DOCUMENT RESUME

ED 408 324 TM 026 577

AUTHOR Wolfe, Edward W.; Chiu, Chris W. T.
TITLE Detecting Rater Effects with a Multi-Faceted Rating Scale Model.
PUB DATE Mar 97
NOTE 37p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Item Response Theory; Mathematical Models; Norms; *Performance Based Assessment; *Rating Scales; Scaling; Simulation
IDENTIFIERS Accuracy; Calibration; Large Scale Assessment; *Rasch Model; *Rater Effects

ABSTRACT

How common patterns of rater errors may be detected in a large-scale performance assessment setting is discussed. Common rater effects are identified, and a scaling method that can be used to detect them in operational data sets is presented. Simulated data sets are generated to exhibit each of these rater effects. The three continua that depict the most commonly cited rater effects are: (1) accuracy/randomness; (2) harshness/leniency; and (3) centrality/extremism. Rasch measurement theory provides one way of examining these rater effects within a normative framework. Rasch measurement places each facet of the measurement context on a common underlying linear scale, resulting in measures that can be subjected to traditional statistical analyses while allowing for unambiguous substantive interpretations of the meaning of examinee performance as it relates to rater performance and task functioning. In addition, Rasch calibrations of examinees, tasks, and raters are sample free in that they remove the influence of sample variability. The Multi-Faceted Rating Scale Model (MFRSM) of J. M. Linacre (1989) was used with simulated datasets that illustrate rater effects. Rater effects could be detected in the normative framework through MFRSM, and these effects seemed to operate on several continua. Further research is needed to determine how large a departure from the pool of raters needs to be before it can be detected in a normative framework. (Contains 5 figures, 9 tables, and 13 references.) (SLD)

Running Head: DETECTING RATER EFFECTS

Detecting Rater Effects with a Multi-Faceted Rating Scale Model

Edward W. Wolfe

Educational Testing Service, Princeton, New Jersey

Chris W.T. Chiu

Michigan State University

Author Notes

Edward W. Wolfe, Center for Performance Assessment; Chris W.T. Chiu, Measurement and Quantitative Methods.

Portions of this research was supported by a post-doctoral fellowship at Educational Testing Service. This manuscript was presented at the Annual Meeting of the National Council on Measurement in Education, March, 1997 in Chicago, Illinois.

Correspondence concerning this article should be addressed to Edward W. Wolfe, Educational Testing Service, Mail Stop 11-P, Princeton, New Jersey, 08541-0001. Electronic mail may be sent via Internet to ewolfe@ets.org.

Abstract

Raters may influence examinee scores in a number ways when judgments are made about examinee responses to open-ended assessment tasks. In this paper, we discuss how a variety of rater effects can be detected with a multi-faceted rating scale model. We describe several common rater effects. When these effects are portrayed in a normative context, the extent to which individual raters differ from the entire pool of raters is of concern. To this end, we demonstrate the development of such a normative framework for examining rater effects through a series of ratings simulations. We also identify several interesting outcomes of our analyses and suggest directions for future research concerning rater effects.

Detecting Rater Effects with a Multi-Faceted Rasch Model

As performance-based and portfolio assessments become more and more popular as a means of linking large-scale educational testing to classroom instruction, more attention has been directed toward understanding how the use of raters influences the validity and reliability of test scores. Raters may introduce error into examinee scores for a variety of reasons--unfamiliarity with or inadequate training toward the rating scale, fatigue or lapses in attention, deficiencies in some areas of content knowledge, or personal beliefs that conflict with the values espoused by the scoring rubric. In any case, rater errors result in patterns of ratings that can be used to identify specific problems with a particular rater. For example, raters who make errors because of fatigue are likely to make more random errors as time progresses. On the other hand, raters who are unable to differentiate the number of categories contained in the rating scale are likely to assign a disproportionate number of scores in the middle of the rating scale.

The purpose of this study is to identify how common patterns of rater errors may be detected in a large-scale performance assessment setting. In the following sections, we identify several rater effects and describe a scaling method that can be used to detect these effects in operational data sets. We also present the results of analyses of several simulated data sets that are generated to exhibit each of these rater effects. Finally, we summarize the results across these various types of effects and propose further studies of rater effects.

Rater Effects

Previous research concerning rater effects has identified several ways that raters may introduce error into examinee scores (Saal, Downey, & Lahey, 1980, provide a good review of much of this work). Our work focuses on three continua that depict the most commonly-cited of these effects. In the sections that follow we describe how rater accuracy/randomness, harshness/leniency, and centrality/extremism manifest themselves in the ratings assigned by raters. We also describe a multi-faceted rating scale model that can

be used to analyze and detect various aberrant patterns in rating data. First, however, we describe the framework within which rater effects will be examined in our study.

As is true for all measurement contexts, the information provided by a "score" cannot be interpreted without a referent against which to compare that score. If we were to tell you that an examinee was assigned a score of "5" on a writing assessment, you have no way of interpreting the information that we provided to you. To do so, you would need answers to a few questions. What is the lowest and highest possible score on the assessment? What is the average examinee score? What are the characteristics of a piece of writing that is worthy of a "5"? To interpret the examinee's score, you need a framework for interpreting the meaning of the examinee's score. The same is true when it comes to examining rater effects.

In order to determine whether raters are unduly influencing examinee scores, we need a framework for examining rater effects. Currently, there are two such frameworks. In a normative framework, the more common of the two, rater effects are examined in the context of the pool of raters from which individual raters are drawn. Hence, a normative framework for examining rater effects describes how much individual raters differ from the "average" rater in the pool. As a result, the normative framework can also be referred to as an agreement framework because we are concerned with how well the ratings of individual raters agree with the ratings assigned by all of the other raters in the pool. The second framework is criterion-referenced in nature. That is, we can examine rater effects in the context of some external point of reference that is assumed to be a valid indicator of the examinee's proficiency. These externally-generated scores are most commonly assigned by a benchmark committee, but they may be determined based on examinee scores on some other assessment instrument. We can refer to the criterion-referenced framework as one that depicts rater errors rather than effects because we are examining the accuracy of a rater's ratings rather than simply the agreement of those ratings with ratings assigned by other raters. In this study, we employ a normative framework for describing rater effects.

However, the effects that we describe in the following sections may appear in either framework.

Accuracy/Randomness

A common concern of those who interpret ratings is the extent to which raters make random errors. We represent this concern with an accuracy/randomness continuum. On the extreme side of accuracy, the ratings assigned by a rater never contain error. These ratings are always accurate representations of the examinee's proficiency. On the other end of the continuum, a rater's ratings are not representative of the examinee's proficiency at all. In fact, the scores are simply random numbers. When raters differ in their position on the accuracy/randomness continuum, we cannot be sure how representative any individual rating is of the examinee's proficiency. The accuracy/randomness continuum can be represented graphically as the difference between a particular rater's estimate of an examinee's proficiency (β_r) and that examinee's actual proficiency (β). This relationship is depicted in Figure 1. When a rater is always accurate, we would expect the difference between the rater's rating and the examinee's actual proficiency (β_r-β) to be zero and to fall on the center line in Figure 1 regardless of the examinee's level of proficiency. On the other hand, we would expect $|\beta_r-\beta|$ to be quite large (sometimes positive and sometimes negative) across the range of examinee proficiency for a random rater (as depicted by the darkened band in Figure 1). Most raters, however, will fall somewhere between these two extremes, and a "reasonable" amount of random variation will be established for evaluating the performance of individual raters (as shown by the dashed lines above and below the line representing perfect accuracy).

