raters simulated as severe raters, the mean values of $\lambda_r$ for those severe raters increase from 2.37 to 4.45 when the magnitude of effect increases. In addition, $\lambda_r$ seems to vary little across the prevalence levels, but it tends to decrease when the severity strength is large and the rater quality is high. On the other hand, the values of $\lambda_r$ for normal, central or inaccurate raters do not seem to vary by level of rater quality, severity strength and prevalence levels. Since central and inaccurate raters have similar patterns as normal raters, only normal raters are reported in Figure 1.

Table 2

*Descriptive Statistics of Simulated Ratings*

| Prevalence | Effect Strength | Rater Pool Quality | Index | True | Normal | Severe | Central | Inaccurate |
|---|---|---|---|---|---|---|---|---|
| 6% | small | high | Mean | 2.00 | 2.01 | **1.52** | 2.01 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.09 | **0.83** | 1.11 |
| | | | R(T,X) | | 0.89 | 0.87 | 0.85 | **0.65** |
| | | low | Mean | 2.00 | 2.01 | **1.60** | 2.01 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.09 | **0.82** | 1.11 |
| | | | R(T,X) | | 0.74 | 0.74 | 0.72 | **0.65** |
| | large | high | Mean | 2.00 | 2.01 | **1.06** | 2.00 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.01 | **0.59** | 1.10 |
| | | | R(T,X) | | 0.89 | 0.84 | 0.80 | **0.47** |
| | | low | Mean | 2.00 | 2.01 | **1.20** | 2.00 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.03 | **0.58** | 1.10 |
| | | | R(T,X) | | 0.75 | 0.72 | 0.66 | **0.46** |
| 18% | small | high | Mean | 2.00 | 2.01 | **1.52** | 2.01 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.09 | **0.83** | 1.11 |
| | | | R(T,X) | | 0.89 | 0.87 | 0.85 | **0.65** |
| | | low | Mean | 2.00 | 2.01 | **1.60** | 2.01 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.09 | **0.82** | 1.11 |
| | | | R(T,X) | | 0.74 | 0.74 | 0.72 | **0.65** |
| | large | high | Mean | 2.00 | 2.01 | **1.07** | 2.00 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.01 | **0.59** | 1.10 |
| | | | R(T,X) | | 0.89 | 0.84 | 0.79 | **0.47** |
| | | low | Mean | 2.00 | 2.01 | **1.20** | 2.01 | 2.01 |
| | | | SD | 1.12 | 1.11 | 1.04 | **0.59** | 1.10 |
| | | | R(T,X) | | 0.75 | 0.73 | 0.66 | **0.47** |

Table 3

*Descriptive Statistics of Indices*

| Prevalence | Effect Strength | Rater Pool Quality | Index | Normal | Severe | Central | Inaccurate |
|---|---|---|---|---|---|---|---|
| 6% | small | high | $\lambda_r$ | -0.05 | **2.37** | -0.06 | -0.04 |
| | | | $SD(\tau_{rk})$ | 5.68 | 5.70 | **8.27** | 5.73 |
| | | | $R(X,\theta)$ | 0.91 | 0.91 | 0.89 | **0.68** |
| | | low | $\lambda_r$ | -0.02 | **0.83** | -0.01 | -0.02 |
| | | | $SD(\tau_{rk})$ | 2.24 | 2.25 | **3.54** | 2.25 |
| | | | $R(X,\theta)$ | 0.77 | 0.76 | 0.75 | **0.68** |
| | large | high | $\lambda_r$ | -0.09 | **4.45** | -0.08 | -0.09 |
| | | | $SD(\tau_{rk})$ | 5.42 | 5.34 | **10.68** | 5.48 |
| | | | $R(X,\theta)$ | 0.91 | 0.88 | 0.83 | **0.50** |
| | | low | $\lambda_r$ | -0.03 | **1.66** | -0.03 | -0.03 |
| | | | $SD(\tau_{rk})$ | 2.21 | 2.24 | **5.39** | 2.24 |
| | | | $R(X,\theta)$ | 0.77 | 0.75 | 0.70 | **0.49** |
| 18% | small | high | $\lambda_r$ | -0.13 | **2.04** | -0.13 | -0.12 |
| | | | $SD(\tau_{rk})$ | 5.09 | 5.11 | **7.42** | 5.12 |
| | | | $R(X,\theta)$ | 0.91 | 0.91 | 0.89 | **0.68** |
| | | low | $\lambda_r$ | -0.05 | **0.79** | -0.05 | -0.05 |
| | | | $SD(\tau_{rk})$ | 2.22 | 2.22 | **3.50** | 2.23 |
| | | | $R(X,\theta)$ | 0.77 | 0.77 | 0.75 | **0.68** |
| | large | high | $\lambda_r$ | -0.23 | **3.56** | -0.20 | -0.22 |
| | | | $SD(\tau_{rk})$ | 4.53 | 4.42 | **8.91** | 4.59 |
| | | | $R(X,\theta)$ | 0.91 | 0.88 | 0.83 | **0.50** |
| | | low | $\lambda_r$ | -0.1 | **1.53** | -0.11 | -0.09 |
| | | | $SD(\tau_{rk})$ | 2.12 | 2.14 | **5.15** | 2.15 |
| | | | $R(X,\theta)$ | 0.77 | 0.75 | 0.70 | **0.49** |

