ACT Research Report Series 90-14

Detecting and Correcting for Rater Effects in Performance Assessment

Mark R. Raymond
Walter M. Houston

December 1990

ACT

DETECTING AND CORRECTING FOR RATER EFFECTS IN PERFORMANCE ASSESSMENT

Mark R. Raymond
Walter M. Houston

American College Testing

An earlier version of this report was presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, April 1990, in Boston, MA.

ABSTRACT

Performance rating systems frequently make use of multiple raters in an effort to improve the reliability of ratings. It is common to use two, three, or four raters in evaluating the clinical performance of health-care trainees, in rating performance on oral and essay examinations, and in evaluating on-the-job performance. However, unless all candidates (i.e., students, employees, examinees) are rated by the same raters, some candidates will be at an unfair advantage or disadvantage due solely to the fact that they were rated by more lenient or more stringent raters. In order to obtain fair and accurate evaluations of candidate performance, such sources of systematic rating error must be taken into consideration. This paper describes four procedures to detect and correct for rater effects. A demonstration of each procedure is also provided, using a small set of rating data to illustrate the impact of the procedures. The results of the demonstration, which are consistent with the findings of other published research, indicate that each of the four procedures produces more accurate estimates of true levels of performance than the traditional approach of summing the observed ratings. The results encourage further research on the utility of methods to correct for rater effects in performance assessment.

DETECTING AND CORRECTING FOR RATER EFFECTS IN PERFORMANCE ASSESSMENT

Obtaining accurate and reliable performance ratings is a challenge faced in most educational and employment settings. The reliability of ratings given in practical settings are typically quite low, ranging from about .30 to .75 (e.g., Cason & Cason, 1984; King, Schmidt, & Hunter, 1980; Muzzin & Hart, 1985; Rothstein, 1990). Even though there has been a constant dissatisfaction with performance ratings, they remain the most widely used method for evaluating work performance (Landy & Farr, 1980). Ratings of performance are frequently used as the sole basis for making extremely important personnel decisions--decisions related to salary increases, promotability (Landy & Farr, 1980), and whether one is fit for practice in professions such as medicine or psychology (Crawford, 1984; Muzzin & Hart, 1985). It can certainly be argued that the reliability of performance ratings is generally inadequate for making decisions as important as these.

The most direct way to address the problem of low reliability is to obtain ratings from multiple raters (Crocker & Algina, 1986; Landy & Farr, 1983; Stanley, 1961). According to psychometric theory, if the reliability of a single rating is .50, then the reliability of two, four, and six ratings will be approximately .67, .80, and .86, respectively. This result can be obtained through classical measurement theory using the well-known Spearman-Brown formula (e.g., Lord & Novick, 1968) or through generalizability theory (e.g., Brennan, 1983; Shavelson, Webb, & Rowley, 1989). The practice of using multiple evaluators to improve reliability is analogous to constructing tests and surveys that consist of multiple questions.

Although the use of multiple raters generally improves reliability by reducing the relative magnitude of random error, it does not eliminate the

type of error that may arise when candidates within a group are evaluated by different raters. Unless the same raters evaluate all candidates, there is the possibility that some candidates will receive positively or negatively biased evaluations due to the fact that they were rated by a relatively lenient or harsh rater (Guilford, 1954; Wilson, 1988). The analogous circumstance in testing occurs when two or more forms of a test with unequal levels of difficulty (i.e., nonparallel test forms) are used to assess a group of individuals. If the two test forms are not adjusted through procedures such as statistical equating, the scores of examinees who take different forms cannot be regarded as comparable.

Performance ratings that arise in situations where candidates are evaluated by different sets of raters contain two types of measurement error: systematic and random. The random error component is what is typically referred to as rater unreliability. If all candidates in a group are evaluated by the same raters, then the reliability coefficient can be estimated by the following equation:

$$\rho^2 = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_e^2/n_r} \tag{1}$$

where ρ^2 is the generalizability (reliability) coefficient, σ_i^2 refers to the variance component due to candidates, σ_e^2 refers to the variance component due to the error (i.e., residual variance), and n_r indicates the number of raters evaluating each candidate. These components of variance can be computed from the mean squares reported for a candidate by rater ANOVA (Brennan, 1983; Crocker & Algina, 1986; Shavelson, Webb, & Rowley, 1989).

The aforementioned method for obtaining a reliability coefficient applies to complete rating designs. If, however, an incomplete rating design is used, a design in which candidates are evaluated by different raters, then the reliability estimate must consider the error due to raters being

differentially lenient or harsh in their ratings. This systematic error is typically referred to as leniency error. The reliability for such incomplete designs is computed by:

$$\rho^2 = \frac{\sigma_i}{\sigma_i^2 + \left(\sigma_r^2 + \sigma_e^2\right)/n_r} \tag{2}$$

where σ_r^2 is the variance component due to raters. If all raters are equally lenient (i.e., the variance component due to raters is zero), then the two forms of the reliability index will be equal. This will seldom be the case, however. The present paper is concerned with methods to identify the systematic error component in performance ratings, and statistically control for such effects when they occur.

A very casual consideration of this rating problem might lead one to conclude that incomplete rating designs with multiple raters are uncommon. To the contrary, incomplete rating design are quite common. The situation in which ratings for a group are obtained from different sets of raters arises in numerous evaluation contexts, such as the following: organizations that make use of peer ratings, subordinate ratings, or ratings from multiple supervisors; educational settings in which different groups of students evaluate instructors; the evaluation of faculty by review committees; settings in which trainees are evaluated by multiple mentors, peers, or more senior students; oral or practical examinations for licensure; interviews of job applicants; scoring of essay examinations; ratings of physical or psychological status obtained in clinical settings; the review of grant applications and proposals; and the accreditation of institutions. There are also other evaluative situations in which incomplete designs occur, even though it may not be recognized. For example, businesses that have a hierarchical management structure may have an evaluation plan whereby departmental assistant managers each rate one-half the staff in a department.

In such circumstances, the manager of the department will also likely be in a position to evaluate many or all of the staff, resulting in an incomplete rating design with multiple raters.

