Psychology Science, Volume 46, 2004 (1), p. 35-51

Identifying rater effects using latent trait models

EDWARD W. WOLFE[1]

Abstract

This study describes how latent trait models, specifically the multi-faceted Rasch model, may be applied to identifying rater effects in performance ratings. Three general types of rater effects are presented (Accuracy/Inaccuracy, Severity/Leniency, and Centrality/ Extremism), and indices that are useful indicators of those rater effects are identified. Each of these indices is examined in the context of ratings assigned by highly trained raters to essays written for an Advanced Placement examination, and individual raters suspected of exhibiting these various rater effects are identified. Each of these rater effects exists in the data in both non-ignorable rates and magnitudes.

Key words: Educational Measurement, Performance ratings, Psychometrics, Rater effects, Ratings

[1] Edward W. Wolfe, Measurement & Quantitative Methods, 450 Erickson Hall, Michigan State University, East Lansing, MI 48824, E-mail: wolfee@msu.edu

Identifying Rater Effects Using Latent Trait Models

Human judgment is used to assign ratings in a wide variety of domains. Judges rate the athletic performance of gymnasts and divers. Connoisseurs rate the quality of wine and beer. Employers rate the suitability of job applicants. Judges rate the degree to which animals exhibit the idealized characteristics of that animal's pedigree. Educators rate the performance of students on constructed-response achievement test items. When important decisions are made based upon such ratings, it is essential that the assigned ratings are accurate and fair. As a result, rater selection, training, and monitoring procedures are often employed for the purpose of minimizing the impact of rater inaccuracy or bias on ratings.

One example of efforts such as these arises in educational achievement tests that require the examinee to write an essay in response to an examination question. Such questions are common on the Test of English as a Foreign Language (TOEFL), the Scholastic Aptitude Test (SAT), and the Advanced Placement (AP) examinations. In these large-scale, high-stakes examination settings, raters typically must meet minimum qualifications in order to be selected as a rater, such as having earned a graduate degree and having teaching experience in the content area of the examination. Once selected for a rating project, raters undergo hours or even days of training to learn the criteria upon which ratings will be assigned, review rated exemplars of examinee responses, practice assigning ratings to example responses, discuss the ratings they assign with other raters to clarify the fine points of the rating criteria, and receive feedback on their progress toward mastering the rating criteria. Often, raters must demonstrate mastery of the rating criteria by achieving a minimum level of agreement with expert raters on pre-scored example responses prior to assigning ratings that are reported to students. In addition, raters are frequently monitored throughout the rating session to verify that their ratings maintain desirable levels of accuracy, and raters who exhibit drift away from the adopted standards for performance may be retrained or removed from the rating session.

Regardless of the efforts taken to minimize inaccuracy and bias in ratings, idiosyncrasies exist in the behaviors of raters, and systematic patterns in these behaviors are termed *rater effects*. Because rater effects are systematic, they are detectable as patterns in the ratings assigned by the raters. This manuscript identifies several classes of rater effects, explains how a latent trait measurement model can be used to detect these rater effects, and presents an empirical example of such an application.

Rater Effects

Most research concerning rater effects has examined those effects using three methodologies. One line of research has focused on rater cognition and the relationship between a rater's cognitive processing and focus and that rater's proficiency (Breland & Jones, 1984; Freedman, 1979; Freedman & Calfee, 1983; Pula & Huot, 1993; Vaughan, 1991; Wolfe, 1997; Wolfe, Kao, & Ranney, 1998). These types of studies tend to emphasize differences between expert and novice raters, and the results mirror expert-novice differences observed in other domains (Glaser & Chi, 1988). Specifically, those studies have shown that expert raters tend to focus on very specific characteristics of the essays and conceptualize the rating process holistically.

Another line of research has focused on characteristics of raters, the rating task, and the rating environment that relate to the presence of rater effects in ratings (Dean, 1980; Hoyt, 1999; Hoyt, 2000; Jako & Murphy, 1990; McIntyre, Smith, & Hassett, 1984; Murphy & Anhalt, 1992; Tziner & Murphy, 1999; Welch & Swift, 1992; Yu & Murphy, 1993). Some of these studies have suggested that the manner in which information is presented to raters and the way that raters process that information may introduce certain types of rater effects, such as proximity errors or halo effects. Others have suggested that certain types of errors have a more severe impact on the reliability and validity of ratings.

The third line of research has focused on the impact of rater effects on ratings and developing methods for statistically modeling and correcting for rater effects (Braun, 1988; de Gruijter, 1984; Engelhard, 1992, 1994, 1996; Houston, Raymond, & Svec, 1991; Longford, 1996; Lunz, Wright, & Linacre, 1990; Murphy & Balzer, 1989; Raymond & Viswesvaran, 1993; Vance, Winne, & Wright, 1983; Wolfe, Chiu, & Myford, 2000). Several of these authors have relied on analyses of raw scores and applications of generalizability theory. More recent efforts have focused on the development and utilization of latent trait applications. The study presented here is consistent with these more recent efforts. Prior to presenting those latent trait procedures, however, several types of rater effects that may be detected using these methods will be presented.

