Adjustments for Rater Effects in Performance Assessment

Walter M. Houston, Mark R. Raymond, and Joseph C. Svec
American College Testing

Alternative methods to correct for rater leniency/stringency effects (i.e., rater bias) in performance ratings were investigated. Rater bias effects are of concern when candidates are evaluated by different raters. The three correction methods evaluated were ordinary least squares (OLS), weighted least squares (WLS), and imputation of the missing data (IMPUTE). In addition, the usual procedure of averaging the observed ratings was investigated. Data were simulated from an essentially τ-equivalent measurement model, with true scores and error scores normally distributed. The variables manipulated in the simulations were method of correction (OLS, WLS, IMPUTE, averaging the observed ratings), amount of missing data (50% missing, 75% missing), rater bias (low, high), and number of examinees or candidates ($N = 50$, $N = 100$). The accuracy of the methods in estimating true scores was assessed based on the square root of the average squared difference between the estimated and known true scores. The three correction methods consistently outperformed the procedure of averaging the observed ratings. IMPUTE was superior to the least squares methods. *Index terms: EM algorithm, incomplete data, incomplete rating designs, least squares adjustments, performance assessment, rater calibration.*

The reliability of performance ratings given in practical settings is typically quite low, often falling below .60 (King, Schmidt, & Hunter, 1980; Rothstein, 1990). The most direct way to address the problem of low reliability is to obtain ratings from multiple raters. For example, according to classical psychometric theory (Lord & Novick, 1968), if the reliability of a single rating is .50, then the reliability of two, four, and six parallel ratings will be approximately .67, .80, and .86, respectively. The practice of using multiple evaluators to improve reliability is analogous to constructing tests and surveys that consist of multiple questions.

The use of multiple raters generally improves reliability by increasing the ratio of true score variance to error variance. However, this practice does little to address the systematic error that may arise when candidates within a group are evaluated by different raters. Unless the same raters evaluate all candidates, some candidates will likely receive a positively or negatively biased evaluation due to the fact that they were rated by a relatively lenient or harsh rater (Guilford, 1954; Wilson, 1988). A similar circumstance occurs in testing when two or more forms of a test with unequal levels of difficulty (i.e., nonparallel forms) are used to assess a group of individuals. If the two test forms are not adjusted through procedures such as equating, the scores of examinees who take different forms cannot be regarded as comparable.

Ratings are obtained from multiple but different sets of raters in circumstances such as the following: situations 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; assessment centers; clinical settings and military settings in which trainees are evaluated by senior students; oral examinations; interviews of job applicants; scoring of essay examinations; and the accreditation of institutions or programs.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 4, December 1991, pp. 409-421

0146-6216/91/040409-13$1.90

To date, several plausible models have been proposed to correct for leniency/stringency effects in performance ratings. However, little research has compared the relative effectiveness of the various approaches. The research that has been performed demonstrates that application of least squares procedures to actual rating data results in substantial adjustments (de Gruijter, 1984; Raymond, Webb, & Houston, 1991). Within the context of essay examinations, Braun (1988) demonstrated that using least squares procedures to correct for rater bias significantly increased the reliability of essay ratings. Whether the correction procedures actually improve the accuracy of the ratings has yet to be demonstrated in a study where the true scores are known.

The purpose of the current research was to conduct a study in which true scores were known in order to compare the effectiveness of alternative procedures to correct for rater effects under systematically varied conditions. The three methods used in this study to correct for leniency/stringency effects were ordinary least squares (OLS) (de Gruijter, 1984; Wilson, 1988), weighted least squares (WLS) (Wilson, 1988), and imputation of the missing data through the EM algorithm (IMPUTE) (Beale & Little, 1975; Dempster, Laird, & Rubin, 1977). These methods all require that each evaluator rate two or more examinees and that each examinee is rated by two or more evaluators.

The Measurement Model

The following measurement model may be used to describe ratings resulting from a subjective evaluation of performance:

$$X_{ij} = \tau_i + \delta_j + \varepsilon_{ij} \quad (i = 1 \text{ to } n;\ j = 1 \text{ to } p) \quad , \tag{1}$$

where X_{ij} is the observed rating given examinee i by rater j,
τ_i is the true score for examinee i,
δ_j is the rater effect for rater j, defined as the mean of rater j across all examinees minus the grand mean of all p raters across all examinees, and
ε_{ij} is random error.
Note that by definition $\sum_{j=1}^{p} \delta_j = 0$.