The best solution to the problem of rater bias is to have the same panel of raters evaluate all candidates. However, numerous logistic and economic constraints usually render this solution infeasible. The purpose of this paper is to describe some practical, cost-effective procedures to correct for leniency/stringency effects in circumstances in which candidates are evaluated by two or more, but not the same, raters. Procedures based on item response theory (de Gruijter, 1984; Wright & Masters, 1982), multivariate analysis from incomplete data (Beale & Little, 1975; Houston, Raymond, Svec, in press; Little & Rubin, 1987; Raymond, 1986), and least squares regression (Cason & Cason, 1985; de Gruijter, 1984; Wilson, 1988) have been proposed in the literature. The procedures described in this paper either estimate scores candidates would have received had they been evaluated by all raters, or adjust observed scores to correct for effects due to rater variability.

Four methods are described in this paper: ordinary least squares (OLS), weighted least squares (WLS), the Rasch model, and data imputation via the E-M algorithm. Each procedure can be applied in circumstances in which multiple evaluators rate multiple candidates, but candidates are evaluated by some subset of raters. All methods require that each rater evaluate two or more candidates and that each candidate is evaluated by two or more evaluators. Although the student (n) by rater (p) data matrix may be severely incomplete, a certain degree of overlap must exist. That is, there should not be a subset of raters and candidates such that those candidates are evaluated only by that subset of raters, and the raters evaluate only that subset of students.

Correction Methods

Ordinary least squares (OLS)

Alternative regression-based procedures to identify and correct for rater effects have been proposed by Cason and Cason (1985), de Gruijter (1984), and Wilson (1988). A regression method for analyzing incomplete rating data postulates that an observed rating for a candidate is a function of the candidate's true ability and a leniency or stringency effect associated with the rater providing that particular rating. The model also assumes an error component. Consider the following model:

$$y_{ij} = \alpha_i + \beta_j + e_{ij} \tag{3}$$

where y_{ij} is the score given to candidate i by rater j,

α_i is the true score for candidate i,

β_j is the bias (i.e., leniency) index for rater j, and

e_{ij} is random error.

The model assumes that the error terms have an expected value of zero and that the variance of the errors across raters is equal.

Let a_i be an estimator of α_i, a candidate's true level of performance. Let b_j be an estimator of β_j, the magnitude of leniency of stringency error for rater j. We refer to this error as bias throughout the paper, consistent with the notion that the error is systematic as opposed to random. If candidate i is rated by all raters, then any estimator of α_i that sums or averages the observed ratings is free from rater bias effects (i.e., is an unbiased estimator). If, however, candidate i is not rated by all raters, then estimators of α_i will contain a bias component, unless $\beta_j = 0$ for all j, which is an unlikely circumstance.

The matrix formulation for the OLS model follows. Let K be the total number of observed ratings assigned by j raters to i candidates.

$$\mathbf{y} = \mathbf{X}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix} + \mathbf{e} \tag{4}$$

where $\mathbf{y}$ is a (K x 1) vector of observed ratings,

$\mathbf{X}$ is a (K x (n+p-1)) design matrix,

$\boldsymbol{\alpha}$ is an (n x 1) vector of true ratings for candidates,

$\boldsymbol{\beta}$ is a (p - 1) x 1 vector of rater bias indices, and

$\mathbf{e}$ is an (K x 1) vector of random errors.

The design matrix, $\mathbf{X}$, consists of n + p - 1 columns; the column for the last rater is dropped to avoid a linear dependency in the columns of $\mathbf{X}$. Because $\mathbf{X}$ is of full column rank, the parameters can be estimated by any standard multiple regression algorithm. Table 1.A presents an example of an incomplete rating matrix for a sample of three candidates and three raters, and Table 1.B presents the corresponding design matrix. For all candidates and all raters except for the last rater, the numeral 1 is used to indicate the candidate and rater with which each observed rating is associated; otherwise a zero is used. The ratings associated with the last rater are implied by coding the other p - 1 raters with a minus one (-1). This coding strategy produces the convenient and useful result that the β_j are in deviation form (i.e., $\sum_{j=1}^{p} \beta_j = 0$). Parameter estimates are then obtained through ordinary least squares regression, where $\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. As the vector in $\mathbf{X}$ corresponding to the last rater has been dropped, the parameter estimate for that rater will be missing from the OLS solution. The estimate for that rater is obtained by $-\sum_{j=1}^{p-1} b_j$, the negative of the sum of the parameter estimates for the other p-1 raters.

<u>Weighted least squares (WLS)</u>

The OLS procedure provides an unbiased estimate of the vector of true ratings. If, however, the consistency of scoring varies across raters, then

the usual regression assumption of equal error variances across all candidates and raters is violated; consequently, the variances of the parameter estimates will be inflated (Draper & Smith, 1981, p. 110). The practical consequence of the inconsistency is that the parameter estimates of candidates who were evaluated by inconsistent raters will be less accurate than the estimates associated with consistent raters.

Wilson (1988) suggested a two-stage regression procedure consisting of ordinary least squares, as described above, followed by weighted least squares. The weights for the second stage, which give less influence to inconsistent raters in the determination of the parameter estimates, are derived as follows. For each candidate/rater pairing that results in a rating, a residual is computed to indicate the accuracy with which that rater's observed rating corresponds to the rating predicted by the model. For example, if the value of a_i for candidate i is 3.50 and the value for rater b_j is .80, the model predicts that a value of 4.30 would be assigned by that rater to that candidate. The predicted value of 4.30 is then compared to the observed rating, and the difference between the predicted and observed rating is computed and squared. If evaluator j provides 10 ratings, the mean squared residual (MSR_j) based on those 10 ratings provides an index of evaluator consistency. The reciprocals of the mean squared residual ($1/MSR_j$) for all raters can then be used to derive weights for use in a generalized least square analysis to obtain revised estimates of the candidates' true scores. The WLS parameter estimates are given by

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{y} \tag{5}$$

where $\mathbf{W}$ is a K by K diagonal matrix of weights, with the elements of $\mathbf{W}$ corresponding to the value of $1/MSR_j$ for each rater. It may be noted that obtaining estimates via WLS is similar to weighted scoring in classical

measurement theory or in item response theory, whereby the influence of each test item in determining an examinee's total score is a function of each item's discrimination index (Stanley & Wang, 1970), or each item's slope (Lord, 1980).