Second, there are significant differences in the standard deviation of rater thresholds index, $SD(\tau_{rk})$, between central and other raters. Trends in the values of these statistics are depicted in Figure 2. For central raters, the mean values of $SD(\tau_{rk})$ are consistently higher than those for other raters. In addition, these values increase as the rater quality becomes higher and the centrality effect becomes stronger. Again, $SD(\tau_{rk})$ do not seem to be sensitive to the prevalence levels, but its values drop when the severity strength is large and the rater quality is

high. For other raters, $SD(\tau_{rk})$ only increases as the rater quality increases while its values vary little across magnitude of centrality effect and prevalence levels.

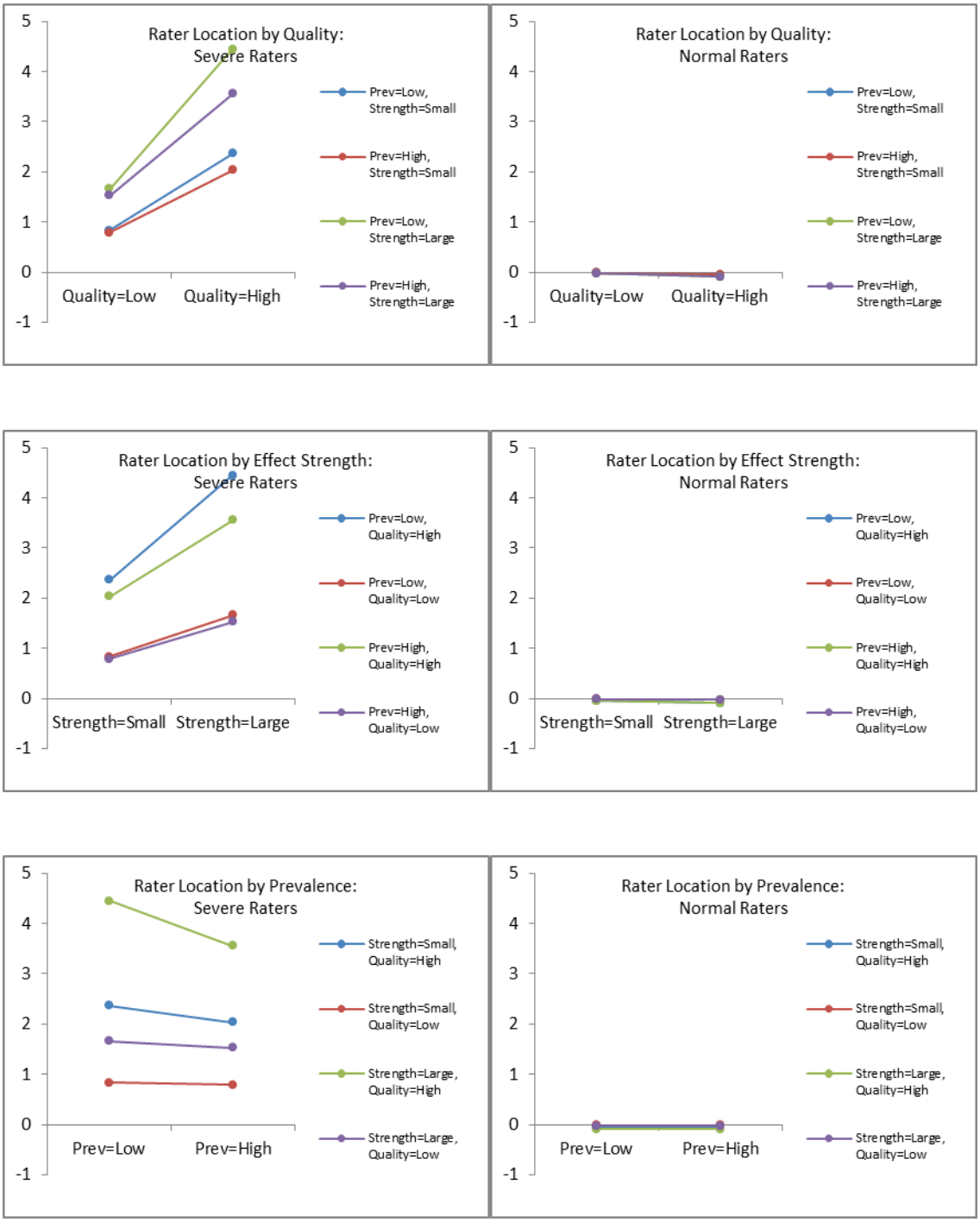


Figure 1. Severity Index by Quality of Rater Pool, Effect Strength or Severity Prevalence

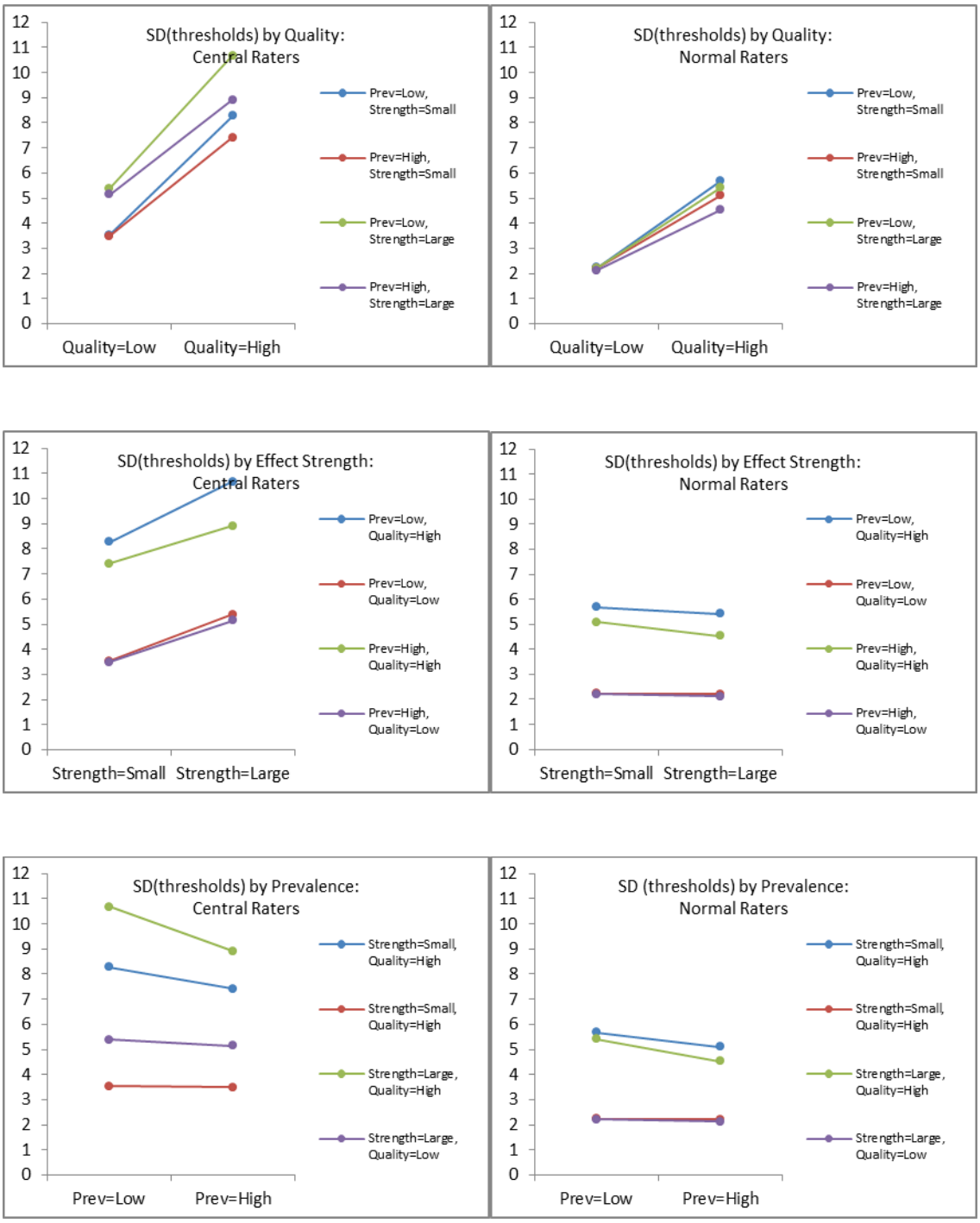


Figure 2. Centrality Index by Quality of Rater Pool, Effect Strength or Centrality Prevalence