For examinee i, the vector of observed ratings is given by

$$\mathbf{x}'_{i(1\times p)} = \tau_i \mathbf{1}'_{(1\times p)} + \boldsymbol{\delta}'_{(1\times p)} + \boldsymbol{\varepsilon}'_{i(1\times p)} \quad . \tag{2}$$

Across the n examinees, the matrix of observed ratings can be expressed as

$$\mathbf{X}_{(n\times p)} = \boldsymbol{\tau}_{(n\times 1)} \mathbf{1}'_{(1\times p)} + \mathbf{1}_{(n\times 1)} \boldsymbol{\delta}'_{(1\times p)} + \mathbf{E}_{(n\times p)} \quad , \tag{3}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}'_n \end{bmatrix}, \boldsymbol{\tau}_{(n\times 1)} = \begin{bmatrix} \tau_1 \\ \cdot \\ \cdot \\ \cdot \\ \tau_n \end{bmatrix}, \boldsymbol{\delta}_{(p\times 1)} = \begin{bmatrix} \delta_1 \\ \cdot \\ \cdot \\ \cdot \\ \delta_p \end{bmatrix} \quad , \tag{4}$$

and

$$\mathbf{E}_{(n\times p)} = \begin{bmatrix} \boldsymbol{\varepsilon}_1' \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\varepsilon}_n' \end{bmatrix} . \tag{5}$$

The model assumes that, across raters, the error terms are independent and normally distributed and that, for a given rater, the variances of the error terms are constant across all examinee true scores. These two assumptions are formally stated as:

Assumption 1. $\boldsymbol{\varepsilon}_i$ ($i = 1$ to n) are iid $N_p[\mathbf{0},\boldsymbol{\Sigma}]$, where $\boldsymbol{\Sigma}$ is a $p \times p$ diagonal matrix with the jjth diagonal element corresponding to the error variance for rater j ($j = 1$ to p). Thus, for a given τ_i, the elements of $\mathbf{x}_i$ were assumed to be conditionally (locally) independent (Novick, 1966). Assumption 1 implies that $(\mathbf{x}_i|\tau_i,\boldsymbol{\Sigma},\boldsymbol{\delta})$ are iid $N_p[\mathbf{1}_{(p\times 1)}\tau_i + \boldsymbol{\delta},\boldsymbol{\Sigma}]$ for $i = 1$ to n.

For purposes of the simulation, it was also assumed that, across examinees, the true scores were normally distributed:

Assumption 2. τ_i is distributed, across examinees, as a random variable with $E(\tau_i) = \mu_T$ and $\text{Var}(\tau_i) = \sigma_T^2$.

As in much of classical test theory, errors across repeated measurements (raters) were assumed to be stochastically independent. The measurement model postulated is a strengthened version of the essentially τ-equivalent model described in Lord and Novick (1968). Conditional on τ_i, errors across raters were independent with variances not required to be equal. For the τ-equivalent model, it was also assumed that the errors were normally distributed. If the diagonal elements of $\boldsymbol{\Sigma}$ were required to be equal, the model would be parallel, with an added assumption of normality.

For examinee i, the mean of the p observed ratings is:

$$\frac{1}{p}\mathbf{x}_i'\mathbf{1}_{(p\times 1)} = \frac{1}{p}(\tau_i \mathbf{1}'\mathbf{1} + \boldsymbol{\delta}'\mathbf{1}_{(p\times 1)} + \boldsymbol{\varepsilon}_{i(1\times p)}'\mathbf{1}_{(p\times 1)}) . \tag{6}$$

Then,

$$E\left(\frac{1}{p}\mathbf{x}_i'\mathbf{1} \mid \tau_i, \boldsymbol{\Sigma}, \boldsymbol{\delta}\right) = \tau_i + \frac{1}{p}\boldsymbol{\delta}'\mathbf{1} + \frac{1}{p}E(\boldsymbol{\varepsilon}_i'\mathbf{1}) = \tau_i , \tag{7}$$

where the expectation is with respect to the replication space and $\mathbf{1}$ is a column vector of unities. By Equation 7, $(1/p)\mathbf{x}_i'\mathbf{1}$ is an unbiased estimator of τ_i. Also,

$$\text{Var}\left(\frac{1}{p}\mathbf{x}_i'\mathbf{1} \mid \tau_i, \boldsymbol{\Sigma}, \boldsymbol{\delta}\right) = \frac{1}{p^2}\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} . \tag{8}$$

Equation 8 is the squared conditional standard error of measurement.