Rasch Model

The Rasch model, also referred to as the one-parameter latent trait model for dichotomously scored items, states that the probability of a person providing a correct answer to a test item is a function of the person's ability and the difficulty of the test item (Wright & Stone, 1979). The Rasch model, like the two- and three-parameter latent trait models (Lord, 1980), specifies that the relationship between an individual's ability and their probability of answering an item correctly takes the form of a logistic function. Whereas the two-parameter model allows the slopes for each item to vary, and the three-parameter model allows the slope and asymptote to vary, the Rasch model assumes that the slopes are equal and the lower asymptotes are zero.

Extensions of the Rasch model to rating data are developed in Wright and Masters (1982), de Gruijter (1984), and Linacre (1989). The following presentation is based on that of Wright and Masters (1982). The Rasch approach for rating data models the probability that a person is assigned a rating category m+1 rather than a rating in category m. In addition to ability (person) and difficulty (rater) parameters, this application of the Rasch model contains another set of parameters, called threshold parameters, that correspond to differences in difficulty between adjacent categories of the rating scale. For a rating scale with M distinct categories there are a

total of M-1 threshold parameters. These parameters are assumed to be the same for all raters. The Rasch model may be written as follows.

$$\phi_{ijm} = \frac{\exp\,[\alpha_i - (\beta_j + \tau_m)]}{1 + \exp\,[\alpha_i - (\beta_j + \tau_m)]} \quad i=1 \text{ to } n;\ j=1 \text{ to } p;\ m=1 \text{ to } (M-1); \qquad (6)$$

where ϕ_{ijm} denotes the probability of examinee i being assigned by rater j a rating in category m+1 versus category m;

α_i is the ability level of examinee i;

β_j is the bias index for rater j;

τ_m is the difficulty associated with category m+1 versus category m, referred to as a threshold parameter by Wright and Masters (1982).

From expression (6), Wright and Masters (1982) derive a general expression for the probability that examinee i is assigned by rater j a rating in category m. The letters a, b, and t are used to denote estimates of α, β, and τ, respectively.[1] Once estimates of α_i, β_j, and τ_m are obtained, it is possible to predict for any rater/candidate pair, the rating that any rater j would assign to any candidate i. An important characteristic of the Rasch model is that it assumes a curvilinear relationship (logistic function) between observed ratings and actual performance. The logistic transformation helps minimize floor and ceiling effects in the observed ratings by stretching the tails of the score distribution.

Parameter estimates for the model are obtained through an iterative unconditional maximum likelihood algorithm. Iterations progressively adjust parameter estimates in order to maximize the likelihood function simultaneously over persons and raters. Initial estimates of a, b, and t obtained via the model are compared to the actual ratings (or a logistic transformation of the ratings) and the difference, or error, between the actual and modelled

---

[1]The parameter estimates for the Rasch model are on a logit scale, whereas estimates provided by OLS and WLS are on the same scale as the observed ratings. For convenience, similar notation has been used for all models.

ratings are computed. Subsequent iterations attempt to minimize this error by adjusting the estimates.

A variation of the Rasch rating scale model as described in expression (6) has been extended to multifaceted rating designs (e.g., designs in which topics, raters, and other factors may vary). The multifacted model is described in Linacre (1989) and Lunz, Wright, and Linacre (1990). The present paper is limited to two factor designs (raters, candidates). For such designs, the algorithm for multifacted designs provides results similar to the Wright and Masters (1982) rating scale model.

Imputed Ratings

Whereas the literature on correcting for rater effects is recent and fragmented, the literature on multivariate analysis from incomplete data is more consistent and spans several years (e.g., Beale & Little, 1975; Buck, 1960; Gleason & Staelin, 1975; Little & Rubin, 1987; Raymond, 1986; Raymond & Roberts, 1987). Methods for handling incomplete multivariate data make use of information available about each case to estimate or impute the missing data. Imputed ratings can be obtained in a variety of ways. A simple, pragmatic, and generally effective approach is based on multiple regression. With this approach, variables (raters) are regressed onto some or all of the other variables. Regression equations are then used to impute the missing data from the data that are present. The regression procedure can be iterated to obtain revised estimates, under the assumption that the imputed values obtained from successive revisions will be more accurate than the initial imputed values. Iterative versions of the regression approach appear very helpful when the amount of missing data approaches or exceeds about 10% (Beale & Little, 1975; Raymond, 1986; Raymond & Roberts, 1987).

A more theoretical version of the *ad hoc* regression approach is based on the E-M algorithm (Dempster, Laird, & Rubin, 1977). This iterative algorithm,

which is based on maximum likelihood estimation procedures, assumes the data are multivariate normal. Whereas the regression method imputes missing observations, the E-M algorithm obtains sufficient statistics using maximum likelihood estimation. Estimates of the means, variances, and covariances obtained from the incomplete data matrix are used to impute the missing data. Results obtained from the E-M algorithm are very similar to those obtained from iterated regression. The major difference is that the E-M algorithm computes a small (usually) correction to the variances and covariances. Several studies indicate that regression-based methods or the E-M algorithm are more effective than listwise deletion, pairwise deletion, or the mean substitution method for analyzing incomplete multivariate data.

Raymond (1986) suggested that methods for imputing missing data might be suitable for addressing the problem of incomplete rating data. The best estimate of any candidate's true score will be the sum or mean of the observed ratings and the imputed ratings. Similarly, an index of bias for each rater is obtained by computing the mean rating of observed and imputed ratings across candidates and then subtracting from each rater's mean the mean across all candidates and raters. However, the method of imputing may or may not prove suitable for levels of missing data encountered in incomplete rating designs, especially for small sample sizes. These procedures appear to merit further study.

In general applications, the method of imputing assumes that the pattern of missing data in the candidate (n) by rater (p) data matrix is quasi-random. That is, although strict randomization is not required, the pattern of missing data must be ignorable (Little & Rubin, 1987). The probability that an observation is missing should not depend on candidate or rater characteristics that might influence the ratings. For example, if low ability candidates are more frequently assigned to one group of raters than to another group, then

the mechanism by which the data are missing cannot be ignored. In the current application, in which emphasis is on imputing missing observations and not on making inferences about the underlying parameters, the normality assumption is not required.