Harshness/Leniency

We represent the most commonly-investigated rater effects on a continuum labeled harshness/leniency. This continuum is concerned with whether some raters are assigning

systematically higher or lower ratings than are other raters. If a rater exhibits harshness, then the ratings assigned by that rater will tend to underestimate the examinee's proficiency across the proficiency continuum. In Figure 2, harshness is represented as a shift toward the negative end of the y-axis in the range of ratings assigned by a rater (i.e., β_r-β is a negative number). On the other hand, if a rater exhibits leniency, then the ratings assigned by that rater will tend to overestimate the examinee's proficiency across the proficiency continuum (shown in Figure 3 as a shift toward the positive end of the y-axis). When raters differ in their positions on the harshness/leniency continuum, we can never be sure whether the examinee's score is a function of examinee proficiency or rater character.

Centrality/Extremism

We represent another common rater effect on a continuum labeled centrality/extremism. This continuum concerns the extent to which raters are under- or over-using the categories contained in the rating scale. If a rater exhibits centrality, then the ratings assigned by that rater tend to cluster in the center of the rating scale. Figure 4 shows that centrality effects result in accurate rating in the central range of the ability continuum, but over-estimates of examinee proficiency for non-proficient examinees (i.e., β_r-β is positive) and under-estimates of examinee proficiency for highly proficient examinees (i.e., β_r-β is negative). Extremism occurs when raters tend to cluster ratings in the extreme categories of the rating scale. As shown in Figure 5, this results in accurate rating in the tails of the proficiency distribution, but large amounts of error associated with the ratings assigned to examinees with middling levels of proficiency (who are forced into the tails of the rating scale). When raters differ in their positions on the centrality/extremism continuum, we can never be sure when the ratings assigned by a particular rater will be accurate and when they will contain error.

Rasch Measurement

Rasch measurement theory provides one way of examining the rater effects described above within a normative framework. Rasch measurement is a latent trait modeling technique that has proven useful for solving a variety of measurement problems. Applying Rasch measurement to the analysis of rating data results in several beneficial conditions. First, Rasch measurement places each facet of the measurement context (e.g., examinees, tasks, and raters) on a common underlying linear scale. This results in measures that can be subjected to traditional statistical analyses while allowing for unambiguous substantive interpretations of the meaning of examinee performance as it relates to rater performance and task functioning. Second, the Rasch calibrations of examinees, tasks, and raters are sample-free. That is, procedures used to estimate examinee proficiency, task difficulty, and rater harshness remove the influence of sampling variability from scaled scores so that valid generalizations can be made beyond the current sample of examinees, collection of tasks, or pool of raters. In applied settings, this feature is useful because it allows an examinee's proficiency to be determined even if that examinee does not respond to all of the assessment tasks or if that examinee is rated by only a portion of the raters in the pool. Third, Rasch procedures can be used to derive expected response patterns that are useful for evaluating the extent to which individual examinees, tasks, or raters are behaving in ways that are inconsistent with the measurement model. As a result, the suitability of the model for the measurement context, as well as the validity of the measures of individuals, can be evaluated by examining the fit between the observed data and the expected response patterns.

Multi-Faceted Rating Scale Model

The Multi-Faceted Rating Scale Model (MFRSM) (Linacre, 1989a) describes the probability that a specific examinee (n) will be rated with a specific rating scale step (x) by a specific rater (k) on a specific task (i). The mathematical form of this probability (Equation 1) depicts the relationship between these elements in terms of a logistic odds ratio (logit).

This probability depends on four parameters: the examinee's proficiency (β_n), the rater's harshness (λ_k), the task's difficulty (δ_i), and the difficulty of each scale step (i.e., the threshold between two adjacent rating scale levels, τ_x). Calibration of rating data produces separate parameter estimate and a standard error for that estimate for each examinee, rater, task, and scale step in the measurement context.

$$P(x|\beta,\lambda,\delta,\tau) = \frac{\exp\sum_{j=0}^{x}\left[\beta_n - \lambda_k - \delta_i - \tau_j\right]}{\sum_{x=0}^{m}\exp\sum_{j=0}^{x}\left[\beta_n - \lambda_k - \delta_i - \tau_j\right]}, \qquad x=0,1,\ldots,m \qquad (1)$$

where, $P(x|\beta,\lambda,\delta,\tau)$ is the probability that the response of examinee n to task i is assigned rating scale category x by rater k when the has m+1 rating options.

This model assumes that a common rating scale structure applies to each task (i.e., that τ_j is constant across tasks). The model also assumes that the data conform to the predictions of the MFRSM. Departures in the data from model-generated expected values indicate potentially misfitting examinees, raters, or tasks. One can examine the residuals of the observed data from the model's predictions to identify individual ratings that are unexpected. To determine whether there is a disproportionate number of unexpected ratings associated with a particular examinee, rater, or task, the empirical percent of unexpected observations associated with each particular element is calculated. If this percentage is considerably larger than would be attributable to chance, then it is likely that the element in question (e.g., a rater) does not conform to the model.

To further evaluate the degree to which the response patterns associated with individual elements of the measurement context (e.g., individual examinees, raters, or tasks) are inconsistent with the MFRSM, two fit statistics are generated for each parameter

estimate (Wright & Masters, 1982). Both of these fit statistics are based on the mean of the squared standardized residuals of the observed scores from their expected scores. The outfit statistic is simply the mean of these standardized residuals. Outfit statistics are sensitive to departures in the data in the extreme scoring categories. The infit statistic, on the other hand, weights each standardized residual by its variance. As a result, infit statistics are more sensitive to unexpected responses that fall near the center of the rating scale. The infit and outfit statistics have an expected value of 1.00 and can range from 0.00 to ∞. A 0.1 increase in a fit statistics is associated with a 10% increase in unmodelled error. In general, elements with fit statistic values ranging from 0.6 to 1.5 are considered to show adequate fit to the model (Wright & Linacre, 1994).

Data Simulation and Analysis

Now we describe a series of simulations that demonstrate how each of the rater effects that we previously identified manifests itself in the logits, fit statistics, and percent of unexpected observations that are produced by Facets (a piece of software that performs MFRSM scaling; Linacre, 1989b). For our study, we generated ten replications of six types of datasets, each exhibiting one of six rater effects: (a) comparison, (b) randomness, (c) harshness, (d) leniency, (e) centrality, and (f) extremism. The following sections describe how each of these datasets was operationalized and how the datasets were analyzed using Facets.