Accuracy/Inaccuracy

One of the most common concerns of those who utilize ratings to make decisions is whether raters have been sufficiently trained or have sufficient expertise to assign accurate, rather than inaccurate, ratings. The ability to assign accurate ratings (i.e., the demonstration of rater *accuracy*) may be the result of experiences that the rater has had (e.g., training, education, work experiences), cognitive factors (e.g., thinking styles, learning abilities), and characteristics of the rating criteria (e.g., degree of similarity of the criteria and the rater's beliefs and understanding of the domain in question) and the rating environment (e.g., freedom from distractions, types of social interactions that occur in that setting). Hence, the most suitable raters maybe those who have thinking styles and belief systems that are consistent with the training approach and rating criteria. In addition, rater training, rater monitoring, and the structure of the rating task may facilitate accuracy by providing a suitable rating environment. *Inaccuracy*, the converse of accuracy, is a rater effect that is typically avoided by those who supervise the assignment of ratings.

Rater accuracy and inaccuracy can be defined statistically by considering the impact that each of these effects has on the residuals of the observed scores from their expectations. The expected rating is the value of the expectation of assigned ratings across an innumerably large number of contexts (raters, time, items, etc.). This expectation is sometimes referred to as the "true" rating (in true score test theory) or as the expected score derived from the rated individual's parameter-based ability (in latent trait test theory). While rater accuracy results in a high degree of consistency between assigned ratings and expected, or known-to-be-valid, ratings[1], rater inaccuracy results in low levels of consistency between assigned ratings and expected ratings. In this sense, accuracy manifests itself as small, randomly distributed patterns of residuals, where the residual is defined as the difference between the observed rating (X) and its expectation [E(X)], $residual = X - E(X)$. That is, rater accuracy should result in a

standard deviation of the residuals that is small, and a correlation between the residual and E(X) that is near zero. Rater inaccuracy would also result in independence between residuals and E(X) (i.e., $r_{residual,expected}$ is near zero), but, the standard deviation of the residuals would be large. These patterns are displayed in Figure 1.

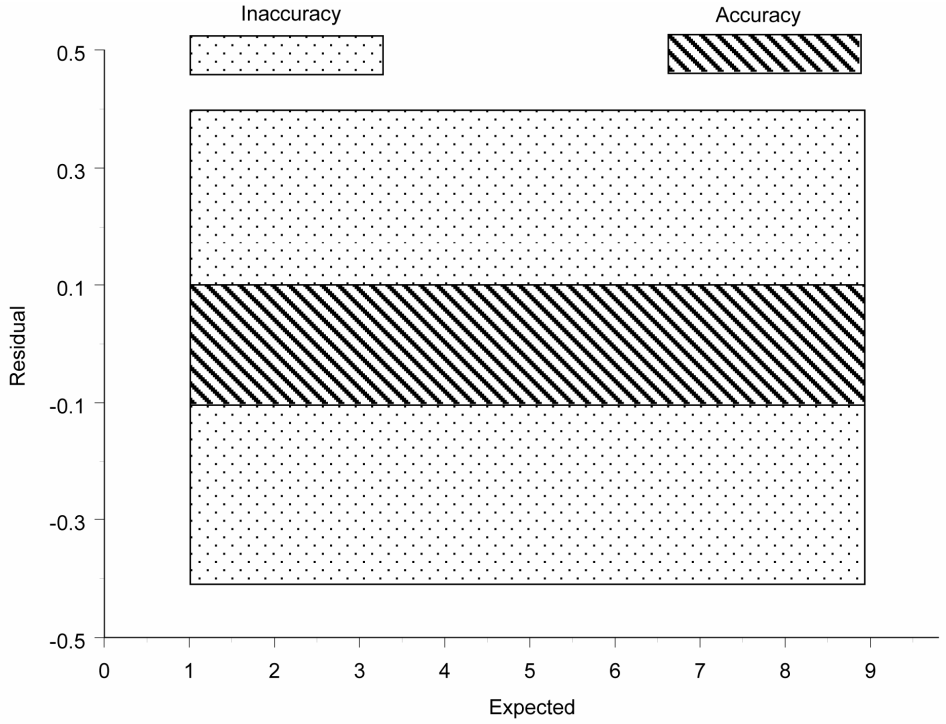

Figure1:
Rater Accuracy and Inaccuracy Residual Patterns

Harshness/Leniency

The most commonly researched rater effect occurs when raters utilize the same criteria but adopt different levels of stringency within that framework. In these cases, the rank ordering of ratings assigned by different raters is very consistent, but the average ratings are different. Raters who tend to assign lower scores are said to be *severe* or *harsh* while raters who tend to assign higher scores are said to be *lenient*. In the purest case, harsh and lenient raters assign ratings that are perfectly correlated with expected ratings. As a result, the standard deviation of the observed residuals would be near zero, and the residuals and E(X) will be independent of one another. Rater harshness and leniency can be explicitly modeled in the latent trait framework described here as locations of raters on the underlying linear contin-

uum, and the residuals from this model take those locations into account. As a result, the mean residual would be near zero as well.