Illustration of Correction Methods

Overview

The purpose of this paper is not only to specify the models and methods to correct for rater effects, but also to demonstrate the performance of the four methods (OLS, WLS, the Rasch model, and the E-M algorithm). To do so, we constructed a small sample of complete rating data with realistic properties, and randomly deleted two-thirds of the values. The incomplete rating data were then subjected to the four methods, and the candidate and rater scores produced by each method were then compared to the values obtained from the complete data. The remainder of this section of the paper explains the procedures for generating the simulated rating data and presents an overview of the analyses used to compare the four methods. The empirical demonstration is limited to one set of data and is not intended to document the effectiveness of the four methods; the data simply serve as a vehicle for illustrating the procedures. However, findings based on the present data are consistent with the results of extensive simulations (Houston, Raymond, & Svec, in press).

Data Generation

Data were simulated to provide the types of ratings that might be obtained from performance evaluations conducted in settings such as military training schools, physician residency programs, and other work settings in business, industry or government. Prior research and anecdotal reports

suggest that performance ratings may often have the following characteristics: 1) a 5- or 7-point Likert scale; 2) rater variances that range from about 0.5 to 1.5 on those scales; and 3) rater reliability ranging from about .30 to .80, with an average of about .50 (e.g., Berk, 1986; Cason & Cason, 1984; King, Schmidt, & Hunter, 1980; Rothstein, 1990; Streiner, 1985; Wakefield, 1985). Given the preceding guidelines, data were simulated for N=25 individuals evaluated by p=6 raters. We readily acknowledge the small sample size. Although the size of the data matrix might correspond to a cohort of students enrolled in an internship or training program, it was intentionally limited in size to facilitate a demonstration of the alternative methods.

Data were simulated in four steps. First, multivariate normal data were generated. The level of correlation among variables was induced so that the average rater reliability would be about .55, with three of the raters being considerably more reliable than the other three. Second, the data were converted to a 1 to 5 scale with a mean of about 3.25 and a standard deviation of approximately 1 for each rater. Third, varying degrees of bias were introduced by adding a constant (-1.00, -.50, -.25, +.25, +.50, +1.00) to the ratings of each rater. Fourth, data were rounded to the nearest integer and, where necessary, truncated to fit within the limits of the 1 to 5 Likert-type rating scale. The complete rating data appears in Table 2. The fifth and final step consisted of randomly deleting four of the six observations for each candidate, thereby producing a 67% incomplete rating design. The coefficient of generalizability for the complete ratings was .52, and the 15 interrater correlations among the six raters ranged from .31 to .74.

Analyses

The four procedures (OLS, WLS, Impute, and Rasch) were applied to the incomplete rating matrix (as indicated by the asterisks (*) in Table 2). Each method provides an adjusted rating, a_i, taken to be an estimate of each

candidates' true level of performance, and a rater bias index, b_j. In addition, the usual approach of simply obtaining the mean of observed ratings was also applied to the incomplete rating data. This method is referred to as uncorrected. The values of a_i and b_j based on incomplete data were compared to their corresponding estimates based on the complete data.

The methods are also capable of producing indices of candidate response consistency and rater consistency. For the least-squares procedures, the index is MSR as described earlier. The Rasch model produces two fit indices referred to as infit and outfit statistics. The E-M algorithm, which was used to impute ratings, produces an adjusted variance-covariance matrix and correlation matrix. These matrices can also be used to make inferences about the degree of consistency among raters. The various indices of candidate and rater fit were recorded and are also discussed.

Results of Applying Correction Methods

Candidate Information

Table 3 presents the ratings based on complete data, uncorrected incomplete data, and the incomplete data subsequent to being subjected to one of the four correction procedures. All procedures except for the Rasch model provide estimates that are on the same scale as the original ratings. Rasch model estimates, which are on a logit scale, range from about -5.0 to +6.3. All methods produced ratings that are noticeably different from the uncorrected ratings based on incomplete data. The ratings from incomplete data for candidates #4, #21, and #24 are of particular interest since they were severely misrated based on the uncorrected incomplete data. Although imputing seemed to deal most effectively with these aberrant cases, it performed poorly with other candidates (e.g., #25). All the other methods (OLS, WLS, Rasch) did <u>attempt</u> to bring the aberrant ratings (#4, #21, #24)

back into line, but were not completely successful. The Rasch estimates for candidates #4 and #21 are quite extreme and inaccurate. The inaccuracy is most likely due to the logistic transformation, which stretches the tails of the distribution.

To provide a succinct description of the values of a_i, descriptive statistics were computed for each method, including complete data. The data are presented at the bottom of Table 3. All methods (except for the Rasch model due to its use of a logit scale) provided overall means nearly identical to the complete data means. The standard deviations based on the correction methods appear to demonstrate some differences, though. In particular, the variability for imputed ratings is quite restricted. This occurs because of the tendency for imputed ratings to be regressed toward the overall mean. If the equations used to produce the imputed values are not very accurate (i.e., if the R-square is substantially less than 1.00), then the imputed values will be substantially regressed. It is obvious that if observed ratings are uncorrelated, then the imputed values will be equal to the observed mean.

Correlations among the estimated true ratings obtained from the various methods are provided in Table 4. Of most interest are the correlations between each of the procedures based on incomplete data with the complete data. All methods that provide adjusted ratings show a stronger relationship with the complete data than the do the unadjusted ratings. For this particular data set, using WLS offered no increase in accuracy compared to OLS. There are other interesting features of the correlation matrix in Table 4, the most noteworthy being the correlation of .986 between OLS and Rasch estimates.[2]

---

[2]In recognition of the fact that it is not strictly appropriate to compare the Rasch ratings to ratings based on a linear model, two precautions were taken. First, Rasch estimates based on complete data were used as a basis for comparing the results of the Rasch model applied to incomplete data. Second, the adjusted ratings produced by the other methods were subjected to a

Deviation scores were computed by finding the absolute value of the difference between ratings based on complete data and ratings based on the various methods for adjusting the incomplete data. The data were standardized prior to making this comparison to assure that deviation scores were on the same scale (again, refer to footnote 2). The data in Table 5 more or less confirm the correlations. The average error produced by not correcting for rater effects is about .56 S.D. units; whereas the magnitude of the error for the various correction procedures is, on the average, about .40 S.D. units. The increase in error encountered by not using corrected ratings, relative to using corrected ratings, is about 40%. The standard deviations and the ranges listed in Table 5 are also of interest in evaluating the various methods, as is the last column, which indicates the number of very aberrant ratings produced by each method. Noticeably larger errors are associated with the uncorrected ratings.