Data Generation

Ten data sets were generated with SAS (SAS Institute Inc., 1985) using the following algorithm. Each data set contained 8000 lines of data (1000 examinees $\times$ 4 tasks $\times$ 2 raters), and error was modeled for each component of the measurement design (examinees, tasks, and raters). The magnitude of these error terms was set so that the MFRSM calibrations under the comparison condition were similar to those observed in

operational datasets. Ten datasets were generated so that the influence of sampling variability could be taken into account in our analyses. These ten datasets serve as our comparison data from which we generate an additional 50 datasets according to the algorithms presented in the following sections. In these datasets, the ratings associated with 90% of the raters (control) exhibited no special characteristics. The ratings associated with the remaining raters (effect) exhibited one of the six types of effects.

1. Generate identifiers for 1000 examinees, each responding to 4 assessment tasks, with each response being scored by 2 raters who are randomly selected from a pool of 50.
2. Generate a true score (T_n) for each examinee from a N(0,1) distribution.
3. Generate an error term (E_{nik}) for each examinee-by-task-by-rater combination from a N(0,1) distribution.
4. Generate an task effect terms (I_i) for each examinee-by-task combination for items 1, 2, 3, and 4 from a N(-0.2,1), N(-0.1,1), N(0.1,1), and N(0.2,1) distribution, respectively.
5. Generate a rater character term (RC_k) for each examinee-by-rater-by-task combination for 90% of the raters (control raters). Let 20% of the control raters be assigned a rater character from a N(-0.2,0.75) distribution. Let another 60% of the control raters be assigned a rater character from a N(0,0.75) distribution. And, let the remaining 20% of the control raters be assigned a rater character from a N(0.2,0.75) distribution. Generate a rater effect term (RE_k) for the remaining 10% of the raters (effect raters) based on the descriptions in the following sections.
6. Compute a total score (TS_{nik}) for each examinee-by-task-by-rater combination for the control raters according to Equation 2.
7. Compute a total score (TS_{nik}) for each examinee-by-task-by-rater combination for the effect raters based on the descriptions in the following sections.
8. Transform the true score to a rating (i.e., a six-point integer scale ranging from one to six) based on the rules specified in Table 1.

$$TS_{nik} = T_n + I_i + RC_k + E_{nik} \quad (2)$$

Comparison/Randomness

The comparison data, which is equivalent to the accuracy effect that we described previously, had no special effect added to the effect raters' ratings. That is, RE_k was simply sampled from a N(0,0.75) distribution. For the randomness condition, the variability of the rater effect was increased substantially so that every examinee-by-rater-by-task combination would contain more random error. Hence, we sampled RE_k from a N(0,1.5) distribution for the random effect raters. For both the comparison and the randomness datasets, the true score was computed according to Equation 3. True score to rating conversion was based on the rules shown in Table 1.

$$TS_{nik} = T_n + I_i + RE_k + E_{nik} \quad (3)$$

Harshness/Leniency

The effect rater terms for the harshness and leniency data sets were generated from a N(0,0.75) distribution. To simulate the harsh/lenient effects, a constant (1.00) was either subtracted from (harshness) or added to (leniency) the total score for the effect raters. Equation 4 shows how the true score was computed for the harshness data. True score to rating conversion was based on the rules shown in Table 1.

$$TS_{nik} = T_n + I_i + RE_k + E_{nik} - 1 \quad (4)$$

Centrality/Extremism

The effect rater terms for the centrality and extremism data sets were generated from a N(0,0.75) distribution. To simulate the centrality/extremism effects, the total score for the effect raters was either divided (centrality) or multiplied (extremism) by a constant of 2.00 (resulting in either one-half or twice the error attributable to raters). Equation 5 shows how

the true score was computed for the centrality data. True score to rating conversion was based on the rules shown in Table 1.

$$TS_{nik} = \frac{T_n + I_i + RE_k + E_{nik}}{2} \quad (5)$$

Rating Scale Analyses

These procedures resulted in 60 datasets--10 for each the 6 conditions. Because the datasets for each condition corresponded to a common comparison dataset, we were able to directly identify how the addition of rater effects influences the ratings associated with specific raters. To this end, we compared the rater facet parameters of the five effect raters in each data set to the parameters of five control raters chosen to be matched-pairs based on comparison condition statistics. Each of the 60 datasets were scaled using Facets, multi-faceted Rasch scaling software (Linacre, 1989b). For each data set, we defined a MFRSM that contained four facets: (a) examinee, (b) task, (c) rater, and (d) scale step. The examinee facet was scaled so that higher values of examinee logits were associated with higher scores. The task, rater, and scale step facets were oriented so that lower scores were associated with higher logit values. The task, rater, and scale step facets were centered at zero, and the examinee facet was non-centered. Analyses focused on only the rater facet. We examined the average logit values, the infit and outfit statistics, and the percent unexpected observations for the five control and five effect raters in each dataset.

Based on previous research concerning rater effects, we expected the rater effects in our simulated data to manifest themselves in the following ways. Randomness was expected to result in larger fit statistics for effect raters because the introduction of noise in a particular response pattern will increase the amount of unmodelled error across the range of the underlying scale (Smith, 1996; Wright, 1991). Harshness and leniency were expected to increase and decrease the mean logit for effect raters, respectively (Engelhard, 1994; Lunz, Wright, & Linacre, 1990). Centrality results in a clustering of ratings about

the midpoint of the rating scale, and these muted rating vectors were expected to result in smaller fit statistics for effect raters (Engelhard, 1994). Although previous work has not investigated the extremism effect, it seems likely that an increase of error variance associated with ratings in the center of the distribution would result in an increase in infit statistics, the fit statistics that are more sensitive to unmodelled variance in the center of the scoring scale.

Results

Comparison/Randomness

Table 2 summarizes the parameter estimates for the rater facet obtained under the comparison condition for the five control and five effect raters. These figures are the means and standard deviations of the average rater parameter values across the ten replicated datasets for the comparison condition. That is, we first averaged the logits, infit statistics, outfit statistics, and percentages of unexpected responses for the five control and five effect raters in each of the ten comparison datasets. Then we computed the average and the standard deviation of these means across the ten datasets. Note that the average logit for both groups is around 0.00 as would be expected, and the average fit statistic is near 1.00. The fact that the average fit statistic is slightly less than 1.00 indicates that the rater vectors are slightly less stochastic than would be predicted by the MFRSM. Also, note that the average percent of unexpected scores associated with both groups is around 11%. The statistics shown in Table 2 are used as benchmarks for evaluating how the addition of various types of rater effects to each dataset influences the Facets output.

Table 3 summarizes the parameter estimates for the rater facet obtained under the randomness condition for the five control and five effect raters. Note that the average logit for both groups is close to its expectation of 0.00. For the control group, the fit statistics are also close to their expected values of 1.00. However, the percent of unexpected observations is somewhat lower than its value in the comparison condition. This spill-over

of rater effects from the raters exhibiting the effect to raters who do not exhibit the effect is common in the normative framework that we use in this example. For the effect group, the addition of randomness to a rater's ratings increases both the infit and outfit statistics substantially. There is about 27% more unmodelled variance (error) in these raters' ratings. As a result, the percent of unexpected observations associated with the effect raters is much larger (23.80%) than the comparison benchmark of 11%.