In norm-referenced measurement frameworks (i.e., those in which ratings derive their meaning from their relative standing in the distribution of ratings), rater harshness and leniency do not influence relative standings of objects of measurement when comparisons are made between raters. When all raters rate all objects of measurement and only rater harshness or leniency exists (i.e., no other rater effects, such as inaccuracy), a simple average or sum of the ratings assigned by all raters reproduces the distribution of expected ratings with only a constant shift of that distribution. As a result, norm-referenced comparisons would result in perfectly reliable decisions about the objects of measurement in the case of pure harshness and leniency with a complete rating design. However, when rating designs are incomplete (i.e., when only a portion of the raters assign ratings to any given object of measurement[2] – a characteristic that is common in large-scale rating projects for the sake of minimizing cost), the existence of rater harshness or leniency may influence the relative standing of individual objects of measurement, depending on the level of harshness or leniency of the subset of raters who assign ratings to that object of measurement. Clearly, the influence of rater harshness and leniency can be minimized by increasing the number of raters who assign ratings to any given object of measurement and ensuring that the assignment of raters to objects of measurement is random. Generalizability Theory provides a measurement model that allows for the prediction of reliability under various rating designs (Brennan, 1992; Shavelson & Webb, 1991).

A similar, but slightly more complicated, problem exists when harshness and leniency enter into the ratings assigned in a criterion-referenced interpretive framework. In this case, decisions are not accurate under complete or incomplete rating designs. In the case of complete rating designs, the shift of the distribution of ratings from the distribution of expected ratings by a constant results in a different proportion of objects of measurement above and below the established cut score. In the case of incomplete rating designs, the shift of subsets of objects of measurement associated with each rater or subset of raters will result in similar changes in the proportion of objects of measurement falling above or below the cut score, but the magnitude of these shifts will depend on the degree to which harshness or leniency is pervasive in that rater subset's ratings. In either case, increasing the number of raters will only result in a reproduction of the distribution of expected ratings if the distribution of rater harshness and leniency is symmetrically distributed around a value of zero.

Centrality/Extremism

A less common concern, but one that is likely to be prevalent in rating projects in which raters are monitored using methods that rely on inter-rater comparisons, occurs whether raters have sufficiently mastered the rating criteria but they apply the extreme categories of the rating scale differentially. Specifically, raters who tend to assign fewer scores at both the high and low ends of the rating scale are said to exhibit *centrality*. This results in a concentration of assigned ratings in the middle of the rating scale. A similar rater effect, one that will not be explored in this article, named *restriction of range* exists when centrality is combined with leniency or harshness (Saal, Downey, & Lahey, 1980). That is, restriction of range results in a restricted dispersion of ratings around a non-central location on the rating

scale. The converse of rater centrality occurs when raters tend to overuse the extreme rating scale categories – a rater effect called *extremism.*

Rater centrality and extremism manifest themselves in both the pattern and the spread of the residuals. When centrality occurs, the observed ratings regress toward the center of the rating scale. As a result, residuals tend to be large and positive for low expected ratings and large and negative for high expected ratings, as shown in Figure 2. However, to understand the impact of this trend on summary statistics for the residuals, one must consider the distribution of the expected scores. Just above the x-axis of Figure 2, a unimodal, centered, and symmetrical distribution of expected ratings has been superimposed. Notice that the greatest density of expected ratings corresponds to the point on the residual axis where residuals are small. As a result, taking the standard deviation of the residuals across this particular sample of expected ratings results in a residual standard deviation that is near zero. Also, note that the pattern of residuals is negatively correlated with the expected ratings due to rater centrality (i.e., $r_{residual,expected}$ approaches -1.00).

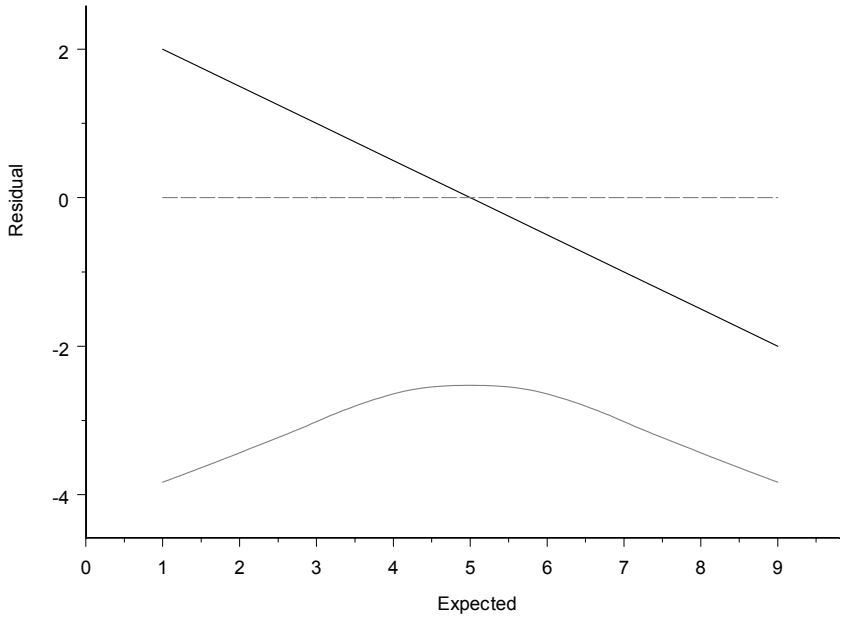

Figure 2:
Rater Centrality Residual Patterns

Figure 3 demonstrates the influence of rater extremism on the residuals. When extremism occurs, the observed ratings are pushed into the tails of the rating scale. As a result, residuals tend to be near zero for extreme expected ratings. The absolute value of the residuals increases as the expected rating approaches the center of its distribution. Taking into account the distribution of expected ratings superimposed near the x-axis of Figure 3, it seems that

the standard deviation of the residuals will likely be large because the largest absolute residuals are located at the most dense location of the distribution of expected ratings. In addition, because of this density pattern, $r_{residual,expected}$ will approach 1.00.

