Application of Latent Trait Models to Identifying Substantively Interesting Raters

American Educational Research Association
New Orleans, LA

Edward W. Wolfe
Aaron McVay

April 2011
Author Note


Correspondence concerning this article should be addressed to Edward W. Wolfe, 2510 N. Dodge St., Mailstop 125, Iowa City, IA 52245-9945. E-mail: ed.wolfe@pearson.com.
Abstract

Historically, research focusing on antecedents of good ratings and research focusing on statistical indicators of good ratings has been conducted by separate communities of researchers. This study demonstrates how existing latent trait modeling procedures can identify groups of raters who may be of substantive interest to those studying the experiential, cognitive, and contextual aspects of ratings. We employ two data sources in our demonstration—simulated data and data from a large-scale state-wide writing assessment. We apply latent trait models to these data to identify examples of rater leniency, centrality, inaccuracy, and differential dimensionality; and we investigate the association between rater background variables and rater effect flags.

*Keywords*: rater cognition, rater effects, rating quality
Application of Latent Trait Models to Identifying Substantively Interesting Raters

Important decisions are made in numerous contexts based on ratings that are assigned by trained judges. When the consequences of decisions that are based on these ratings are serious, those involved in the decision making process must verify that the ratings assigned by the judges are of suitable quality to warrant the intended inferences. In response to this need, a considerable amount of prior research has focused on variables that are believed to have a potential impact on the quality of ratings. For example, in an effort to determine why some raters assign ratings that are of a higher quality than ratings assigned by other raters, researchers have studied rater experience and prior knowledge, rater training procedures, and rater cognition. An equally volume of research exists focusing on statistical indices that may be useful in attempts to depict the quality of assigned ratings and to differentiate the various patterns of ratings that may detract from rating quality.

Unfortunately, there has been little effort to bring these two lines of research together. For example, studies of the characteristics of good raters and conditions that promote rating quality often do not utilize clearly defined indicators of rating quality. This limits the degree to which researchers can draw clear and valid inferences about the association between rater characteristics and rating conditions and the quality of the assigned ratings. In addition, different studies typically utilize different operational definitions of rating quality. While one study may differentiate raters based on judgment-based criteria of “expertise” and rater experience, another study may differentiate raters based on statistical criteria. Even when statistical indicators of rating quality are adopted, the generalizability of the results of studies of raters and rating conditions is limited unless researchers carefully consider what types of rater effects those indicators are best suited to detect. On the other hand, researchers who focus on developing and
evaluating statistical indicators of rater effects often conduct their studies using simulated data, making it impossible to take into determine the association between rater characteristics, rating conditions, and rating quality. Of the statistical studies that utilize real data, it is uncommon for the researchers to consider the complex causal antecedents for any observed rater effects.

As a result, our current knowledge concerning raters, rating conditions, and rating quality is limited. We know a good bit about what characteristics differentiate raters, and we know a good bit about which statistical indicators are best suited for detecting specific types of rater effects. What we lack is an understanding of the overlap of these two lines of research. How do rater characteristics and rating conditions influence the manifestation of rater effects? This article provides guidelines for how statistical indicators of clearly defined rater effects can be applied to identify groups of raters who would be substantively interesting to compare using procedures that are commonly adopted in studies of rater characteristics and rating contexts. Specifically, we identify several useful statistical indicators, explain how specific rater effects manifest themselves in the values of those indicators, and provide guidance in applying these indices to the task of identifying substantively interesting groups of raters. We demonstrate the applicability of these indices by analyzing both simulated and real data.

Studies of Raters and Rating Contexts

A brief review of studies of raters and rating contexts reveals that little attention has been directed toward determining the relationship between the cognitive, experiential, and contextual aspects of the process of rating and the quality of ratings. For example, several studies have been conducted that have simply focused on developing models that depict the rating process, rater cognition, rater training, or influences on rating quality (Barrett, 2001; Elders, Knoch, Barkhuizen, & von Randow, 2005; Freedman, 1979; Freedman & Calfee, 1983). Some studies
that compare behaviors of raters have focused on making comparisons based on rater experience or loosely defined “expertise” (Breland & Jones, 1984; Huot, 1988; Pula & Huot, 1993; Vaughan, 1991). A few studies have examined the association between rater characteristics and measures of rating quality that are not necessarily indicative of specific rater effects (Shohamy, Gordon, & Kraemer, 1992; Wolfe, Kao, & Ranney, 1997; Wolfe, 1997). Only a few recently published studies exist that explore the relationship between characteristics of raters and the rating context and specific measures of rater effects (Eckes, 2005; 2008).