Harshness/Leniency

Table 4 summarizes the parameter estimates for the rater facet obtained under the harshness condition for the control and effect raters. As was true for the randomness condition, the harshness effect influences the rater parameters for both the control and the effect raters. The average logit for the control raters is slightly lower than the expectation of 0.00. That is, the addition of harshness to the effect raters made the control raters seem lenient. However, the fit statistics and percent of unexpected observations associated with control raters are similar to the expected values. The average logit and percent of unexpected observations for the effect raters are both higher than one would expect. The fit statistics, on the other hand, do not seem to be influenced.

The addition of leniency to effect raters' scores has the opposite effect of the addition of rater harshness. Table 5 summarizes the parameter estimates for the rater facet obtained under the leniency condition for the control and effect raters. As one would expect, control raters show a slight increase (i.e., appear to be slightly more harsh) from their expectation when the effect raters exhibit lenient scoring. On the other hand, effect raters show a decrease in their average logits and an increase in the percent of unexpected observations with which these raters are associated. Fit statistics are close to the expectation of 1.00 for both groups.

Centrality/Extremity

Table 6 summarizes the parameter estimates for the rater facet obtained under the centrality condition for the control and effect raters. Based on an examination of the fit

statistics, the ratings associated with the effect raters seem to exhibit less stochasticity than do the ratings of the control raters. Although the control raters have fit statistics near the expected value of 1.00, effect raters exhibit about 15% less unmodelled stochasticity in their ratings. Only by examining the percent of unexpected observations for each group, can we identify that a problem exists. The percent of unexpected observations for the control raters is close to the comparison benchmark of 11%. However, the percentage for the effect raters is much larger (21.90%), which is contrary to what we might predict based on the fit statistics for these raters. This illustrates a danger of interpreting fit statistics without examining the percent of unexpected observations with which individual raters are associated. We would come to very different conclusions about the quality of the effect raters' ratings if we were to examine both the fit statistics and the percent of unexpected observations versus examining only the fit statistics.

Based on the trends in the previous examples, one would expect a spill-over effect in under the extremism condition as well as observing a fit statistics that are contrary to those observed in the centrality data sets. Table 7, which summarizes the parameter estimates for the rater facet obtained under the extremism condition for the control and effect raters, verifies this expectation. Note that the average logit for each group of raters is close to the expected value and that the fit statistics and percent of unexpected responses for the control group are also close to their expectations. On the other hand, the fit statistics for the effect group are somewhat larger than expected. While the outfit statistic indicates about 16% unmodelled variance in the ratings of the effect raters, the infit statistic indicates about 31% unmodelled variance. This makes sense. Because extreme scoring results in more error in the middle of the rating scale than in the tails of the distribution, we would expect the infit statistic (which is more sensitive in the center of the distribution) to be influenced by the extremism effect. The outfit statistic, which is more sensitive to unexpected ratings falling in the tails of the distribution, is not influenced as greatly by the extremism effect.

As was true for the centrality effect, extreme rating results in a percentage of unexpected observations for the effect raters that is higher than the comparison benchmark of 11%.

Across Effect Comparisons

Table 8 summarizes the influence that each type of rater effect had on the parameter estimates for the rater facet of the control and effect raters. As shown, rater harshness, leniency, and centrality produce unique patterns in these parameter estimates when compared to the comparison benchmarks. Harshness manifests itself as an increase in the mean logit of effect raters and a slight decrease in the mean logit of control raters. The same pattern is true for the percent of unexpected observations associated with these groups--the percent for the control raters decreased slightly while the percent for the effect raters increased slightly. A converse pattern is observed in the logits for under leniency condition. Lenient effect raters had a rather large decrease in their mean logits, while control raters showed a slight increase in their mean logits. The changes in the percent of unexpected observations for each group are similar to, and in the same directions as, those observed under the harshness condition.

The centrality effect manifests itself in a somewhat different manner. For these datasets, there were no large changes in any of the rater facet estimates for the control group. And, although the mean logits for the effect group did not change, a unique pattern arises in their fit statistics and percent of unexpected observations. For these raters, both the infit and outfit statistics are smaller (i.e., show about 15% less unmodelled variance) than they were under the comparison condition. In addition, the percent of unexpected observations with which centrality effect raters were associated is considerably larger than that of the comparison condition. Interestingly, the variability of the percent of unexpected observations for the effect raters also increased under the centrality effect. That is, the stability of the percent of unexpected ratings increased under the centrality condition.

The rater facet summaries for the randomness and extremism datasets do not show such distinctive patterns. For both of these effects, there is little change in the mean logits

or fit statistics for the control raters, and there is little change in the mean logits for the effect raters. Also, both random and extreme raters show an increase in both the fit statistics and the percent of unexpected observations for effect raters. Closer examination of these figures reveals subtle differences between the random and extreme raters. Recall that the infit and outfit statistics for the random raters are approximately equal. This means that similar amounts of unmodelled error are found in both inlying and outlying scores. On the other hand, the infit statistics are somewhat larger than the outfit statistics for the extreme raters, indicating that there is more unmodelled error in inlying scores than there is in outlying scores. There is also more variability in the fit statistics for extreme scorers than for random scorers. That is, random raters' fit statistics are more consistent across different samples while extreme raters' fit statistics are somewhat variable.

Table 9 shows what these effects mean in terms of ratings assigned by a particular rater. This table shows the ratings assigned by a single rater to twelve examinees (chosen to be representative of the range of ratings assigned by this rater) under each of the six rater effect conditions. It should be noted that the rater facet parameter estimates shown in Tables 2 through 7 would not result from a Facets analysis of data such as these because most of these parameter estimates are highly sensitive to sample size. However, this sample of ratings demonstrates two things. First, the rater facet summary statistics are associated with clearly identifiable patterns in the rating data. For example, the introduction of harshness and leniency clearly resulted in a tendency toward the assignment of higher and lower ratings, respectively. On the other hand, the addition of centrality and extremism effects resulted in a greater and smaller proportion of ratings falling in the central categories of the rating scale (e.g., "4" and "5"). And, second, the rater parameter summary statistics that we discussed in this paper may be fairly sensitive to even small departures from modeled expectations that can be attributed to rater characteristics. As shown by the summary statistics for the ratings, even small rating deviations from the comparison mean can be indicated by increases or decreases in the rater logit values, and small rating deviations

from the comparison standard deviation can be indicated through increases or decreases in the rater fit statistics.

Conclusions

We draw the following conclusions from the results of our simulations.