Then,

$$E\left(\frac{1}{p}\mathbf{x}_i'\mathbf{1} \mid \boldsymbol{\Sigma}, \boldsymbol{\delta}\right) = E_{f(T)}E\left(\frac{1}{p}\mathbf{x}_i'\mathbf{1} \mid \tau_i, \boldsymbol{\Sigma}, \boldsymbol{\delta}\right) = E_{f(T)}(\tau_i) = \mu_T \tag{9}$$

and

$$\mathrm{Var}\left(\frac{1}{p}\mathbf{x}_i'\mathbf{1} \mid \boldsymbol{\Sigma}, \boldsymbol{\delta}\right) = \mathrm{Var}_{f(T)}\,\mathrm{E}\left(\frac{1}{p}\mathbf{x}_i'\mathbf{1} \mid \tau_i, \boldsymbol{\Sigma}, \boldsymbol{\delta}\right) + \mathrm{E}_{f(T)}\,\mathrm{Var}\left(\frac{1}{p}\mathbf{x}_i'\mathbf{1} \mid \tau_i, \boldsymbol{\Sigma}, \boldsymbol{\delta}\right)$$

$$= \mathrm{Var}_{f(T)}(\tau_i) + \mathrm{E}_{f(T)}\left(\frac{1}{p^2}\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}\right) = \sigma_T^2 + \frac{1}{p^2}\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} \quad . \tag{10}$$

Equation 10 makes use of the assumption that the conditional standard error of measurement is homoscedastic across true scores.

The reliability of the average of the p observed ratings $[(1/p)\mathbf{x}_i'\mathbf{1}]$ is by definition,

$$\rho_{\bar{x}\bar{x}'} = \frac{\sigma_T^2}{\sigma_T^2 + \frac{1}{p^2}\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} \quad . \tag{11}$$

Other quantities of interest include the elements of the interrater reliability matrix.

Let $\mathbf{P}_{(p \times p)}$ denote the population interrater reliability matrix with elements $\rho_{jj'}(j = 1 \text{ to } p; j' = 1 \text{ to } p)$. The jj'th element of $\mathbf{P}$ is given by

$$\rho_{jj'} = \frac{\sigma_T^2}{(\sigma_T^2 + \boldsymbol{\Sigma}_{jj})^{1/2}(\sigma_T^2 + \boldsymbol{\Sigma}_{j'j'})^{1/2}} \quad . \tag{12}$$

The Correction Procedures

Consider the situation in which each examinee is rated by p^* raters, where $2 \leq p^* < p$. Unless δ_j is equal to 0 for every rater, then the δ_j do not necessarily sum to 0 over p^*. In this case, estimators that sum or average the obtained ratings (e.g., Equation 6) contain a bias component due to rater effects. Nevertheless, the most common method employed for estimating an examinee's true score is to average the ratings obtained by the p^* raters. This method is denoted NOTHING.

Ordinary least squares. The OLS approach to estimating examinees' true scores assumes that the linear measurement model in Equation 1 holds (i.e., the error terms have expected values of 0 and the variance is constant across all raters and examinees). Thus, the OLS method assumes a classically parallel measurement model—the diagonal elements of $\boldsymbol{\Sigma}$ are assumed to be equal. Let K denote the total number of observed ratings assigned by p raters to n candidates. K is the total number of nonmissing elements in the matrix $\mathbf{X}$ from Equation 3. The matrix formulation for OLS is

$$\mathbf{y} = \mathbf{A}\begin{bmatrix}\boldsymbol{\tau} \\ \boldsymbol{\delta}^*\end{bmatrix} + \boldsymbol{\varepsilon} \quad , \tag{13}$$

where $\mathbf{y}$ is a $K \times 1$ vector of observed ratings,
$\mathbf{A}$ is a $K \times (n + p - 1)$ design matrix,
$\boldsymbol{\tau}$ is a $n \times 1$ vector of true ratings for examinees,
$\boldsymbol{\delta}^*$ is a $(p - 1) \times 1$ vector of rater bias effects, and,
$\boldsymbol{\varepsilon}$ is a $K \times 1$ vector of random error terms.