Studies of Ratings

Much of the previous research concerning the quality of ratings has focused on identifying each of several rater effects, patterns of ratings produced by ratings that contain measurement error (Engelhard, 1994; Myford & Wolfe, 2003, 2004; Wolfe, 2004, 2005). The rater effect label applied to a particular rating pattern communicates the type of error that the ratings contain. One of the most commonly studied rater effects is rater severity or leniency—systematic error that causes the assigned ratings to be lower or higher (respectively) than ratings that do not contain error. That is, a lenient rater’s pattern of ratings across a sample of examinees may depict higher levels of performance than is warranted by the abilities of those examinees as shown in the following example:

Accurate Ratings  
[0,0,0,1,1,2,2,3,3,4,4,4,4,4,4]

Lenient Ratings  
[0,0,1,1,2,2,2,3,3,3,4,4,4,4,4,4].

Another rater effect that is a common concern in educational assessments is rater centrality—systematic error that causes the assigned ratings to be more tightly clustered than ratings that do not contain error. A centrality rater’s pattern may depict levels of performance that are closer to the middle of the rating scale than is warranted by the abilities of the examinees:
Accurate Ratings \[0,0,0,1,1,2,2,2,3,3,3,4,4,4]\n
Central Ratings \[0,1,1,2,2,2,2,2,3,3,3,4]\n
Rater inaccuracy occurs when seemingly random error that causes the assigned ratings to be inconsistent with accurate ratings. An inaccurate rater’s pattern may not adequately represent the true abilities of the examinees:

Accurate Ratings \[0,0,0,1,1,2,2,2,3,3,3,4,4,4]\n
Inaccurate Ratings \[4,4,3,4,1,2,4,4,2,3,4,2,2,1,4]\n
Differential dimensionality occurs when a ratings for a subgroup of raters systematic error that covaries between raters, a special case of multidimensionality, which constitutes a violation of the local independence assumption upon which most unidimensional latent trait models rely. Differential dimensionality may occur when some raters allow features of the examinee response that are not evidence of the examinee’s ability to influence the ratings (e.g., quality of handwriting or personal knowledge of the examinee). As a result, the ratings of this subgroup of raters will depict a combination of the examinee’s true ability and the secondary trait causing the ratings of the subgroup of raters to be internally consistent even though those ratings do not accurately represent the true abilities of the examinees.

A body of literature exists describing how rater effects may be detected in rating data, and much of that literature focuses on applications of the Rasch rating scale model for that purpose (Andrich, 1978). That model depicts the logit value of an examinee (\(n\)) 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 (\(\theta_n\)), rater (\(\lambda_r\)), and rating scale category threshold (\(\tau_k\)) onto a common underlying continuum.
\[
\ln \left( \frac{\pi_x}{\pi_{x-1}} \right) = \theta_n - \lambda_r - \tau_k,
\]

where \( k \) references the threshold between score category \( x \) and \( x - 1 \). It is worth noting that, in the current application, raters score examinee responses to a single item, omitting the commonly employed “item difficulty” parameter. Parameters for this model can be estimated using joint maximum likelihood estimation procedures as implemented in commercial software, such as *Facets* (Linacre, 2010) or *Winsteps* (Linacre, 2011).

Several statistical estimates associated with the Rasch rating scale model are useful for evaluating rater effects. As stated previously, the most commonly studied rater effects is rater severity or leniency, an effect that causes the scores assigned by a particular rater to be lower or higher, respectively, than warranted. The Rasch rating scale model explicitly estimates parameters to depict rater severity or leniency using the \( \lambda_r \) component of the model. Specifically, \( \lambda_r \) depicts the relative location of the mean score assigned by rater \( r \). Hence, in order to determine whether a particular rater assigns scores that are more severe or lenient than other raters, a data analyst would examine the value of \( \lambda_r \) for that rater.