1. Rater effects can be detected in a normative framework by examining MFRSM rater calibration logit values, fit statistics, and the percent of unexpected observations associated with individual raters, and the magnitude of the values for an individual rater correspond to the extent to which that rater's ratings differ from the ratings assigned by the remaining raters in the pool.
2. The rater effects that we investigated seem to operate on a several continua as evidenced by the predictability with which affected indicators changed. That is, randomness seems to increase the infit and outfit statistics and the percent of unexpected observations while accuracy seems to reduce these statistics. Harshness and leniency seem to raise and lower the value of the logit, respectively. And, centrality and extremism seem to increase and lower the value of the fit statistics, respectively, while both these effects seem to increase the percent of unexpected observations.
3. Examining rater effects in a normative framework is possible, but the information provided by such analyses is ambiguous. As evidenced by the spill-over effects we observed in several of our analyses, the only question we can answer is "Which raters look suspect?" rather than "Which raters are incorrect?". For example, when we observe evidence of harshness effects in a normative framework, we do not know whether the raters who have large logit values are indeed harsh or whether the remaining raters are rating leniently.
4. It is important to examine all of the information available about raters prior to drawing conclusions about their rating behaviors. As demonstrated by the results of our centrality effect analyses, examination of fit statistics can be misleading in the absence

of information about the percent of unexpected observations associated with individual raters.

5. Still, some of the information about rater effects may be confusing. We find it odd that there were no differences between the infit and outfit statistics for the centrality effect even though the infit statistics for the extremism effect was larger, as one might predict based on the nature of that effect. Perhaps this phenomena occurred because our data were designed to approximate other performance assessment datasets that we have analyzed. Most of these datasets have a pronounced leptokurtic (i.e., peaked) shape. This peakedness may cause the outfit statistic (which is sensitive to error variance in the tails of the score distribution) to show less change under the centrality effect because fewer cases are influencing that statistic.

Future Research

We believe that further research concerning the examination of rater effects with the multi-faceted rating scale model should focus on the following issues.

1. Power analyses should be done to determine how large a departure from the pool of raters needs to be before it can be detected in a normative framework. These power analyses should take into account both the proportion of raters exhibiting the effect and the size of the effect.
2. This methodology should be extended so that other types of rater effects can be examined using the MFRSM. For example, Engelhard (1994) considered halo effects where raters fail to independently rate examinee performance across multiple test items. Wolfe and Myford (1997, March) suggests that rater effects may manifest themselves as a function of time (e.g., fatigue, practice, recency, or primacy).
3. Because the spill-over effects that we observed are unavoidable in a normative framework and because these spill-over effects introduce ambiguity into the interpretation of rater effects, further work should be done to develop models for examining rater errors in a criterion-referenced framework. Although Engelhard (1996)

proposed one model for detecting rater errors in a criterion-referenced context, this model has seen only limited application in operational settings.

References

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

Engelhard, G.J. (1996). Evaluating rater accuracy in performance assessments. Journal of Educational Measurement, 33, 56-70.

Linacre, J.M. (1989a). Many-facet Rasch measurement. Chicago, IL: MESA Press.

Linacre, J.M. (1989b). A user's guide to Facets: Rasch measurement computer program. Chicago, IL: MESA Press.

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.

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.

SAS Institute Inc. (1985). SAS user's guide: Basics. Cary, NC: Author.

Smith, R.M. (1996). Polytomous mean-square fit statistics. Rasch Measurement Transactions, 10, 516-517.

Wolfe, E.W., & Myford, C.M. (1997, March). Detecting Order Effects with a Multi-Faceted Rasch Model. Manuscript presented at Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Wolfe, E.W. (1996). Unbeknownst to the right honorable judge—Or how common judging errors creep into organized beer evaluations. Brewing Techniques, 4(2), 56-59.

Wright, B.D. (1991). Diagnosing misfit. Rasch Measurement Transactions, 5, 156.

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

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

Table 1

Rules for Transforming True Scores to Ratings

True Score	Ratings
$\underline{TS}_{nik} < -3.5$	1
$-3.5 \leq \underline{TS}_{nik} < -2.0$	2
$-2.0 \leq \underline{TS}_{nik} < 0.0$	3
$0.0 \leq \underline{TS}_{nik} < 2.0$	4
$2.0 \leq \underline{TS}_{nik} < 3.5$	5
$\underline{TS}_{nik} \leq 3.5$	6

Table 2

Rater Facet Summary Under Comparison Condition

Parameter	Control Raters	Effect Raters
Logit	0.01	0.00
	(0.04)	(0.04)
Infit	0.96	0.98
	(0.06)	(0.04)
Outfit	0.98	0.98
	(0.03)	(0.04)
Percent Unexpected	12.20	10.30
	(9.96)	(3.06)

Note: $N_{control}$=5. $N_{effects}$=5. The mean parameter estimate (and standard deviation) of the ten replications are shown.

Table 3

Rater Facet Summary Under Randomness Condition

Parameter	Control Raters	Effect Raters
Logit	0.01	0.03
	(0.04)	(0.14)
Infit	0.94	1.27
	(0.06)	(0.04)
Outfit	0.94	1.28
	(0.06)	(0.04)
Percent Unexpected	8.70	23.80
	(8.12)	(5.87)

Note: $N_{control}$=5. $N_{effects}$=5. The mean parameter estimate (and standard deviation) of the ten replications are shown.

Table 4

Rater Facet Summary Under Harshness Condition

Parameter	Control Raters	Effect Raters
Logit	-0.10	1.01
	(0.05)	(0.06)
Infit	0.98	0.96
	(0.03)	(0.06)
Outfit	0.98	0.97
	(0.03)	(0.07)
Percent Unexpected	10.70	12.20
	(9.15)	(3.61)

Note: $N_{control}$=5. $N_{effects}$=5. The mean parameter estimate (and standard deviation) of the ten replications are shown.

Table 5

Rater Facet Summary Under Leniency Condition

Parameter	Control Raters	Effect Raters
Logit	0.12	-1.00
	(0.05)	(0.05)
Infit	0.98	0.98
	(0.03)	(0.03)
Outfit	0.98	0.99
	(0.03)	(0.03)
Percent Unexpected	10.80	11.00
	(8.15)	(4.03)

Note: $N_{control}$=5. $N_{effects}$=5. The mean parameter estimate (and standard deviation) of the ten replications are shown.

Table 6

Rater Facet Summary Under Centrality Condition

Parameter	Control Raters	Effect Raters
Logit	-0.01	-0.01
	(0.06)	(0.05)
Infit	0.99	0.83
	(0.03)	(0.03)
Outfit	0.99	0.84
	(0.03)	(0.03)
Percent Unexpected	11.20	21.90
	(8.04)	(6.62)

Note: $N_{control}$=5. $N_{effects}$=5. The mean parameter estimate (and standard deviation) of the ten replications are shown.

Table 7

Rater Facet Summary Under Extremism Condition

Parameter	Control Raters	Effect Raters
Logit	0.01	0.01
	(0.04)	(0.04)
Infit	0.95	1.31
	(0.03)	(0.17)
Outfit	0.95	1.16
	(0.02)	(0.37)
Percent Unexpected	9.10	21.90
	(8.09)	(6.62)

Note: $N_{control}$=5. $N_{effects}$=5. The mean parameter estimate (and standard deviation) of the ten replications are shown.