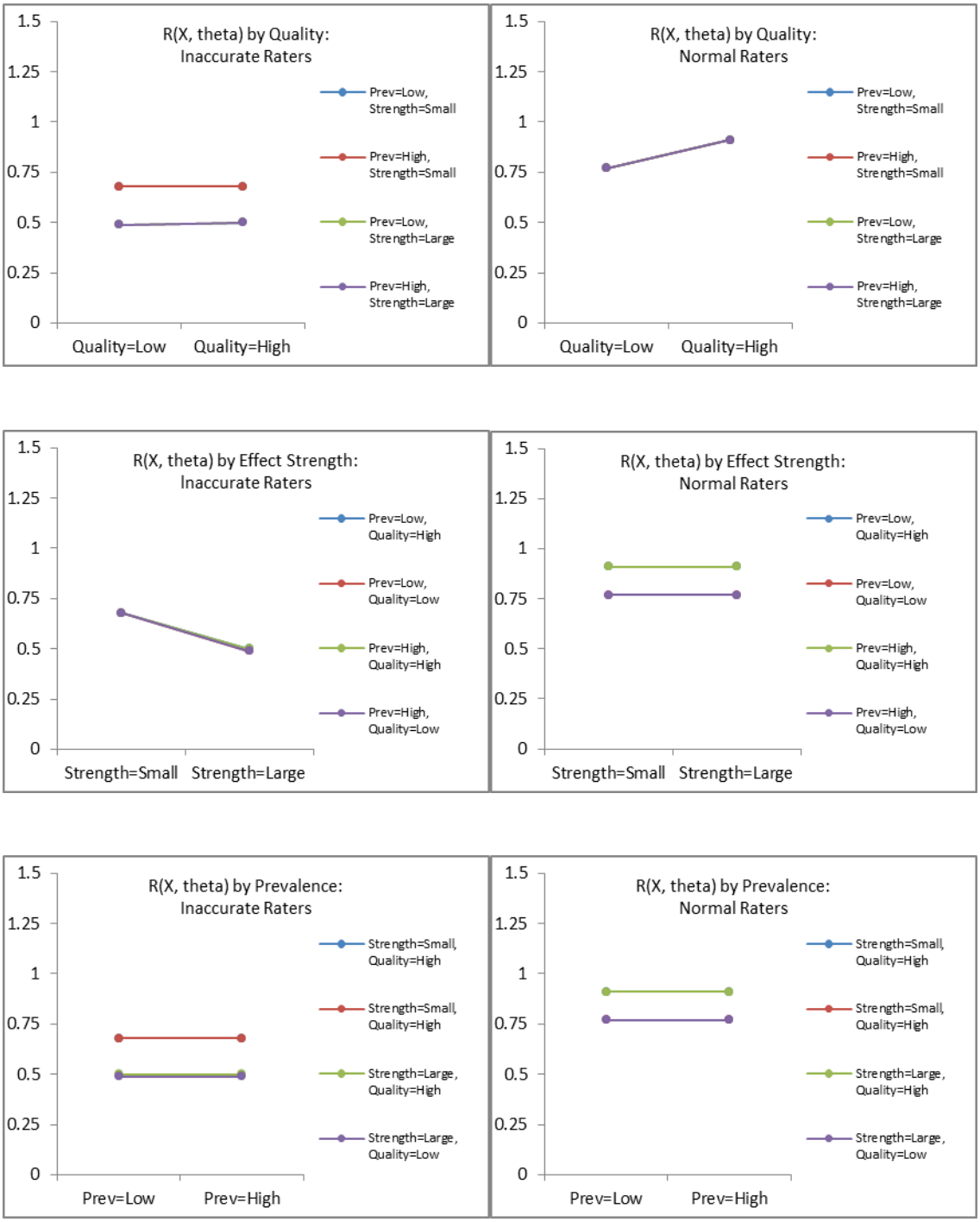


Figure 3. Inaccuracy Index by Quality of Rater Pool, Effect Strength or Inaccuracy Prevalence

Third, rater inaccuracy manifests itself as attenuated values of $R(X,\theta)$, and the differences in this index between inaccurate raters and other raters become larger when the rater quality increases. These trends are depicted in Figure 3. In addition, this index is only sensitive to the strength of the inaccuracy effect. For example, it decreases from .68 to .50 as the effect becomes stronger. It is also worth noting that the values of $R(X,\theta)$ for central raters seem to be slightly smaller compared to those for normal and severe raters. However, there are still significant differences between central and inaccurate raters so that inaccurate raters can be easily detected using this index.