Another potentially useful index is the *score-estimate correlation* \( (r_{x,\hat{\theta}}) \), also known as the *point-measure correlation*. The score-estimate correlation is the latent trait analog to the *item-total correlation*, often referred to as the *point-biserial correlation* when scores are dichotomous. In applications to detecting rater effects, the score-estimate correlation is computed as the correlation between the scores assigned by a particular rater to a group of examinees \( (x_r) \) and the ability estimates of those examinees \( (\hat{\theta}) \). The score-estimate correlation 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. Hence, the score-
estimate correlation should be sensitive to rater effects which create inconsistencies between these pairs of measures, such as rater inaccuracy or differential dimensionality. When the consistency is low, the value of the correlation coefficient should approach zero. It is important to note that, the score-estimate correlation should not be influenced by rater effects that preserve the consistency of these two measures, such as rater leniency or centrality.

A third set of indices that has been used to evaluate rater effects are four model-data fit indices associated with the Rasch rating scale model. These include the unweighted mean-squared fit statistics and the standardized version of this fit statistic. The mean squared fit statistics (Wright & Masters, 1982) are based on the standardized residual of the observed response for each person-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 \( r \),

\( E_{nr} = \sum_{k=0}^{m} k \pi_{nrk} \), the expected score assigned to person \( n \) by rater \( r \),

\( W_{nr} = \sum_{k=0}^{m} (k - E_{nr}) \),

\( k = \) the scored responses, ranging from 0 to \( m \), and

\( \pi_{nrk} = \) the model-based probability that person \( n \) be assigned a score in category \( k \) by rater \( r \).

**Unweighted mean squared fit statistics** for raters are computed as the average of the squared standardized residuals across all persons scored by a rater,
This statistic can also be standardized via the Wilson-Hilferty cube root transformation (Wilson & Hilferty, 1931) to obtain the **standardized unweighted mean square fit statistics** \((\text{ZUMS})\).

Historically, rule-of-thumb upper and lower limits for acceptable mean square fit values have been established for flagging items, such as 0.7 and 1.3 for multiple-choice items, 0.6 and 1.4 for rating scales, and ±2.0 for the standardized versions (Wright & Linacre, 1994).

Previous research has demonstrated that fit statistics may be sensitive to rater centrality effects, inaccuracy, and differential dimensionality (Engelhard, 1994; Wolfe, Chiu, & Myford, 2000; Wolfe & Moulder, 2001) making it difficult to differentiate these effects from each other (Myford & Wolfe, 2003, 2004; Wolfe, 2004, 2005). To address this problem, preliminary work has been done to develop an index that is primarily sensitive to centrality effects (Wolfe, 2004, 2005). Specifically, the **expected-residual correlation** \((r_{\text{exp, res}})\) is based on the notion that the residuals (i.e., observed score – expected score) produced by scoring that exhibit centrality will be positive for examinees of low ability and negative for examinees of high ability. That is, a scatterplot of the expected score (X axis) and residuals (Y axis), shown in Figure 1, should have a negative slope when centrality exists. Hence, analysts may be able to differentiate cases of rater inaccuracy and differential dimensionality from cases of rater centrality by first identifying suspect raters based on rater fit indices and then differentiating cases based on whether the expected-residual correlations are negative or not. Suspect raters that are associated with negative expected-residual correlations would be flagged for centrality while those associated with zero or positive expected-residual correlations would be flagged for inaccuracy or differential dimensionality.
Rater inaccuracy and differential dimensionality can be differentiated according to whether there are violations of the local independence (i.e., correlated residuals). Note that rater fit indices are sensitive to inflated values of residuals, and inaccuracy and differential dimensionality should both inflate the absolute values of residuals. Hence, inaccuracy and differential dimensionality should both manifest themselves in rater fit indices. Differentiation of these two effects requires determining whether the residuals themselves are correlated within a subset of raters. Because there is seldom reason to suspect a particular subset of raters as candidates for exhibiting differential dimensionality, analysts rely upon exploratory methods to uncover potential cases of this rater effect. Specifically, ratings analysts conduct principal component analysis (PCA) of the residuals associated with raters to determine whether extracted components identify groups of raters who produce correlated residuals. Because the process of scaling data using the Rasch model accounts for the common variance between all raters and is equivalent to extracting the first principal component in the ratings, any components that are identified in an analysis of the model residuals constitute secondary dimensions (i.e., internally consistent latent variables beyond the one the one defined jointly by all raters). Hence, if a PCA of the residuals identifies a cluster of raters who exhibit model-data misfit and have substantial loadings on a particular component, then those raters would be those who jointly define a secondary dimension and would, therefore, be flagged for differential dimensionality.