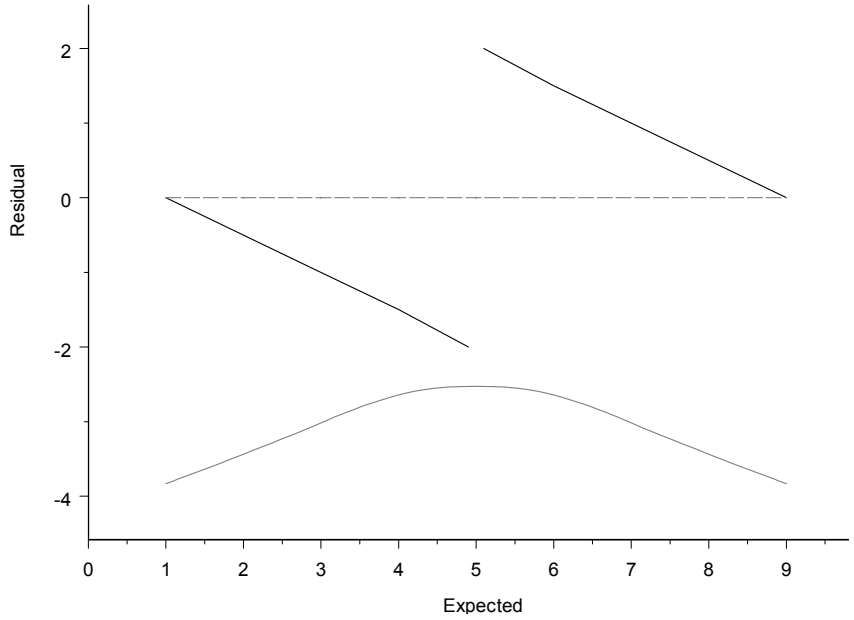

Figure 3:
Rater Extremism Residual Patterns

A Statistical Summary

In the previous discussion, several types of rater effects were described according to the behavior of the residuals that arise when those rater effects occur. Table 1 summarizes those characteristics. Specifically, we expect accurate raters to produce residuals that have a standard deviation close to zero with no correlation between the magnitude of the residual and the expected rating. Inaccurate raters are expected to produce residuals that also exhibit no correlation with the expected ratings, but the standard deviation of the residuals will be large. Lenient and harsh raters will produce residuals that are tightly clustered (i.e., the standard deviation will be close to zero), and the residuals will not be correlated with the expected ratings. As explained in the next section, a parameter of the latent trait model will be useful for differentiating these two effects from the accuracy effect. Central raters will produce residuals with standard deviations near zero. However, unlike accurate raters, central raters will produce residuals that are negatively correlated with expected scores. Similarly, extreme raters will produce residuals with large standard deviations, and the residuals will be positively correlated with expected scores.

Table 1:
Statistical Summary of the Influence of Rater Effects on Residuals

Rater Effect	$SD_{residual}$	$r_{residual,E(X)}$
Accuracy	0	0
Inaccuracy	•	0
Leniency	0	0
Severity	0	0
Centrality	0	-1.00
Extremism	•	1.00

The Multifaceted Rasch Model

The multifaceted Rasch model extends the family of Rasch models to cases in which multiple facets of measurement (i.e., elements of the measurement context that may contribute to measurement error) are modeled (Linacre, 1994; Wright & Masters, 1982). In the case of rater effects, the multifaceted Rasch model may be used to control for error contributed by systematic variability between both items and raters. Although there are several versions of the multifaceted Rasch model, only the Multifaceted Rasch Rating Scale Model (MRRSM) will be discussed in this article.[3] As with all Rasch models, the MRRSM depicts the additive contribution of each element of the measurement context to the log of the odds (the *logit*) of observing one rating scale category versus the next lower rating scale category using parameters that represent the object of measurement and facets of the measurement context such as raters and items. Mathematically, we can express this relationship as

$$LN\left(\frac{P_x}{P_{x-1}}\right) = \theta_n - \lambda_r - \delta_{ik}\,, \tag{1}$$

where θ_n is the location of the object of measurement on the underlying linear continuum, λ_r is the location of rater *r*, and δ_{ik} is the location of the threshold (*k*) between categories *x* and *x*-1 on the rating scale for item *i*. In the case of the MRRSM, the location of each threshold is held constant across all items. Rasch scaling software typically employs maximum likelihood methods to estimate the values of parameters based on observed ratings.

As is evident in this equation, the location of the underlying rater directly influences the probability of observing a particular category for the object of measurement in question. Hence, it is clear that λ_r depicts the harshness or leniency of rater *r*. By examining the relative magnitudes of the λ_r estimates for a particular set of ratings, one can identify raters who are harsh or lenient relative to the pool of raters. Standard errors can be estimated for each of these parameter estimates, and Wald statistics

$$\chi^2_{Wald} = \left(\frac{\lambda_r}{SE_{\lambda_r}}\right)^2, \tag{1}$$

can be computed to identify raters who deviate from a group mean (i.e., null value) of 0 by a statistically significant degree[4].