The vector $\mathbf{y}$ contains the nonmissing elements of the matrix $\mathbf{X}$ in Equation 3, stacked into a column vector. $\boldsymbol{\delta}^*$ contains the first $(p - 1)$ elements of $\boldsymbol{\delta}$. The design matrix $\mathbf{A}$ consists of $(n + p - 1)$ columns. The first n columns are 0 - 1 variables coding the n examinees (i.e., dummy coding); the next $(p - 1)$ columns are 0 - 1 variables coding the first $(p - 1)$ raters. The column corresponding to the last rater is dropped to avoid a linear dependency. The ratings associated with the last rater are

implied by coding the other $(p - 1)$ raters -1.

Least squares estimates of the parameters τ and δ^* are given by

$$\begin{bmatrix} \hat{\tau} \\ \hat{\delta}^* \end{bmatrix} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y} \quad . \tag{14}$$

The estimated rater bias parameter for rater p is equal to $-\sum_{j=1}^{p-1}\delta_j^*$. Under the assumption that the error terms are normally distributed, the estimates given in Equation 14 are maximum likelihood estimates. Standard linear model theory provides confidence intervals on the estimated true scores and rater bias indices.

The estimated true score for examinee i obtained from Equation 14 is equal to the estimated true score obtained from NOTHING minus the average of the estimated rater bias parameters for those raters who rated examinee i. Using this result, it is possible to take into account the sampling variance of the estimated rater bias parameters in estimating an examinee's true score. For example, raters whose estimated rater bias parameters have large sampling variances could be excluded from the estimation of the τ_i.

Weighted least squares. OLS provides an unbiased estimate of the vector of true ratings. If, however, reliability varies across raters, then the usual regression assumption of homoscedastic error variance across all examinees and raters is violated. Consequently, the sampling variances of the parameter estimates will be inflated (Draper & Smith, 1981). Furthermore, the parameter estimates for candidates evaluated by unreliable raters will be less accurate than the estimates associated with reliable raters.

Wilson (1988) suggested a two-stage regression procedure consisting of OLS, as described above, followed by WLS. The weights for the second stage give less influence to inconsistent raters in the determination of the parameter estimates. The weight associated with each observed rating is the reciprocal of the mean squared residual (across examinees) of the rater providing the observed rating. The WLS parameter estimates are given by

$$\begin{bmatrix} \hat{\tau} \\ \hat{\delta}^* \end{bmatrix} = (\mathbf{A}'\mathbf{W}\mathbf{A})^{-1}\mathbf{A}'\mathbf{W}\mathbf{y} \quad , \tag{15}$$

where $\mathbf{W}$ is a $K \times K$ diagonal matrix of weights, with elements corresponding to the mean squared residual associated with each rater. Note that both OLS and WLS assume that examinees and raters are fixed effects.

Imputation of the missing data. Whereas the literature on correcting for rater bias effects is recent and fragmented, the literature on multivariate analysis of missing data is more consistent and spans several years (Beale & Little, 1975; Buck, 1960; Hartley & Hocking, 1971; Little, 1976; Little & Rubin, 1987; Raymond & Roberts, 1987; Rubin, 1976). Although much of the prior research has been concerned with the estimation of underlying parameters when missing data are present (with emphasis on maximum likelihood estimation), the emphasis in the current application is on imputing the missing observations themselves. An examinee's true score is estimated by averaging the obtained and imputed ratings (Raymond, 1986).

Imputed ratings can be obtained in a variety of ways. A pragmatic and generally effective approach is based on multiple regression (Beale & Little, 1975; Buck, 1960; Raymond & Roberts, 1987). With this approach, variables (i.e., raters) with missing data are regressed onto some or all of the other variables. The regression equations are then used to obtain estimates of the missing data from the data that are present. The regression procedure can be iterated to obtain revised estimates. Iterative

regression appears to be most advantageous as the level of missing data approaches or exceeds 10% (Beale & Little, 1975).

A more theoretical version of the ad hoc regression approach is based on the EM algorithm (Dempster et al., 1977). Whereas the regression method imputes missing observations (using least squares criteria with no distributional assumptions), the EM algorithm imputes sufficient statistics using maximum likelihood estimation (under an assumption of multivariate normality). Estimates of the means, variances, and covariances obtained from the incomplete data matrix are used to impute the missing data. Then, using the observed and imputed missing data, the parameter estimates are updated and new estimates of the missing data are imputed. The procedure iterates until the parameter estimates stabilize. Results obtained from the EM algorithm are very similar to those obtained from iterated regression (Beale & Little, 1975). The major difference is that the maximum likelihood approach computes a (usually) small correction to the variances and covariances (during the maximization step of the algorithm).