**Purpose**

The purpose of this study is to demonstrate how the Rasch rating scale model can provide evidence of specific and important rater effects. To that end, we first apply that model to simulated data that contains known rater effects to show that these procedures work as intended.
We then apply those methods to real data to demonstrate that these rater effects can be observed in operational settings.

**Simulation Example**

**Data Generation & Scaling**

We simulated rating data for 1000 examinees, each assigned a rating on a five-point scale (0 to 4) by each of 100 raters. Ability generating parameters ($\theta_{\text{generating}}$) were drawn from a N(0,1) distribution, and these generating thetas were transformed in three ways. First, we generated a rater-specific ability ($\theta_{nr}$) for each examinee-by-rater combination to allow us to introduce rater-specific randomness into the simulated data. Specifically, we generated data from a second, independent N(0,1) error distribution ($\theta_{\text{error}}$), and created a rater-specific ability value via

$$\theta_{nr} = (\theta_{\text{generating}} \times \rho_{xx}) + \left( \theta_{\text{error}} \times \sqrt{1 - \rho_{xx}^2} \right),$$

where $\rho_{xx}$ is the desired inter-rater reliability coefficient for the simulated data. Second, we generated a centrality ability ($\theta_{\text{centrality}}$) by regressing the values of $\theta_{nr}$ via

$$\theta_{\text{centrality}} = \theta_{nr} \times \omega_{\text{central}}$$

where $\omega_{\text{central}}$ is the desired degree to which a central rater shrinks the distribution of abilities. Third, we generated a secondary ability distribution ($\theta_{\text{secondary}}$) for each examinee-by-rater combination to allow us to introduce multidimensionality into the simulated data. Specifically, we generated data from another independent N(0,1) distribution ($\theta_{\text{dimension}}$), and created a multidimensional ability value via

$$\theta_{\text{secondary}} = (\theta_{nr} \times \rho_{\text{dimension}}) + \left( \theta_{\text{dimension}} \times \sqrt{1 - \rho_{\text{dimension}}^2} \right),$$
where $\rho_{dimension}$ is the desired correlation between the primary and secondary dimensions in the simulated data.

Each rater was assigned to one of five groups. *Normal* raters ($n = 84$) were simulated to exhibit ratings according to a Rasch rating scale model, using the values of $\theta_{nr}$, in which rater severity was drawn from a N(0,.15) distribution (i.e., raters who exhibit very little severity or leniency) and $\rho_{sx}$ equaled 0.90. *Lenient* raters ($n = 2$) were simulated to exhibit ratings according to a Rasch rating scale model, using the values of $\theta_{nr}$, in which rater severity was set to equal -1.00 and $\rho_{sx}$ equaled 0.90. *Central* raters ($n = 2$) were simulated to exhibit ratings according to a Rasch rating scale model, using the values of $\theta_{central}$ (with $\omega_{central}$ equal to 0.50), in which rater severity was drawn from a N(0,.15) distribution and $\rho_{sx}$ equaled 0.90. *Inaccurate* raters ($n = 2$) were simulated to exhibit ratings according to a Rasch rating scale model, using the values of $\theta_{nr}$, in which rater severity was drawn from a N(0,.15) distribution and $\rho_{sx}$ equaled 0.40. *Differential dimensionality* raters ($n = 10$) were simulated to exhibit ratings according to a Rasch rating scale model, using the values of $\theta_{dimension}$, in which rater severity was drawn from a N(0,.15) distribution (i.e., raters who exhibit very little severity or leniency), $\rho_{sx}$ equaled 0.90, and $\rho_{dimension}$ equaled 0.70. Scored data were generated by anchoring resulting examinee and rater parameters at these values via the SIFILE= option of *Winsteps* (2011). Once simulated data were generated, parameters were estimated using *Winsteps*, and we computed the values of each rater effect index discussed in the previous section (i.e., $\lambda_r$, $r_{sx,\theta}$, $UMS_r$, $ZUMS_r$, and $r_{exp,res}$).
Simulation Results

Table 1 summarizes the simulated raw score distributions for each rater effect group. The distributions of raw ratings for the Differential and Inaccuracy groups are nearly identical to the distribution of ratings for the Normal group, with the mean and standard deviation of the ratings for each of these groups being only slightly different than the values produced by raters in the Normal group. As would be expected, the mean rating for the Lenient group (2.77) is greater than the mean rating of the normal group (1.97). Also, the standard deviation of the ratings for the Central group (SD = 0.97) is less than the standard deviation for the Normal group (SD = 1.11). In addition, the correlation between the average rating assigned by raters in the Normal group displays a predictable pattern across rater effect groups. The correlation is greatest for the raters in the Normal and Lenient groups and is somewhat less for the Differential, Central, and Inaccuracy groups.