Model-based expected values can also be computed for each rater-by-measurement object-by-item combination,

$$E_{nir} = \sum_{n=1}^{N} \sum_{i=1}^{I} k\pi_{nirk} \,, \tag{3}$$

where

$$\pi_{nirk} = \frac{\exp \sum_{j=0}^{x} \theta_n - \lambda_r - \delta_{ik}}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} \theta_n - \lambda_r - \delta_{ik}}, \tag{4}$$

$$\delta_{i0} \equiv 0 \text{ so that } \sum_{j=0}^{0} \theta_n - \lambda_r - \delta_{ik} = 0 \,,$$

and x is a count of the number of rating scale thresholds exceeded by an object of measurement (ranging from 0 to m).

The residuals of observed ratings from these model-based expectations, $X_{nir} - E_{nir}$, can be used to identify rater accuracy/inaccuracy and rater centrality/extremism as specified in the previous section of this article. Specifically, smaller residual standard deviations are associated with rater accuracy and centrality, while larger residual standard deviations are associated with rater inaccuracy and extremism. Rater accuracy and centrality can be differentiated by examining the correlation between the residual and the expected rating associated with that residual ($r_{residual,expected}$). If the correlation between the residual and expected rating is near zero then rater accuracy is indicated. On the other hand, if the residual-expected correlation is negative, then rater centrality is indicated. Rater inaccuracy and extremism can be differentiated in a similar way – the correlation between residuals and expected ratings should be near zero in the case of rater inaccuracy and should be positive in the case of extremism.

Method

Instrumentation

Each year, high school students in the United States participate in *Advanced Placement* (AP) courses, and they may elect to take an examination covering the content of that course in order to receive college credit for participating in the AP course. The AP course in *English Literature and Composition* engages students in reading and critical analysis of imaginative literature. Through the reading of selected texts, students deepen their understanding of the ways writers use language to provide both meaning and pleasure for their readers. The examination for the English Literature and Composition course contains a 60-minute multiple-

choice section and two 60-minute essay questions (Advanced Placement Program, 2003a). The data for this study were taken from a single essay prompt requiring students to read and respond to a humorous text.

Raters

Essay questions for the *English Literature and Composition Examination* were scored by 101 readers who were selected to be highly qualified and were trained extensively to provide fair, uniform, and accurate ratings. Each year, developers of the AP examinations create detailed scoring guidelines, thoroughly train all readers, and implement various checks and balances throughout the AP Reading. The scoring guidelines describe the characteristics of each category of the nine-point rating scale, and exemplars of student essays are used to illustrate those characteristics during rater training. Raters also rate and discuss prescored examples of student essays. Prior to rating operational essays, raters must demonstrate mastery of the scoring guidelines by achieving a minimum level of agreement with a set of prescored student essays. During the operational rating session, numerous steps are taken to ensure that grading is done fairly and consistently (e.g., student identification information is concealed, clerical aids minimize paperwork performed by readers, scores of other readers are masked, readers are routinely retrained and their performance is monitored by expert readers) (Advanced Placement Program, 2003b).

Procedures

In an operational AP reading, only a single rating is assigned to each examinee's essay. To remove the confounding of rater effects and examinee abilities that is created by the nested rating design (i.e., raters nested within examinees), 28 essays were distributed to and rated by the 101 operational readers who participated in this study. These essays were also assigned a consensus score by a panel of six expert raters, but these scores were concealed from the operational raters.

Analyses

Ratings for the 28 essays assigned by the 101 operational raters were scaled to a multifaceted Rasch rating scale model using the *Facets* computer program (Linacre, 2003). Rater location parameter estimates and the associated standard errors, and residuals and expected ratings for each rater-by-examinee combination were created. These statistics were used to create a standard deviation of the residuals and a correlation between the residuals and expected ratings for each rater.

Results

Harshness/Leniency

Figure 4 displays the rater location parameter estimates, sorted from the most lenient rater to the harshest rater, each with a 95% confidence band drawn around it. A Bonferroni correction for multiple comparisons was applied to these confidence bands to control for the experiment-wise Type I error rate. Note that six of the raters (6%) exhibit leniency that is statistically significant in its difference from the group mean of zero. Similarly, five of the raters (5%) exhibit statistically significant harshness. For illustration, the average rating assigned by the harshest rater equals 3.42 while the average rating assigned by the most lenient rater equals 5.68. The average of all ratings assigned to the set of 28 essays by the 101 raters equals 4.39. From these figures, it is clear that the rater harshness and leniency may have a profound impact on the rating assigned to an examinee. In fact, if an arbitrary cut-point was imposed on the raw ratings (e.g., 5 or greater is a passing rating – the 50th percentile of the expert ratings), 75% of the examinees rated by the most lenient rater would pass while only 14% of the examinees rated by the harshest rater would pass[5].