The Rasch and least squares methods produce indices indicating the degree to which the postulated model fits the data. These indices have also been referred to as fit statistics, and can be used for evaluating the degree to which candidates are appropriately measured. The Rasch model produces two statistics referred to as infit and outfit. The OLS method is capable of providing MSR_a, the mean squared residual for each candidate obtained over the raters evaluating that candidate. In general, there was a positive relationship among these indices. For example, the correlation between MSR_a and Rasch infit was .76 for the incomplete data. Furthermore, the correlation between MSR_a for complete data and Rasch infit based on complete data was .95,

---

logit transformation prior to comparison to the Rasch model. The bottom line is that it made very little difference. In fact, for the present data, the Rasch model (incomplete) appeared to fair a little worse by using Rasch-complete as a basis for comparison. In the interest of clarity and brevity, only results obtained from using the original, untransformed, complete data as the basis for comparison are presented here.

suggesting that the two types of indices characterize very similar aspects of the rating data. It is also worth noting that the OLS and Rasch fit statistics based on the incomplete data were generally similar to the corresponding fit statistics based on the complete data.

Rater Information

The indices of rater bias, b_j, appear in Table 6. The indices for all procedures have a mean of zero. Note that the signs for the Rasch model rater indices are reversed. All IRT-based models compute a difficulty index for the measurement instruments for which the more difficult, or stringent instruments have high values and the more lenient instruments have low values. Another notable characteristic of the data in Table 6 is the range for the Rasch values of b_j, which extend from -3.40 to 2.41. This is not unusual given the range of Rasch-based candidate scores (-5.05 to 6.29). However, there appears to be greater instability with the Rasch estimates of b_j. For example, the Rasch estimate of b_j for rater 6 based on complete data was only -2.4. The impute method estimated the rater biases very accurately, which is, of course, one of the primary goals that the method was initially designed to achieve.[3] The uncorrected estimates of rater bias were also reasonably accurate, attesting to the random assignment of raters to candidates. Of the other three methods (OLS, WLS, and Rasch), the OLS and WLS estimates were similar and were more accurate than the Rasch estimates. Although the magnitudes of the rater bias indices vary considerably, the rank ordering of raters according to their indices is identical for all procedures. Given the magnitude of the differences in rater bias, it is not surprising that all methods resulted in the same ranking.

---

[3]As noted earlier, the E-M algorithm's primary purpose is to estimate the complete data means and variance-covariance matrix.

Rater consistency (or rater fit) statistics can be obtained through the Rasch model, least squares procedures, and E-M algorithm. The Rasch model computes infit and outfit statistics for each rater, while the OLS procedure enables the computation of MSR_j to be obtained over all candidates rated by each rater. An index of rater reliability can also be obtained from the impute procedure, since it produces estimates of the complete data interrater correlation matrix. Once the interrater correlations are obtained, it is possible to compute two indices of rater consistency: 1) the average interrater correlation for each rater; or 2) each rater's correlation with the sum of the observed and imputed ratings.

For the complete data, the indices of rater consistency for each of the methods (OLS, Rasch, Impute) were extremely similar ($r > .90$). For the incomplete data, some differences occurred. Again, the MSR_j statistic resulting from OLS and the Rasch infit statistics were very similar in terms of identifying inconsistent raters. Both were generally capable of differentiating inconsistent raters from consistent ones, although rater 5, who was really a good rater in the complete data, was identified as an inconsistent rater by both methods based on the incomplete data. However, the incorrect flagging of rater 5 was consistent with the observed incomplete data for that rater 5, which consisted of eight ratings. A consistency index based on eight ratings should be expected to be unreliable due to sampling error. The correlation matrix resulting from the E-M algorithm bore very little semblance to the complete data correlation matrix. The small sample size and sparsity of the data result in very unstable estimates of interrater correlations.

At least for the present set of data, all correction methods appeared to offer some improvement over the common practice of summing or averaging the observed ratings (i.e., doing nothing at all). The E-M algorithm produced

imputed ratings that made the data matrix appear as if all candidates were rated by all raters. In most instances, the imputed ratings were reasonably accurate, although in general they were regressed toward the mean. The Rasch model and the correction procedures based on least squares analysis (OLS and WLS) also seemed to do an effective job of approximating the ratings a group of candidates would receive had they all been evaluated by the same raters. Furthermore, the Rasch-based and OLS methods provide very similar results in terms of the rank-ordering of both candidates and raters. The candidate and rater fit statistics obtained through least-squares and the Rasch model were also similar. The notable differences between the Rasch and least square models occurred at the extremes of the score scale. The similarity between the methods can be increased by subjecting the rating data to certain transformations (e.g., logit). Weighted least squares did not improve the OLS estimates in most instances. The lack of notable improvement is likely a consequence of small sample sizes, reasonable levels of rater consistency, or both.

Results of Other Studies

The four methods for correcting for rater effects have been applied to other small sets of simulated data, (e.g. N=15, p=5; N=30, p=6; N=25, p=5). Data reported in Wilson (1988) have also been analyzed (N=10, p=8). The results have been consistent in that all correction methods, including imputing missing values, provide more accurate results than not correcting (i.e., simply summing or averaging over the observed ratings). Furthermore, Rasch and OLS provide very similar rank-order results, with neither showing a consistent advantage. WLS appears to have advantages when the weights are obtained over more ratings and when there is considerable variability in MSR_j due to one or two of the raters being considerably more aberrant than the other raters. For example, in the rating data reported in Wilson (1988), two

0of the eight raters had rather extreme values of MSR_j relative to the other raters. The use of WLS by Wilson (1988) resulted in substantial improvements in estimates of true scores, as well as reductions in the standard errors associated with the estimates.

Houston, Raymond, and Svec (in press) conducted an experiment using numerous replicated sets of simulated rating data. The correction methods were investigated under varying conditions of sample size (N=50, 100), levels of incomplete data (50%, 75%), and levels of bias (low, high). The interrater correlation was .46 and the number of replicated simulations per conditions was 30. This research found that OLS, WLS, and the method of imputing, consistently resulted in less error (10% to 50% less error) than using uncorrected ratings (the Rasch model was not studied in that experiment). In addition, the method of imputing was always equal to, or more accurate than, the least squares procedures. OLS and WLS performed similarly.