Table 8

Rater Facet Summary Across Conditions

Parameter Effect	Harsh	Lenient	Central	Random	Extreme
Logit					
Control	-	+	NC	NC	NC
Effect	++	--	NC	NC	NC
Infit					
Control	NC	NC	NC	NC	NC
Effect	NC	NC	-	++	++
Outfit					
Control	NC	NC	NC	NC	NC
Effect	NC	NC	-	++	+
Unexpected					
Control	NC	NC	NC	-	-
Effect	NC	NC	++	++	++

Note: Each cell shows the trend for the parameter in question across ten replications of that condition for both the control (N=5) and effect (N=5) raters. ++ (and --) indicate an increase (or decrease) of 5 standard deviations in logits or fit statistics or a 5% change in unexpected responses from those obtained under the comparison condition. + (and -) indicate an increase (or decrease) of 2.5 standard deviations in logits or fit statistics or a 5% change in unexpected responses from those obtained under the comparison condition. NC indicates no appreciable change from the comparison estimates.

Table 9

Rater Facet Summary Across Conditions

Case	Compare	Random	Harsh	Lenient	Central	Extreme
1	1	1	1	2	2	1
2	2	2	2	3	2	2
3	2	2	1	3	2	1
4	3	3	2	3	3	3
5	3	3	3	4	3	3
6	3	2	2	3	3	2
7	4	4	3	4	4	4
8	4	4	4	5	4	4
9	4	5	4	5	4	5
10	5	5	4	5	4	5
11	5	6	4	6	5	5
12	6	6	5	6	5	6
Mean	3.50	3.58	2.92	4.08	3.42	3.42
SD	1.45	1.68	1.31	1.31	1.08	1.68

Note: Each column shows the ratings assigned to a set of twelve examinees by a single effect rater across the six rater effect conditions.

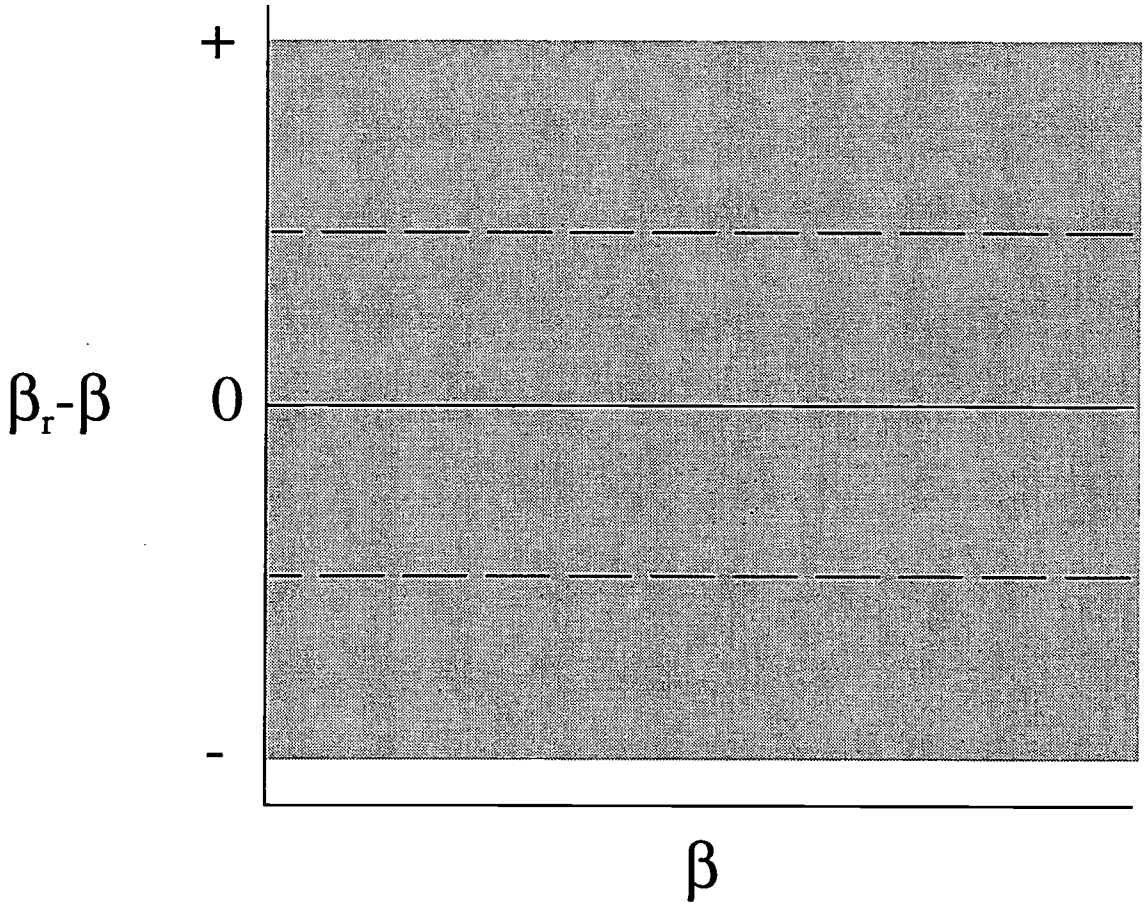
Figure 1. The distribution of rater error (β_r-β) across the proficiency range for accurate (β_r-β =0) and random ($|\beta_r$-$\beta|$ is large) scoring. The solid line represents accurate rating. The darkened band represents random rating. The dashed lines represent a reasonable amount of randomness.