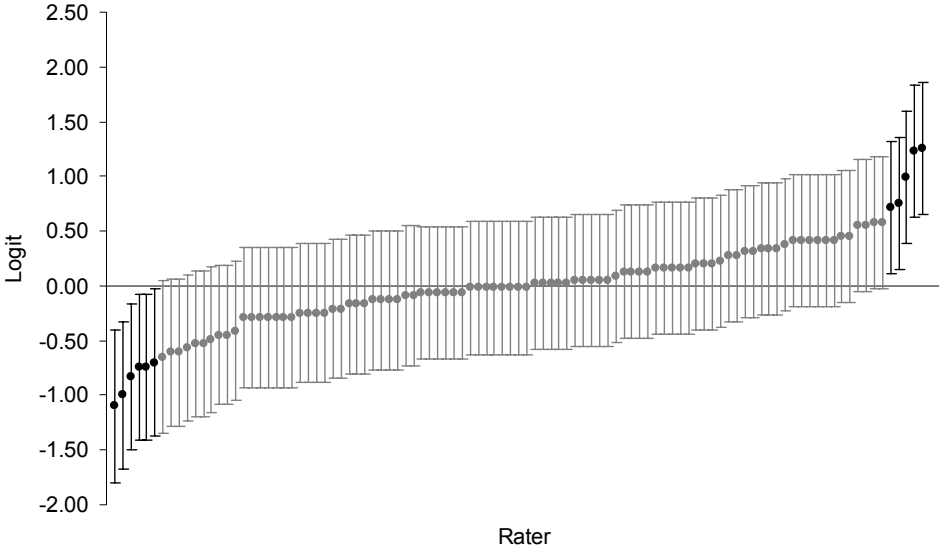

Figure 4:
Rater Harshness and Leniency Estimates

Accuracy/Inaccuracy/Centrality

Figure 5 displays the scatterplot of the standard deviations of the residuals (x-axis, $SD_{residual}$) and the correlations between residuals and expected ratings (y-axis, $r_{residual,expected}$). Overlaid on this plot are regions associated with rater accuracy, inaccuracy, centrality, and extremism. Note that the boundaries of the regions of $SD_{residual}$ axis were arbitrarily set to 0.75 and 1.25 and the boundaries of the regions of $r_{residual,expected}$ are defined by the critical value of that correlation coefficient. As described previously, accuracy is depicted by a small $SD_{residual}$ and a near-zero value of $r_{residual,expected}$, inaccuracy is depicted by a large $SD_{residual}$ and a near-zero value of $r_{residual,expected}$, centrality is depicted by a small $SD_{residual}$ and a negative value of $r_{residual,expected}$, and extremism is depicted by a large $SD_{residual}$ and a positive value of $r_{residual,expected}$.

Overall, 10% of the raters met the accuracy criteria, 1% of the raters met the inaccuracy criteria, 4% met the centrality criteria, and 2% met the extremism criteria. For the purpose of verifying that the criteria identify meaningfully large patterns of rater effects, consider the information shown in Table 2. The top section of that table compares the agreement rate between two raters – one meeting the accuracy criteria and one meeting the inaccuracy criteria – and the rounded average rating of all 101 raters. These figures demonstrate that the accurate rater was within one rating scale category of the average of all raters on 86% of the

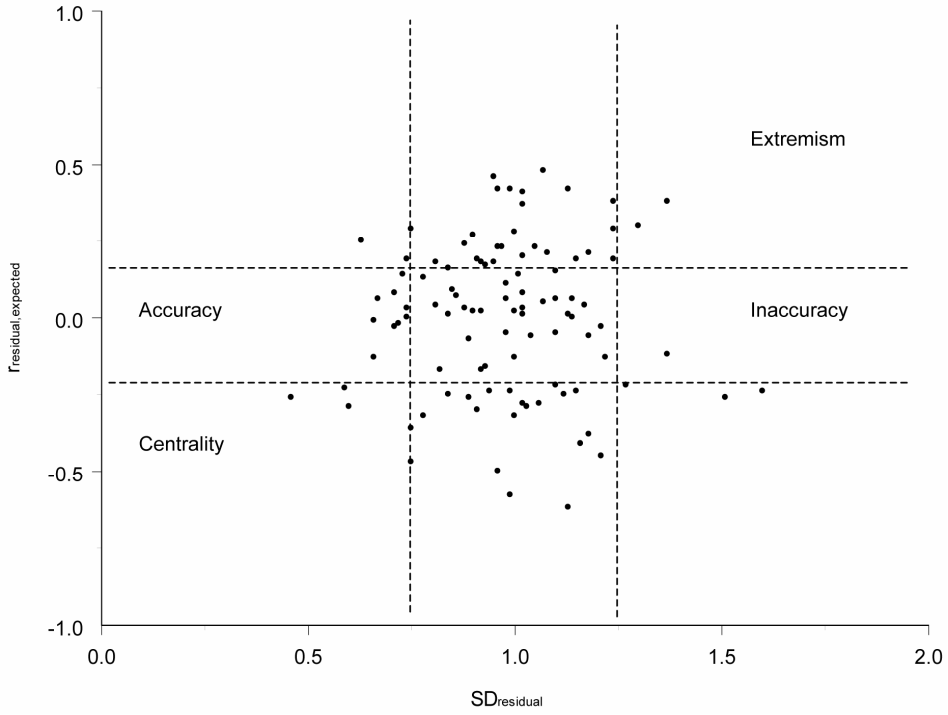

Figure 5:
Rater Accuracy, Inaccuracy, and Centrality Indices

Table 2:
Number of Essays Consistent with Rater Effect

Rater	Deviation from Mean of All Raters	
	0 to 1 point	2 or more points
Accurate	24	4
Inaccurate	20	8
	Proximity to Own Mean versus Proximity of Mean of All Raters to Overall Mean	
	Closer	Farther
Central	16	12
Extreme	6	22

essays while the inaccurate rater was within one rating scale category of the average of all raters on only 71% of the essays. In rating programs that employ adjudication procedures when two raters assign ratings that are more than one point apart, the adjudication rate of the inaccurate rater would be considerably higher than the adjudication rate of the accurate rater.