In applied studies conducted within the context of certification examinations, it has been shown that rater effects are significant (Raymond, Webb, & Houston, in press; Lunz & Stahl, 1990) and that an individual rater's leniency index is reasonably stable over the period of a year (Raymond, Webb,, and Houston, in press). It has also been shown that the use of OLS adjusted ratings can affect the pass-fail decisions of approximately 6% of the candidates taking an oral examination (Raymond, Webb & Houston, in press). Within the context of essay ratings, Braun (1988) has demonstrated that the use of an OLS model to adjust observed ratings can result in substantial improvements in rater reliability. Under circumstances of pronounced rater effects, using an OLS model could actually result in greater gains in reliability than could be obtained by doubling the number of raters.

Preliminary Evaluation of Methods

The use of the E-M algorithm to impute missing ratings appears to be a feasible approach to correcting for rater effects. Although it has been proven to be very effective for managing incomplete multivariate data when 10% to 20% of the data are missing and the samples are moderate to large (Beale & Little, 1975; Raymond, 1986; Raymond & Roberts, 1987), its use with incomplete rating designs is a practice that clearly warrants some scrutiny. Incomplete designs in applied settings can often range from 75% to 90% incomplete. The suitability of the E-M algorithm for such designs is an empirical issue that needs to be addressed in future research. As noted earlier, it has been found to be effective when the rating designs are up to 75% incomplete (Houston, Raymond, & Svec, in press). However, it would not be appropriate to generalize this finding to the many other rating designs that exist in practice. One limitation of the E-M algorithm is that the rater consistency indices (i.e., estimates of complete data rater correlations) are likely to be very unstable with small samples. Another limitation, the fact that imputed estimates are regressed toward the mean, may present a particularly thorny problem in the context of criterion-referenced performance examinations. That is, the pass-fail status of some candidates may change because their imputed ratings were pulled toward the mean of all ratings. This potential complication will also need to be addressed by future research.

The Rasch model appears to have some notable advantages. It has a sound theoretical basis, it has occupied a position in the psychometric literature for many years now, and Rasch-based software for analyzing rating data is readily available. For example, two-factor rating designs can be analyzed with Microscale or FACETS (Linacre, 1989), while multifactorial rating designs can be handled by FACETS. In addition, Rasch-based software produces considerable useful information concerning the characteristics of the rating

data (e.g., standard errors and fit statistics). One possible disadvantage of the Rasch model is that it assumes a curvilinear relationship (logistic function) between observed ratings and actual performance. The logistic transformation presumably minimizes floor and ceiling effects. Although it may be useful or required for rating scales with a limited number of scale values (e.g., less than seven points), the transformation may not be advantageous for rating scales with a large number of scale points. For example, in the context of oral ratings based on a 12-point Likert scale, a logistic transformation offered no improvement in the fit of rating data to an OLS model (Raymond, Webb, & Houston, in press).

The least-squares approaches (OLS and WLS) also have obvious desirable features. Least-squares regression is not a psychometric model with a theoretical basis and numerous assumptions, but is, instead, a general statistical model with few assumptions. The least-square models, like the Rasch-based FACETS program, can also accompany multifactorial designs. Both OLS and WLS provide adjusted ratings and bias indices that are on the same scale as the original data. Therefore, the linkage between the original data and corrected data is relatively direct and obvious (although less obvious for WLS). Another notable benefit of the least-squares procedures is that, like the Rasch-based algorithms, it is possible and easy to obtain a wealth of information concerning the characteristics of the rating data. In addition to obtaining the standard errors of parameter estimates and the consistency indices based on the residuals, one can also generate the variety of diagnostic statistics and plots discussed in the literature on regression diagnostics (e.g., Belsley, Kuh, & Welsch, 1980).

In short, the least-squares methods are more flexible than either the Rasch model or the method of imputing ratings; they can more readily be adapted to varieties of data and rating designs that may be encountered in

practice. Although Lunz, Wright, and Linacre (1990) indicate that an ANOVA (i.e., least-squares) approach to correcting for rater leniency/stringency error is not possible for incomplete designs and categorical rating data, the models presented in this paper suggest otherwise. Any one of a number of curvilinear transformations can be applied should the data warrant such a transformation. Similarly, WLS can be invoked should the assumption of equal variance of errors be violated, and/or if weighted scoring appears to be desirable. Least-squares analysis, particularly OLS, may also be more suitable for small samples. Under such conditions, curvilinear transformations, including the transformation used by the Rasch model, have the capability of converting small, possibly random, differences between candidates or raters into large differences.

Although there is no prepackaged software designed specifically for applying least squares analysis to rating data, it is relatively simple to perform OLS or WLS using a variety of software packages. The authors have used SPSS-PC (MANOVA) and SYSTAT (MGLH) to perform analyses on a microcomputer for small data sets, and SAS (PROC REG, PROC GLM, and PROC MATRIX) on an IBM mainframe for large data sets ($N=120$, $p=40$). The packages vary considerably in terms of the ease with which the regression diagnostics can be obtained.

Recommendations for Future Research

The methods to correct for rater effects show promise; however, much more research is needed. One line of research would involve the investigations of alternative models to correct for rater effects, or extensions of the present models. As noted earlier, multifactor models have been introduced (Braun, 1988) as have curvilinear models (Cason & Cason, 1984; 1985) and multifactor curvilinear models (Linacre, 1989). The method of imputing, which appears to

be useful for two-factor data, might also be extended to multifactor (or multiway) data.

A second line of research would involve conducting simulation studies that consider a variety of correction procedures under a wide range of conditions (type of rating data, sample size, number of raters, level of rater reliability, level of rater bias, percent and pattern of missing data). For example, one study might examine the methods under the various conditions encountered in performance rating, whereas other studies might address the conditions under which essay or oral examinations are typically graded. Overall model fit of the simulated rating data will influence the results of such studies. Two very critical factors in such studies will be the level of rater bias and the degree of interrater correlation. Rater bias can be monitored by evaluating the magnitude of the rater effect resulting from a two factor (candidate by rater) ANOVA. The sums of squares, mean squares, and variance components associated with each effect (candidates, raters, error) can be useful. As the σ_r^2 , the variance component due to raters, approaches zero, the need to correct for rater effects diminishes. As σ_e^2 , the variance component associated with error approaches zero, then the capability to correct for rater bias increases. Therefore, the effectiveness of the correction procedures will be a joint function of σ_r^2 and σ_e^2 . Clearly, as σ_e^2 becomes large relative to other sources of variance, no correction procedure will be effective. In such instances, neither the corrected ratings nor the uncorrected observed ratings will be suitable for decision-making purposes.