In general applications, IMPUTE assumes that the rows of the examinee (n) $\times$ rater (p) data matrix are independent p-variate normal random vectors with the parameter set consisting of a mean vector and variance/covariance matrix. IMPUTE also assumes that the pattern of missing data (i.e., the mechanism by which the data are missing) can be ignored (Little & Rubin, 1987). One way to insure this is to randomly assign examinees to raters. In the context of essay and oral examinations, for example, raters are usually assigned to candidates more or less at random. However, strict randomization is not necessary. What is necessary is that the probability that an observation is missing does not depend on rater or examinee characteristics that might have influenced the rating. For example, if low ability examinees 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.

Maximum likelihood estimates of the underlying parameters are obtained through the EM algorithm (Dempster et al., 1977). The theory needed to conduct statistical inferences on the parameters is given in Orchard and Woodbury (1972). After the parameter estimates are obtained, the missing data are imputed by finding the expected value of the missing data, given the observed data and these point estimates of the parameters. In the current application, the normality assumption is not required because the emphasis is on imputing missing observations and not on making inferences about the underlying parameters.

Method

Overview

The purpose of this experiment was to investigate the behavior of alternative methods for estimating examinees' true scores under varying conditions of the number of examinees, rater bias, and amount of missing data. The four methods were OLS, WLS, IMPUTE, and NOTHING. Two levels of number of examinees were simulated, $N = 50$ and $N = 100$, in conjunction with two levels of incomplete data—50% incomplete and 75% incomplete. The levels of rater bias were induced to have either high variability (with a range of -2 to $+2$) or low variability (with a range of -1 to $+1$).

Complete rating data were simulated according to this design. The data were made incomplete by randomly deleting specified proportions of the ratings and then subjecting the incomplete rating data to each of the four methods. The estimates of the examinees' true scores obtained from each method were then compared to the known true scores.

Data Simulation

Data were simulated to provide ratings typically obtained from performance evaluations conducted

in a variety of settings. Experience and prior research indicated that performance ratings are frequently made using 7-point Likert scales, exhibit rater variances ranging from approximately 1.0 to 2.0 (on a 7-point scale), and possess rater reliabilities ranging from .40 to .60 (e.g., Berk, 1986; King et al., 1980; Neufeld & Norman, 1985; Rothstein, 1990). Therefore, data were simulated using a 7-point rating scale, and the average level of rater reliability was specified to be a conservative .47 for a single rater.

Each simulated dataset consisted of eight raters. A subset of the raters was assigned to evaluate a subset of the 50 or 100 candidates, such that the design was either 50% incomplete ($p^* = 4$) or 75% incomplete ($p^* = 2$). The total number of raters was arbitrarily selected to be $p = 8$ for two reasons. First, this number provided the capability of creating incomplete rating designs that could potentially range from 12.5% incomplete to 75% incomplete (if the constraint that $p^* \geq 2$ is imposed). Second, the use of eight raters, in conjunction with $N = 50$ or $N = 100$ and $p^* = 2$ or $p^* = 4$, resulted in very realistic requirements for the raters. That is, each simulated rater was required to evaluate from 12 to 50 simulated candidates.

Examinee (n) $\times$ rater ($p = 8$) matrices of observed ratings $\mathbf{X}_{(n \times p)}$ were generated using Equation 3. The vector of true scores $\boldsymbol{\tau}_{(n \times 1)}$ was simulated by drawing n independent observations from an N(4,1.2) distribution. The vector of rater biases $\boldsymbol{\delta}'_{(1 \times p)}$ was fixed at one of two levels:

$$\boldsymbol{\delta}'_{\text{HIGH}} = (-2.0\ -1.5\ -1.0\ -.5\ +.5\ +1.0\ +1.5\ +2.0),$$
$$\boldsymbol{\delta}'_{\text{LOW}} = (-1.0\ -.75\ -.5\ -.25\ +.25\ +.5\ +.75\ +1.0).$$

Thus, the standard deviation of the levels of bias were 1.37 (HIGH) and .68 (LOW).