Similarly, the bottom section of that table compares the distance the ratings of two raters – one meeting the centrality criteria and one meeting the extremism criteria – lie from their own mean ratings to the distance each average rating (across the 101 raters) lies from the mean of all of these averages. These figures clearly demonstrate that the extreme rater assigns ratings that are much more widely dispersed than a typical rater – 79% of the ratings assigned by that rater were farther away from that rater's mean than were the average ratings from their own mean. Similarly, albeit a less pronounced effect, the central rater assigned ratings that were closer to that rater's mean (than the average rating was to the mean of all of these average ratings) 57% of the time.

Frame of Reference

It is important to note that the frame of reference for interpreting these rater effects has been the pool of raters. This requires one to assume that, on average, the pool of raters contains only a very small proportion of raters who exhibit rater effects. For example, because rater harshness and leniency are measured as deviations from the average rating across all raters, an implicit assumption required to interpret rater location parameter estimates is that there is no prevalence of harshness or leniency among the raters in the pool. This may or may not be the case. For example, in the data reported here, the average rating across the 101 raters and the 28 essays equals 4.39. On the other hand, the average consensus score assigned by the pool of six expert raters to these same essays equals 3.79. Hence, on average, raters in this study were lenient when compared to the experts. This fact illustrates an important problem in the analysis of rater effects – nearly all of the procedures based on latent trait measurement models employ a normative frame of reference and the use of such a frame of reference requires an assumption that rater effects are randomly distributed and are exhibited by only a minority of the raters in the pool.

Discussion

Rater effects may manifest themselves in a variety of contexts involving human judgment, and the existence of rater effects may threaten the validity of decisions that are made based on those ratings. This article has demonstrated that statistics derived from the multifaceted Rasch rating scale model may provide information that is useful to those who wish to evaluate the quality of ratings. Because much of the effort to study rater effects has focused on rater harshness and leniency, this article simply demonstrates the usefulness of the MFRRSM for detecting these rater effects.

On the other hand, this article also proposes and demonstrates the utility of two residual-based indices that may be useful in the diagnosis of rater accuracy, inaccuracy, centrality, and extremism – effects that have received considerably less attention in the literature relating to rater effects. Specifically, this article provides an explanation concerning how the standard deviation of model-based residuals and the correlation of those residuals from the model-based expectations should behave in the presence of these rater effects, illustrates that the patterns of ratings that are flagged by these indices are consistent with the hypothesized rater effects, and provides evidence that these rater effects even exist in the ratings assigned in highly structured and standardized settings (e.g., settings like AP rating sessions). Previous efforts to measure rater effects using model-to-data fit indices have demonstrated mixed levels of success in differentiating between and identifying the existence of these particular rater effects (Engelhard, 1994; Wolfe et al., 2000). A useful extension of the information presented in this article would be a series of simulation studies designed to document the sampling distributions of these indices and the rates with which these indices accurately and inaccurately nominate (or fail to nominate) ratings that are simulated to exhibit each of these rater effects.

A very important topic that has received little attention in the literature relating to rater effects is the interpretive frame of reference within which rater effects are portrayed. A serious shortcoming of the methods described in this article is their reliance on an implicit assumption that rater effects are distributed in the pool of raters in a non-systematic manner. As a result, in rater pools in which a particular rater effect is pervasive, a minority of highly competent raters may be flagged for exhibiting aberrant rating patterns because the frame of reference is the pool of non-competent raters. Although procedures have been developed for depicting rater effects relative to a framework other than the pool of raters (Engelhard, 1996; Wolfe, 1998), these procedures have not been thoroughly researched and have not been adopted as standard rater monitoring practice.

Finally, this article focuses exclusively on what may be appropriately labeled *static* rater effects (i..e., rater effects that are assumed to be an immutable characteristic of the rater over time). Previous research has demonstrated that rater effects may be *dynamic* due to phenomena such as rater fatigue, implementation of rater monitoring procedures, and learning that occurs after the commencement of a rating project (Congdon & McQueen, 1997; Wolfe, Moulder, & Myford, 2001). An interesting and challenging task for future research would be to extend the procedures introduced in this article to the examination of pratice/fatigue effects (i.e., differential accuracy/inaccuracy over time), recency/primacy (i.e., differential harshness/leniency over time), and differential centrality/extremism over time.