Table 4 provides the Type I and Type II error rates for each index. Overall, all three indices produced low Type I and Type II error rates. First, $\lambda_r$, $SD(\tau_{rk})$ and $R(X,\theta)$ performed very well, resulting in Type I error rates around .05. Specifically, only 4% to 7% of the time a normal

Table 4

*Type I and Type II Error Rates*

| Prevalence | Effect Strength | Rater Pool Quality | $\lambda_r$ | | $SD(\tau_{rk})$ | | $R(X,\theta)$ | |
|---|---|---|---|---|---|---|---|---|
| | | | Type I | Type II | Type I | Type II | Type I | Type II |
| 6% | small | High | 0.04 | 0.00 | 0.05 | 0.00 | 0.05 | 0.00 |
| | | Low | 0.05 | 0.00 | 0.05 | 0.00 | 0.05 | 0.00 |
| | large | High | 0.04 | 0.00 | 0.07 | 0.00 | 0.06 | 0.00 |
| | | Low | 0.06 | 0.00 | 0.05 | 0.00 | 0.05 | 0.00 |
| 18% | small | High | 0.06 | 0.00 | 0.06 | 0.00 | 0.06 | 0.00 |
| | | Low | 0.07 | 0.00 | 0.05 | 0.00 | 0.05 | 0.00 |
| | large | High | 0.05 | 0.00 | 0.07 | 0.00 | 0.05 | 0.00 |
| | | Low | 0.04 | 0.00 | 0.05 | 0.00 | 0.05 | 0.00 |

rater would be erroneously identified as a severe, central or inaccurate rater. Second, they also produced high statistical power (1-Type II error rate) in detecting problem raters. They correctly flagged severe, central or inaccurate raters 100% of the time. It indicates all three indices are effective to detect rater effects.

## Discussion

Our results indicate that rater severity, centrality, and inaccuracy indices used in this study perform effectively in detecting rater effects in data containing multiple rater effects. Specifically, each index exhibits good detection of its own rater effect, with low Type I and Type II error rates, without being confronted by different types of rater effects. For example, when multiple rater effects exist in the data, rater location parameter ( $\lambda_r$ ) is only sensitive to severity, not to centrality and inaccuracy. Standard deviation of rater thresholds ( $SD(\tau_{rk})$ ) is inflated for central raters, not for severe and inaccurate raters. Similarly, inaccurate raters have attenuated values of the correlation between observed ratings and ability estimates ( $R(X,\theta)$ ), but there are no large differences between severe, central and normal raters using this index.

It is also worth noting the limitations of our simulation study. First, we only evaluated three indices under the Rasch partial credit model. Some of the indices may not be useful in other latent trait models. For example, the centrality index used here would not be applicable under the Rasch rating scale model in which rater thresholds are constrained to be constant across raters. Other indices (e.g., the expected score-residual correlation as an indicator of centrality) can be considered in the future studies. Second, we chose the simplest data design, a single-prompt/single-holistic-score design, and we did not consider designs that include examinees who responded to multiple items or multiple scores being assigned to a particular response. Third, no

missing data was allowed in the current simulation. As a result, our results may not generalize to rating contexts in which randomly selected subsets of raters assign scores to each examinee.

We should also note that the method of calculating critical values for flagging raters in our study may be difficult to implement in the operational scoring project. In our simulation, we chose the distribution of each index from normal raters as a null distribution, and used the $5^{th}$ or $95^{th}$ percentile in this distribution as a cut point. However, in the operational scoring project, we do not know which raters are normal raters, and hence additional sampling errors would be introduced in calculating the critical values. In the future, bootstrap or asymptotic distribution of the transformed indices should be explored for this purpose.

## References

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

Linacre, J. M. (2012). WINSTEPS. Beaverton, OR: Winsteps.com.

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

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

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

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

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. (2014). *Methods for monitoring rating quality: Current practices and suggestted changes*. (White paper) Iowa City, IA: Pearson.

Wolfe, E.W., Jiao, H., & Song, T. (in press). A family of rater accuracy models. *Journal of Applied Measurement*

Wolfe, E.W., & Song, T. (in press). Comparison of models and indices for detecting rater centrality. *Journal of Applied Measurement*.