Table 2 contains summary statistics for each rater effect index by rater effect groups. The values of the rater calibration indices ($\lambda_r$) are inflated for the Leniency group in comparison to the values for the remaining groups. On the other hand, the values of the two fit indices ($UMS_r$ and $ZUMS_r$) are similar for the Normal and Lenient groups and are inflated for the remaining groups. The values are slightly inflated for the Central group, although the amount is not to a degree that would result in these raters being flagged for misfit. However, Inaccurate and Differential raters exhibit a considerably greater amount of model-data misfit. It is interesting to note that the values of the score-estimate correlation ($r_{x,\hat{\theta}}$) are similar for the Differential, Central, and Inaccurate groups, calling into question the usefulness of this index in the diagnosis of rater effects. Finally, the value of the expected-residual correlation ($r_{\text{exp,\text{res}}}$) is negative to a noticeable degree for the Inaccurate raters. The value of this index is even more extreme for the
Differential raters. However, as would be expected, the value of the expected-score index is negative to a substantial degree for the Central raters. It is only possible to differentiate Inaccurate and Differential raters by determining whether the residuals that contribute to the observed model-data misfit covary, and this is indicated by the loading of each rater on the first principal component extracted from a PCA of the model-based residuals. As displayed in the bottom row of Table 2, only the Differential raters exhibit substantial average loadings on this first principal component.

Hence, we can see the potential usefulness of these indices for differentiating the four rater effects that were simulated, as summarized in Table 3. Specifically, extreme negative values of \( \lambda_r \) differentiates Lenient raters from all other raters. Conversely, extreme positive values of \( \lambda_r \) would indicate rater severity. Extreme negative values of \( r_{\text{exp,res}} \) in the absence of substantial model-data misfit indicate the rater Centrality. Differential and Inaccurate raters both exhibit substantial model-data misfit, so these two rater effects are differentiated by conducting a PCA on the model-based residuals. Raters who exhibit inflated loadings on the first principal component of the residuals are those who share common variance beyond that accounted for by the measures and are, therefore, a source of differential dimensionality. Researchers who seek to better understand the influence of rater experience, rater cognition, and rating context on the quality of ratings can apply these rules to examples from real data in order to identify substantively interesting groups of raters to compare based on observation of those raters.

**Real Data Example**

**Real Data Method**

We provide a demonstration of these methods on real data, taken from a published study of the impact of rater training context on rating quality, and a detailed description of the sample
of raters and data collection are provided in a publication of that study (Wolfe, Matthews, & Vickers, 2010). They collected data from 40 essay raters under each of these three training conditions ($n = 120$). Raters participated in training activities, assigned scores on a 4 point rating scale to example essays in qualifying tests, and assigned scores through an online distribution system to a common set of 400 secondary student essays that were composed in response to a state-wide writing assessment. In this article, we scale the rating data using a Rasch rating scale model and examined the results for evidence of rater effects within that sample of raters by considering the indices and criteria presented in Table 3. Specifically, we flagged raters based on the following criteria: a Normal classification satisfied none of the following criteria; a Severe/Lenient classification meant that $|\text{r}| > 1.00$; a differential classification meant that $|\text{PC1 Loading}| > 0.50$ and $\text{UMS}_r > 1.40$; a central classification meant that $r_{\text{exp, res}} > -0.30$; an Inaccurate classification meant that $|\text{PC1 Loading}| \leq 0.50$ and $\text{UMS}_r > 1.40$; and an Unclassified classification satisfied the Inaccuracy and either the Severity or Centrality classification criteria.