References

1. Advanced Placement Program. (2003a). English Language and Composition Course Description. Retrieved July 7, 2003, from http://apcentral.collegeboard.com/repository/ap03_cd_english_0203f_4309.pdf
2. Advanced Placement Program. (2003b). Exam Scoring. Retrieved July 7, 2003, from http://apcentral.collegeboard.com/article/0,3045,152-167-0-1994,00.html
3. Braun, H.I. (1988). Understanding score reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 13, 1-18.
4. Breland, H.M., & Jones, R.J. (1984). Perceptions of writing skills. Written Communication, 1, 101-119.
5. Brennan, R.L. (1992). Elements of generalizability theory. Iowa City, IA: American College Testing.
6. Congdon, P., & McQueen, J. (1997). The stability of rater characteristics in large-scale assessment programs. Paper presented at the Ninth International Objective Measurement Workshop, Chicago, IL.
7. de Gruijter, D.N. (1984). Two simple models for rater effects. Applied Psychological Measurement, 8, 213-218.
8. Dean, M.L. (1980). Presentation order effects in product taste tests. Journal of Psychology, 105, 107-110.
9. Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5, 171-191.
10. 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.
11. Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. Journal of Educational Measurement, 33, 56-70.
12. Freedman, S.W. (1979). How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology, 71, 328-338.
13. Freedman, S.W., & Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor & S.A. Walmsley (Eds.), Research on Writing: Principles and methods (pp. 75-98). New York: NY: Longman.
14. Glaser, R., & Chi, M.T.H. (1988). Overview. In M.T.H. Chi, R. Glaser & M.J. Farr (Eds.), The nature of expertise (pp. xv-xxviii). Hillsdale, NJ: Lawrence Erlbaum.
15. Houston, W.M., Raymond, M.R., & Svec, J.C. (1991). Adjustments for rater effects in performance assessment. Applied Psychological Measurement, 15, 409-421.
16. Hoyt, W.T. (1999). Magnitude and moderators of bias of observer ratings: A meta-analysis. Psychological Methods, 4, 403-424.
17. Hoyt, W.T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? Psychological Methods, 5, 64-86.
18. Jako, R.A., & Murphy, K.R. (1990). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. Journal of Applied Psychology, 75, 500-505.
19. Linacre, J.M. (1994). Many-Facet Rasch Measurement. Chicago, IL: MESA.
20. Linacre, J.M. (2003). Facets--Rasch measurement computer program (Version 3.42). Chicago, IL: MESA Press.
21. Longford, N.T. (1996). Adjustment for reader rating behavior in the Test of Written English (No. 55). Princeton, NJ: Educational Testing Service.
22. Lunz, M.E., Wright, B.D., & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 3, 331-345.

23. McIntyre, R.M., Smith, D.E., & Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.
24. 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, 494-500.
25. Murphy, K.R., & Balzer, W.K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74, 619-624.
26. 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.
27. Raymond, M.R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. Journal of Educational Measurement, 30, 253-268.
28. Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.
29. Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.
30. Tziner, A., & Murphy, K.R. (1999). Additional evidence of attitudinal influences in performance appraisal. Journal of Business and Psychology, 13, 407-419.
31. Vance, R.J., Winne, P.S., & Wright, E.S. (1983). A longitudinal examination of rater and ratee effects in performance ratings. Personnel Psychology, 36, 609-620.
32. Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 11-125). Norwood, NJ: Ablex.
33. Welch, J.L., & Swift, C.O. (1992). Question order effects in taste testing of beverages. Journal of the Academy of Marketing Science, 20, 265-268.
34. Wolfe, E.W. (1997). The Relationship between essay reading style and scoring proficiency in a psychometric scoring system. Assessing Writing, 4, 83-106.
35. Wolfe, E.W. (1998). Criterion-referenced rater monitoring through optimal appropriateness measurement. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
36. 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). Stamford, CT: Ablex.
37. Wolfe, E.W., Kao, C.W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. Written Communication, 15, 465-492.
38. Wolfe, E.W., Moulder, B.C., & Myford, C.M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. Journal of Applied Measurement, 2, 256-280.
39. Wright, B.D., & Masters, G.N. (1982). Rating scale analysis: Rasch measurement. Chicago, IL: MESA.
40. Yu, J., & Murphy, K.R. (1993). Modesty bias in self-ratings of performance: A test of the cultural relativity hypothesis. Personnel Psychology, 46, 357-363.

Endnotes

[1] It is important to note that, in this article, all rater effects are defined in a normative sense. That is, the frame of reference for interpreting rater effects is the sample of raters. The

term “known to be valid” is used here to acknowledge that a different frame of reference could be chosen (e.g., one that depicts rater effects relative to an absolute criterion, such as expert ratings that contain no error).

[2] In this article, the term *object of measurement* has been adopted to refer to the elements that are the focus of interest. In achievement testing, those elements are typically examinees. In industrial settings, those elements are typically candidates.

[3] Other multifaceted Rasch models include the more restricted dichotomous model and the more generalized partial credit model. Most of the analyses described in this article can be performed in the context of these other models.

[4] Because of scale indeterminacy, the mean and standard deviation of one distribution of parameter estimates must be set to equal zero. Most Rasch parameter estimation software allows the user to specify which of these distributions serves as the “anchor” for the remaining distributions.

[5] It is important to note that this example was chosen for illustration purposes only. These results are not comparable to what would be observed in the AP passing rates, which are set by individual institutions, because the essay portion of the *English Literature and Composition* test is only one of several components of the composite score.