Simulations will also be useful for identifying the rating designs that will optimize the effectiveness of different correction models. However, it will still be necessary to perform research to answer questions regarding

applications in specific settings using actual rating data. Thus, a third line of research would involve applied investigations designed to uncover special problems that may be encountered in operational use. Only after extensive study should the correction methods be applied operationally.

Since the various correction methods provide an opportunity to evaluate and monitor rating behavior over time, a fourth line of research would need to assess the reliability, validity, and utility of various indices for describing rater performance. All of these methods offer a way to assess ethnic bias, gender bias, or other types of bias in ratings. For example, the least-squares methods provide values of b_j and MSR_j that can be obtained and compared for selected subgroups. Phenomena such as rater bias are difficult to evaluate with the procedures typically used for assessing rating behavior.

The use of any method to correct for rater effects may also have important implications for the manner in which raters are trained. If future research shows that such methods can be effectively applied to rating data, then perhaps training programs should be redesigned to complement and take advantage of the correction procedures. Most rater training programs address the common rating errors of leniency/stringency, halo, and central tendency. Since the correction methods have the capability to correct for leniency/stringency errors under conditions of low levels of error variance, it may be unnecessary, and perhaps counterproductive, to continue to lecture raters about errors of leniency/stringency. Instead, rater training might allocate more time to teaching raters to discriminate among different levels of performance. Both common sense and the prior data suggest that leniency or stringency is a relatively stable personal characteristic that is not changed by training (Bernardin & Pence, 1980; Lunz & Stahl, 1990; Raymond, Webb, & Houston, in press). Training that attempts to overcome this characteristic may be destined for failure. Within the context of the correction procedures,

it is the inconsistent raters, not the lenient or stringent raters, who present the most serious problem.

Assessment situations that make use of subjective ratings of performance are common and pervasive. Important decisions about an individual's career hang in a rather dubious balance comprised of data that is often of very limited reliability. However, assessment techniques such as work samples, practical examinations, essays, and oral examinations have many potential strengths, including the possibility of improved validity and increased public acceptance. Both the general public and influential policy makers are calling for a decrease in multiple-choice testing and an increase in alternative assessment methods that rely on subjective ratings of performance (National Commission on Testing and Public Policy, 1990). If society and the measurement community plan to increase their use of assessment methods that rely on performance ratings, then problems related to rater reliability need to be addressed in order for these methods to make a valuable contribution to the types of decisions for which assessment data are used. The methods discussed in this paper may serve as a foundation to improve the quality of performance rating data.

References

Beale, E. M. L, & Little, R. J. A. (1975). Missing data in multivariate analysis. Journal of the Royal Statistical Society (B), 129-145.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: Wiley and Sons.

Berk, R. A. (1986). Performance assessment: Methods & applications. Baltimore: Johns Hopkins University Press.

Bernardin, H. J. & Pence, E. C. (1980). Effects of rater training: creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 13, 1-18.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City: The American College Testing Program.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, Series B, 22, 302-307.

Cason, G. J. & Cason, C. L. (1984). A deterministic theory of clinical performance rating. Evaluation and the Health Profession, 7, 221-247.

Cason, G. J. & Cason, C. L. (1985). A Regression Solution to Cason and Cason's Model of Clinical Performance Rating: Easier, Cheaper, Faster. Presented at the annual meeting of the American Educational Research Association, Chicago.

Crawford, W. (1984). The California model: Semi-structured oral examinations. Professional Practice of Psychology, 5(2), 86-93.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

de Gruijter, D. N. M. (1984). Two simple models for rater effects. Applied Psychological Measurement, 8(2), 213-218.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (B), 39, 1-38.

Draper, N. R. & Smith, H. (1981). Applied regression analysis. 2nd ed. New York: Wiley.

Gleason, T. C. & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40, 229-252.

Guilford, J. P. (1954). Psychometric methods. 2nd ed. New York: McGraw-Hill.

Houston, W. M., Raymond, M. R., & Svec, J. (in press). Adjustments for rater effects in performance assessment. Applied Psychological Measurement.

King, L. M., Schmidt, F. L., & Hunter, J. E. (1980). Halo in a multidimensional forced-choice evaluation scale. Journal of Applied Psychology, 65, 507-516.

Landy, F. J. & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72-101.

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

Linacre, J. M. (1989). Objectivity for judge-intermediated certification examinations. Presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Little, R. J. A. & Rubin, D. B. (1987). Statistical analysis with missing data. New York: John Wiley.

Lord, R. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Mass: Addison-Wesley.

Lunz, M. E. & Stahl, J. A. (1990). Judge consistency and severity across grading periods. Evaluation and the Health Professions, 13, 425-444.

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.

Muzzin, L. J. & Hart, L. (1985). Oral examinations. In V. R. Neufeld and G. R. Norman (eds.). Assessing clinical competence. New York: Springer.

National Commission on Testing and Public Policy (1990). From gatekeeper to gateway: Transforming testing in America. Boston, MA: Author.

Raymond, M. R. (1986). Missing data in evaluation research. Evaluation and the Health Professions, 9, 395-420.

Raymond, M. R. & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. Educational and Psychological Measurement, 47, 13-26.

Raymond, M. R., Webb, L. C., & Houston, W. M. (in press). Correcting performance rating errors in oral examinations. Evaluation and the Health Professions.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptate level with increasing opportunity to observe. Journal of Applied Psychology, 75, 322-327.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. Psychometrika, 26, 203-219.

Stanley, J. C. & Wang, M. D. (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. Educational and Psychological Measurement, 30, 21-35.

Streiner, D. L. (1985). Global rating scales. In V. R. Neufeld and G. R. Norman (eds.). Assessing clinical competence. New York: Springer.