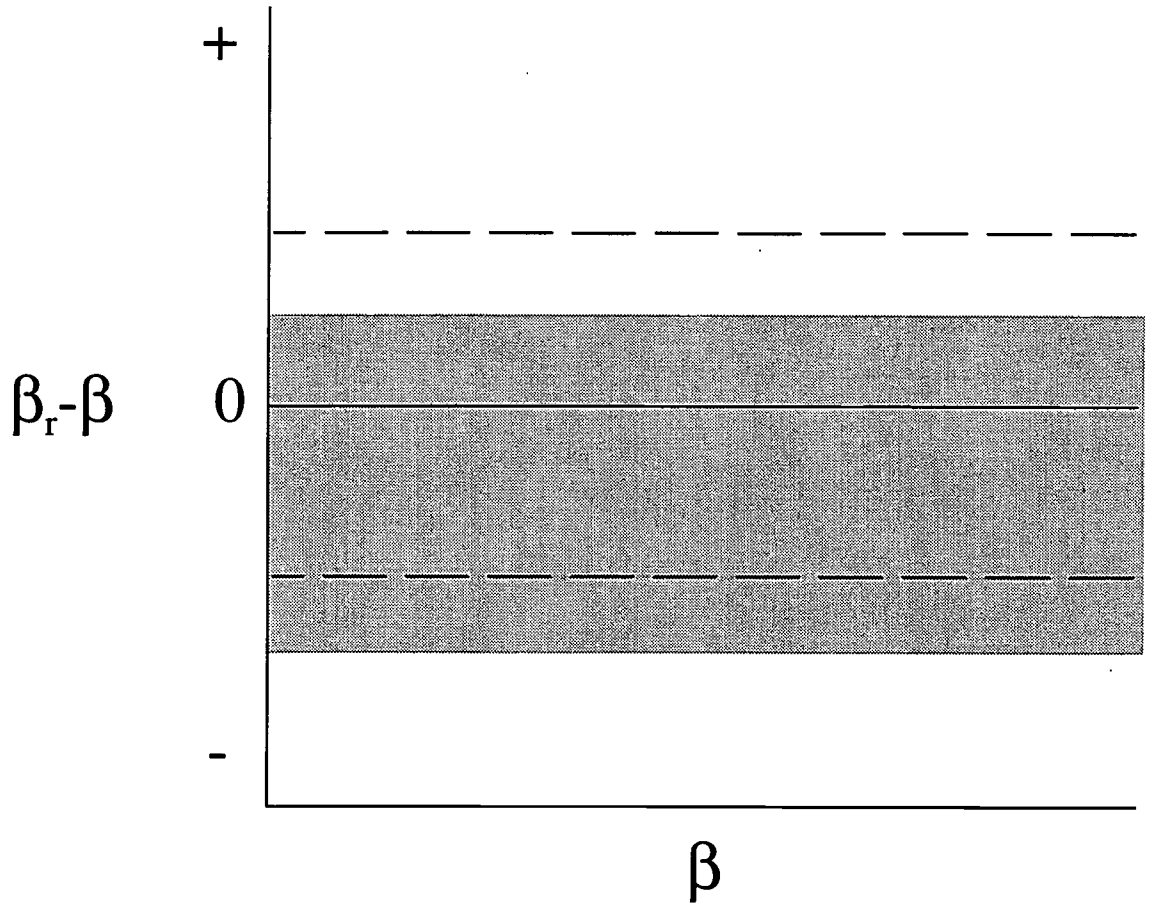
Figure 2. The distribution of rater error (β_r-β) across the proficiency range for harsh (β_r-β is negative) rating. The solid line represents accurate rating. The dashed lines represent a reasonable amount of randomness. The darkened band represents harsh rating.

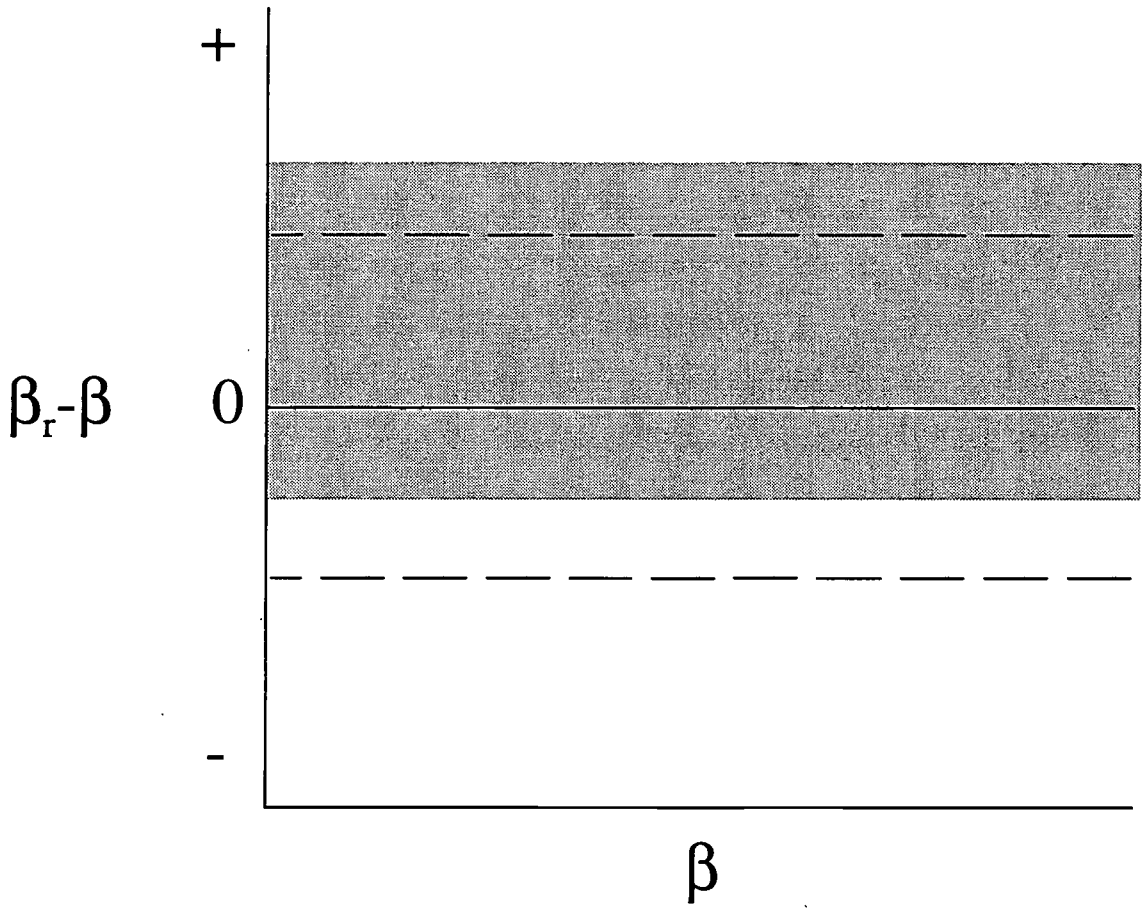
Figure 3. The distribution of rater error (β_r-β) across the proficiency range for lenient (β_r-β is positive) rating. The solid line represents accurate rating. The dashed lines represent a reasonable amount of randomness. The darkened band represents lenient rating.