**Real Data Results**

Table 4 summarizes the values of each rater effect index as observed in the real data. Clearly, evidence of potential rater effects exist in these data. The values of $\lambda_r$ range from -1.76 to 2.04, suggesting that there are both lenient and severe raters in the sample. $\text{UMS}_r$ values range from 0.48 to 1.83, suggesting that there are cases of rater misfit, which could result in raters being flagged for inaccuracy or differential dimensionality. In addition, the expected residual correlation ($r_{\text{exp, res}}$) indices range extend below zero to -0.51, indicating that there are probable cases of rater centrality. Finally, the values of the loadings on the first residual component ($\text{PC1 Loading}$) have a substantial maximum value (0.71).
Table 5 summarizes the proportion of raters who satisfied each rater effect classification criterion. A substantial majority of the raters satisfied none of the criteria and received a Normal classification. These raters assign ratings that do not exhibit any evidence of the rater effects considered in this article. The most common rater effect we observed was severity and leniency, which 18% of the raters in the pool exhibited. A nontrivial percentage of the raters (5%) exhibited evidence of assigning inaccurate ratings (i.e., assigned ratings that exhibited random misfit). Small percentages of the raters exhibited evidence of centrality (3%), and 4% of the raters assigned ratings that exhibited misfit coupled with either severity or potential centrality. None of the raters exhibited evidence of differential dimensionality based on the criteria that we adopted.

Discussion

In this article, we identified several rater effects that can be detected using latent trait modeling procedures, presented the indices and defined criteria for interpreting those indices as indicators of rater effects in ratings, and applied those criteria to simulated and real data. The simulation demonstrated the effectiveness with which those indices can identify known cases of rater effects. The real data application demonstrated that those effects can be observed in real world contexts in non-trivial amounts and that the classifications that we described account for nearly all of the observed rating patterns in our data. In fact, only 4 of the 120 raters in our sample exhibited patterns that did not fit into our classification scheme, and those four simply exhibited a combination of two of the effects that we described.

However, we believe that the potential usefulness and contribution of these procedures in studies of rater cognition, rater training, rater experience, and rating context has not been recognized by researchers of those topics. To date, most studies of the process of assigning
ratings have employed loosely-defined measures of a generic “rating quality” or have utilized no measures at all. Instead, they have tended to use somewhat generic measures of rating quality (e.g., interrater correlations, interrater agreement, or anecdotal accounts of expertise)—a fact that makes it nearly impossible to document the relationship between rating quality and cognitive, experiential, or contextual characteristics of raters and ratings. On the other hand, psychometric researchers have not sought to extend their analysis of ratings to the understanding of these characteristics. Rather, much of their work has focused on developing indices and applying them to real data sets without attempting to build an understanding of how the contextual, cognitive, and experiential characteristics relate to the manifestation of these clearly defined measures of rater effects.

We believe that an important next step in efforts to understand the psychometric and the cognitive, experiential, and contextual features of rating relate by merging these two lines of research that have remained, to a large degree, separate. By doing so, those engaging in substantive research can build more specific models with clearly defined and measure-based outcomes. Similarly, psychometric researchers can better understand how the technical work that they do fits into the efforts of those who select, train, and monitor raters, and conduct the rating process. We acknowledge that this may be a difficult goal to realize. Many of the methods that are used by substantive researchers are resource intensive, often necessitating collection of data from relatively small samples. Similarly, many of the psychometric models used to detect rater effects require raters assign a large number of ratings while following a highly structured data collection design. Regardless, it seems that better research can be produced by utilizing highly specific measures of rater effects as central variables in substantive research relating to raters and the rating process.
References


Table 1

*Descriptive Statistics for Simulated Raw Scores by Rater Effect Group*

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Mean</th>
<th>SD</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>84</td>
<td>11%</td>
<td>21%</td>
<td>37%</td>
<td>22%</td>
<td>9%</td>
<td>1.97</td>
<td>1.11</td>
<td>0.65</td>
</tr>
<tr>
<td>Differential</td>
<td>10</td>
<td>9%</td>
<td>19%</td>
<td>34%</td>
<td>25%</td>
<td>14%</td>
<td>2.16</td>
<td>1.15</td>
<td>0.40</td>
</tr>
<tr>
<td>Lenient</td>
<td>2</td>
<td>3%</td>
<td>8%</td>
<td>29%</td>
<td>30%</td>
<td>30%</td>
<td>2.77</td>
<td>1.04</td>
<td>0.60</td>
</tr>
<tr>
<td>Central</td>
<td>2</td>
<td>7%</td>
<td>23%</td>
<td>43%</td>
<td>22%</td>
<td>5%</td>
<td>1.97</td>
<td>0.97</td>
<td>0.42</td>
</tr>
<tr>
<td>Inaccurate</td>
<td>2</td>
<td>11%</td>
<td>22%</td>
<td>34%</td>
<td>22%</td>
<td>12%</td>
<td>2.02</td>
<td>1.17</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Note: r = the correlation between the average score assigned by raters in the Normal group and a single randomly chosen rater in each remaining group.
### Table 2