Wakefield, J. (1985). Direct observation. In V. R. Neufeld and G. R. Norman (eds.). Assessing clinical competence. New York: Springer.

Wilson, H. G. (1988). Parameter estimation for peer grading under incomplete design. Educational and Psychological Measurement, 48, 69-81.

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

Wright, B. D. & Stone, M. (1979). Best Test Design. Chicago: MESA Press.

Table 1

Example of Incomplete Rating Design and the Corresponding OLS Design Matrix

Table 1.A

Candidate	Rater A	Rater B	Rater C	Rater D
1	2		4	
2		2	5	
3	3			4
4		4		6
5	6			6
6	5	4		

Table 1.B

Observed Score	Candidates C_1	C_2	C_3	C_4	C_5	C_6	Raters R_A	R_B	R_C
2	1	0	0	0	0	0	1	0	0
4	1	0	0	0	0	0	0	0	1
2	0	1	0	0	0	0	0	1	0
5	0	1	0	0	0	0	0	0	1
3	0	0	1	0	0	0	1	0	0
4	0	0	1	0	0	0	-1	-1	-1
4	0	0	0	1	0	0	0	1	0
6	0	0	0	1	0	0	-1	-1	-1
6	0	0	0	0	1	0	1	0	0
6	0	0	0	0	1	0	-1	-1	-1
5	0	0	0	0	0	1	1	0	0
4	0	0	0	0	0	1	0	1	0

Table 2

Complete Rating Data

	Rater					
Candidate	1	2	3	4	5	6
1	1	1*	2	2	2	3*
2	1	1	2*	2	3	3*
3	2	2*	1	3*	3	3
4	2	3	1*	3	2*	3
5	3	2	2*	2	2*	3
6	2*	3	1	3	3	4*
7	3*	2	2*	3	3	4
8	3	2	2	3*	4*	3
9	3*	2*	2	3	4	4
10	3	3	2*	3*	4	4
11	3	3*	2	3*	4	5
12	4*	3	3	3*	3	4
13	4	2	2	3	4*	5*
14	3*	4*	2	3	4	4
15	3	3	1	4*	5	4*
16	4	3	2*	3	4*	5
17	4	3*	2	4*	3	5
18	4	2*	3	4	4*	5
19	3*	2	3	5	5*	5
20	4*	3	2*	5	4	5
21	4	3	2	5*	5	5*
22	2	5*	3	5	4*	5
23	3	3	4	5	5*	5*
24	3*	5	4*	4	5	5
25	5	4*	4	5	5	5*

*The asterisk (*) denotes ratings that remained after deleting 67% of the data.

Table 3

Overall Rating Provided by Each Method*

Candidate	Complete Data	Incomplete Data				
		Uncorrected	Impute	OLS	WLS	Rasch
1	1.83	2.00	2.50	1.81	2.02	-4.12
2	2.00	2.50	2.46	2.45	2.40	-2.37
3	2.33	2.50	2.90	2.65	2.67	-0.15
4	2.33	1.50	2.52	1.59	1.60	-5.06
5	2.33	2.00	2.07	2.09	2.13	-2.70
6	2.67	3.00	3.03	2.76	3.06	-0.87
7	2.83	2.50	3.16	3.00	2.91	0.86
8	2.83	3.50	3.20	3.20	3.04	0.83
9	3.00	2.50	2.96	2.86	2.78	0.45
10	3.17	2.50	3.01	2.79	2.75	0.16
11	3.33	3.00	3.29	3.15	3.03	1.09
12	3.33	3.50	3.56	3.60	3.20	2.09
13	3.33	4.50	3.68	3.85	4.07	2.40
14	3.33	3.50	3.75	3.86	3.99	2.75
15	3.33	4.00	3.08	3.55	3.51	1.50
16	3.50	3.00	3.57	3.09	3.07	0.72
17	3.50	3.50	3.48	3.65	3.67	2.29
18	3.67	3.00	3.11	2.96	2.95	0.28
19	3.83	4.00	3.81	3.90	4.00	2.88
20	3.83	3.00	3.62	3.50	3.31	2.10
21	4.00	5.00	3.66	4.55	4.51	6.29
22	4.00	4.50	3.84	4.45	4.54	4.88
23	4.17	5.00	3.94	4.35	4.35	4.98
24	4.33	3.50	3.87	4.00	4.12	3.28
25	4.67	4.50	3.76	4.31	4.33	4.49
Mean	3.260	3.280	3.272	3.278	3.279	1.161
S.D.	0.736	0.947	0.510	0.805	0.811	2.761

*Note: The complete data ratings for each candidate are based on the mean of p=6 observed ratings. The incomplete ratings were obtained from p=2 ratings (data were 67% incomplete).

Table 4

Correlations of Complete, Uncorrected, and Corrected Ratings (N=25)

		Incomplete Data				
Method	Comp	Uncor	Impute	OLS	WLS	Rasch
Complete	1.000					
Uncorrected	.777	1.000				
Imputed	.875	.812	1.000			
OLS	.881	.939	.914	1.000		
WLS	.858	.943	.901	.984	1.000	
Rasch	.879	.906	.896	.986	.959	1.000

Table 5

Descriptive Statistics for Standardized Absolute Difference Scores (N=25)*

Method	Mean	SD	Min	Max	Percent of Diffs. >.75 SD
Uncorrected	.562	.341	.02	1.23	6
Imputed	.397	.292	.06	1.10	4
OLS	.413	.247	.00	.95	2
WLS	.456	.258	.09	.96	4
Rasch	.388	.291	.02	.99	4

*Obtained by computing the absolute difference between standardized ratings for each method and the standardized ratings on complete data.

Table 6

Estimates of Rater Bias Provided By Each Method

		Incomplete Data				
Rater	Complete Data	Uncorrected	Impute	OLS	WLS	Rasch
1	-0.22	-0.16	-0.26	-0.31	-0.30	1.08
2	-0.50	-0.39	-0.44	-0.41	-0.44	1.43
3	-1.02	-1.16	-1.02	-0.69	-0.66	2.41
4	0.26	0.22	0.28	0.11	0.21	0.04
5	0.50	0.50	0.52	0.50	0.47	-1.57
6	0.98	0.97	0.92	0.80	0.72	-3.40