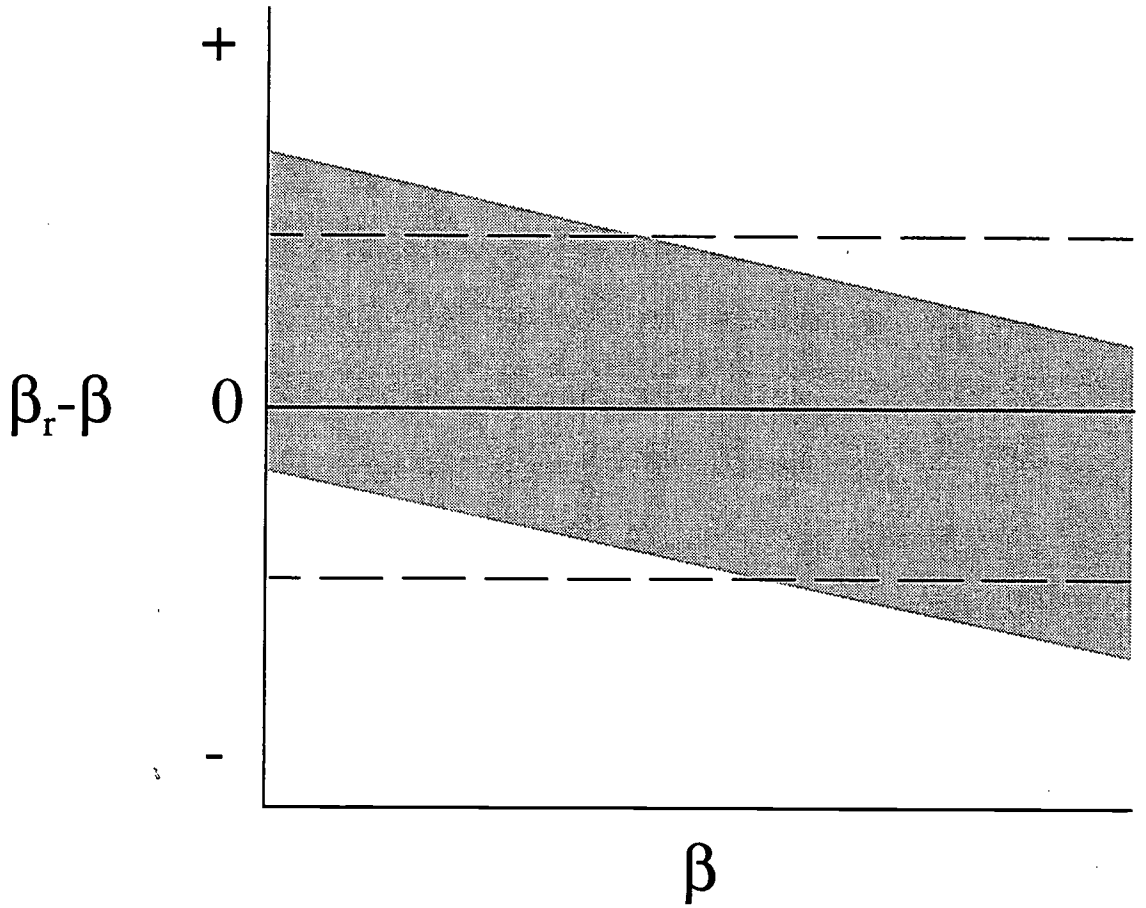
Figure 4. The distribution of rater error (β_r-β) across the proficiency range for central ($|\beta_r$-$\beta|$ is non-zero in the tails of the proficiency distribution) rating. The solid line represents accurate rating. The dashed lines represent a reasonable amount of randomness. The darkened band represents central rating.

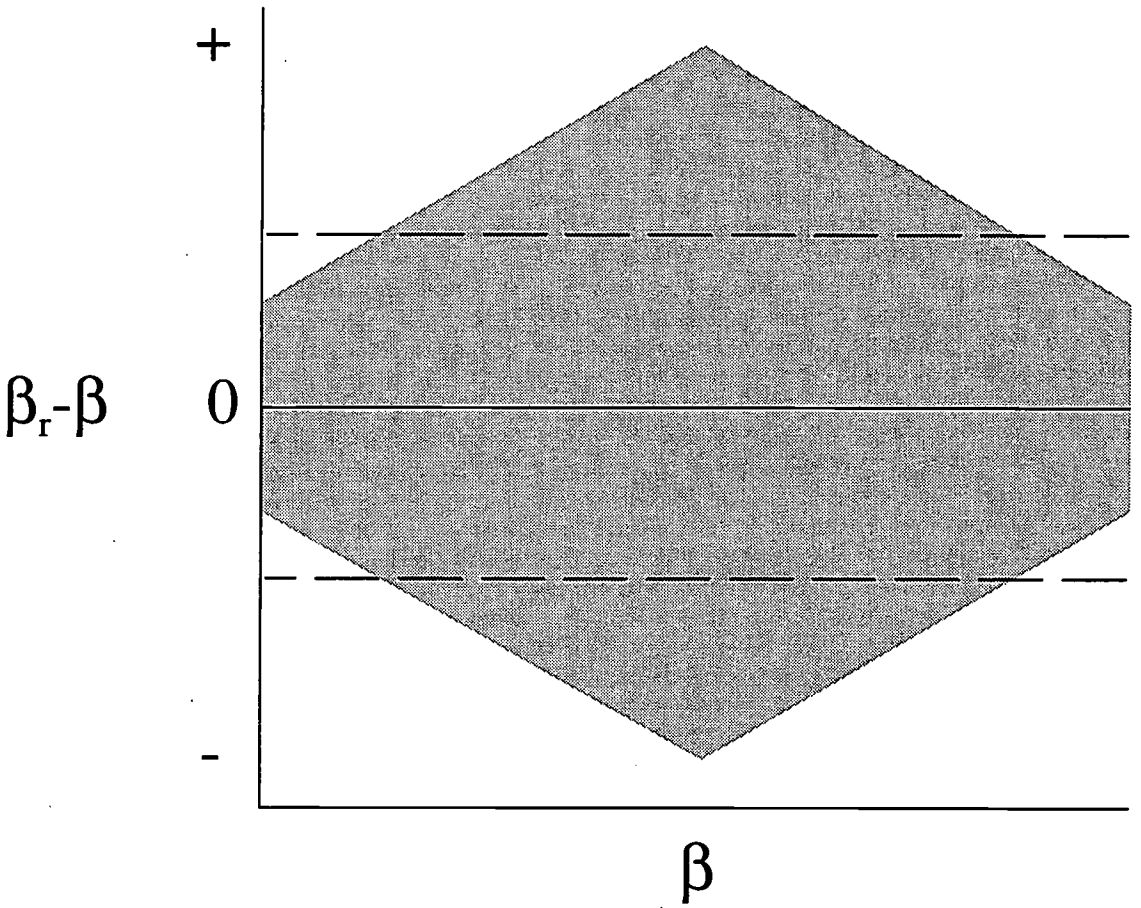
Figure 5. The distribution of rater error (β_r-β) across the proficiency range for extreme ($|\beta_r$-$\beta|$ is large in the center of the proficiency distribution) rating. The solid line represents accurate rating. The dashed lines represent a reasonable amount of randomness. The darkened band represents extreme rating.


+
β_r-β
0
-
β


+
β_r-β
0
-
β


+
β_r-β
0
-
β


+
β_r-β
0
-
β


+
β_r-β
0
-
β

TM026577

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Detecting Rater Effects with a Multi-faceted Rating Scale Model

Author(s): Edward W. Wolfe & Chris W.T. Chiu

Corporate Source: Educational Testing Service

Publication Date: Mar. 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

	The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2 documents	
☒ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	☐ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.
	Level 1	Level 2	

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1.**

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Edward W. Wolfe	Printed Name/Position/Title: Edward W. Wolfe / Post Doctoral Fellow	
Organization/Address: ETS Mailstop 11-P Princeton, NJ 08548	Telephone: 609-734-1855	FAX: 609-734-5115
	E-Mail Address: ewolfe@ets.org	Date: 3/19/97

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the apprcpriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
The Catholic University of America
Washington, DC 20064

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2d Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

(Rev. 6/96)