**Average Simulated Rater Effect Indices by Rater Effect Group**

<table>
<thead>
<tr>
<th>Index</th>
<th>Normal</th>
<th>Differential</th>
<th>Lenient</th>
<th>Central</th>
<th>Inaccurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>84</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$\lambda_r$</td>
<td>0.02</td>
<td>-0.04</td>
<td>-0.94</td>
<td>0.28</td>
<td>0.10</td>
</tr>
<tr>
<td>$UMS_r$</td>
<td>0.94</td>
<td>1.41</td>
<td>0.98</td>
<td>1.13</td>
<td>1.32</td>
</tr>
<tr>
<td>$ZUMS_r$</td>
<td>-1.50</td>
<td>8.59</td>
<td>-0.53</td>
<td>2.98</td>
<td>6.90</td>
</tr>
<tr>
<td>$r_{x,\theta}$</td>
<td>0.65</td>
<td>0.43</td>
<td>0.61</td>
<td>0.44</td>
<td>0.49</td>
</tr>
<tr>
<td>$r_{\exp,\res}$</td>
<td>0.05</td>
<td>-0.22</td>
<td>0.02</td>
<td>-0.35</td>
<td>-0.17</td>
</tr>
<tr>
<td>PC1 Loading</td>
<td>-0.09</td>
<td>0.62</td>
<td>-0.03</td>
<td>0.15</td>
<td>-0.13</td>
</tr>
</tbody>
</table>

**Note:** PC1 Loading is the loading on the first component extracted from a PCA of the model residuals.
Table 3

Statistical Indicators of Rater Effects

<table>
<thead>
<tr>
<th>Index</th>
<th>Differential</th>
<th>Lenient</th>
<th>Central</th>
<th>Inaccurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_r$</td>
<td>--</td>
<td>Negative</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>$UMS_r$</td>
<td>Inflated</td>
<td>--</td>
<td>--</td>
<td>Inflated</td>
</tr>
<tr>
<td>$ZUMS_r$</td>
<td>Inflated</td>
<td>--</td>
<td>--</td>
<td>Inflated</td>
</tr>
<tr>
<td>$r_{exp, res}$</td>
<td>Negative</td>
<td>--</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>PC1 Loading</td>
<td>Inflated</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
### Table 4

*Rater Effect Indices for the Real Data*

<table>
<thead>
<tr>
<th>Index</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_r$</td>
<td>0.00</td>
<td>0.74</td>
<td>-1.76</td>
<td>2.04</td>
</tr>
<tr>
<td>$UMSr$</td>
<td>0.99</td>
<td>0.28</td>
<td>0.48</td>
<td>1.83</td>
</tr>
<tr>
<td>$ZUMSr$</td>
<td>0.41</td>
<td>3.90</td>
<td>-9.20</td>
<td>9.18</td>
</tr>
<tr>
<td>$r_{x,\delta}$</td>
<td>0.76</td>
<td>0.06</td>
<td>0.52</td>
<td>0.88</td>
</tr>
<tr>
<td>$r_{exp,res}$</td>
<td>0.01</td>
<td>0.18</td>
<td>-0.51</td>
<td>0.44</td>
</tr>
<tr>
<td>PC1 Loading</td>
<td>0.01</td>
<td>0.26</td>
<td>-0.39</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Table 5

*Rater Effect Classifications for the Real Data*

<table>
<thead>
<tr>
<th>Classification</th>
<th>Raters</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>86</td>
<td>71.7%</td>
</tr>
<tr>
<td>Severe/Lenient</td>
<td>21</td>
<td>17.5%</td>
</tr>
<tr>
<td>Differential</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Central</td>
<td>3</td>
<td>2.5%</td>
</tr>
<tr>
<td>Inaccurate</td>
<td>6</td>
<td>5.0%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>4</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

Note: Raters were classified according to the following criteria. *Normal*: Satisfies none of the following criteria. *Severe/Lenient*: $|\lambda_i| > 1.00$. *Differential*: $|PC1 Loading| > 0.50 \& UMS_r > 1.40$. *Central*: $r_{exp,res} > -0.30$. *Inaccurate*: $|PC1 Loading| \leq 0.50 \& UMS_r > 1.40$. *Unclassified*: Satisfies Inaccuracy & either (Severity or Centrality).
Figure Captions

*Figure 1.* Residual-Expected Score Relationship for Central Raters