The matrix of random errors $\mathbf{E}_{(n \times p)}$ was simulated by drawing N independent observations from a N(0,1) distribution p times and forming the matrix $\mathbf{E}^*$. $\mathbf{E} = \mathbf{E}^*\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{\Sigma}$ is the $p \times p$ diagonal matrix of rater error variances. The diagonal matrix $\boldsymbol{\Sigma}$ was assigned the following values:

$$\text{Diag}\ (\boldsymbol{\Sigma}) = (1.0\ 1.5\ 1.0\ 2.0\ 2.0\ 1.0\ 1.5\ 1.5).$$

The specified error variances were randomly assigned to raters.

Thus, the rows of $\mathbf{E}$ were iid $N_8(\mathbf{0}, \boldsymbol{\Sigma})$, corresponding to Assumption 1. The matrix $\mathbf{X}$ was formed using Equation 3. In the population (and when every rater evaluated every examinee), the observed data matrices were characterized by the following reliability parameters:

$$\rho_{\bar{x}\bar{x}'} = .87,$$

and

$$\text{Diag}\ (\mathbf{P}) = (.55\ .44\ .55\ .38\ .38\ .55\ .44\ .44).$$

The off-diagonal elements of $\mathbf{P}$ ranged from .38 to .55, with average correlation among raters of .47. The elements of $\mathbf{X}$ were rounded to integers and truncated to conform to the limits of the 1 to 7 rating scale. After rounding and truncating, the average level of rater reliability was equal to .43 and $\rho_{\bar{x}\bar{x}'}$ was equal to .86 (averaged across all simulated datasets). Thus, rounding and truncating did not appreciably affect reliability.

Design and Procedure

The variables in the simulation study were method (OLS, WLS, IMPUTE, NOTHING), amount of missing data (50%, 75%), rater bias ($\boldsymbol{\delta}' = \boldsymbol{\delta}'_{\text{LOW}}, \boldsymbol{\delta}' = \boldsymbol{\delta}'_{\text{HIGH}}$), and number of examinees ($N = 50$, $N = 100$). The levels of the first two factors were repeated measures; units of analysis were sets of rating data nested in levels of the last two factors. For each unique combination of the levels of the nested factors, 30 complete data matrices were simulated. The total number of simulated data matrices was, therefore, $2 \times 2 \times 30 = 120$.

For every simulated complete dataset, the following procedure was implemented. From the complete dataset, specified proportions of the data were deleted. First, 50% of the data were deleted

so that each examinee was rated by four of the eight raters. (For each examinee, four observations were drawn from a uniform distribution on the integers 1 to 8 until four unique integers were obtained. The ratings corresponding to these four raters were deleted.) The four methods of estimating the τ_i were applied, and the accuracy of the estimated true scores for each method was assessed based on the index of accuracy described below. 75% of the data were then deleted from the complete data so that each examinee was rated by two of the eight raters. The four measures were applied and the index of accuracy calculated. Thus, the unit of analysis was the simulated dataset, with each experimental condition consisting of 30 $N \times p$ sets of simulated rating data.

The index of accuracy employed was the root-mean-squared-error (RMSE), which is the square root of the average squared difference between the true scores estimated by a given method and the known true scores. For purposes of testing all hypotheses, RMSE was transformed to ln(RMSE). This transformation permitted RMSE to satisfy the standard normality assumptions required for analysis of variance (ANOVA) procedures.

The levels of the two repeated measures factors were treated as multivariate observations in the subsequent analysis (i.e., each combination of the amount of missing data and method was viewed as a separate dependent variable). Thus, the design was a 2 (Rater Bias) $\times$ 2 (Number of Examinees) multivariate ANOVA. Full rank parameterization was used to test the relevant hypotheses using sets of linear contrasts.

Results

Table 1 reports the mean RMSE for each of the four methods and for the two levels of amount of missing data, across the 30 replications within each level of the two independent factors—rater bias (bias) and number of examinees. The effect of N does not appear to be large for any dependent variable. The effect of the bias factor is important only for NOTHING (for both 50% and 75% missing data). For example, at 50% missing data for NOTHING, the mean RMSE was .630 at low bias and .739 at high bias. Univariate ANOVAs on each dependent variable confirmed the conclusions that the effect of N was not important and that the effect of the bias factor was important for (i.e., detrimental to) only the method NOTHING. None of the Bias $\times$ N interaction effects were statistically significant ($\alpha = .01$).

Table 1
Mean RMSE Between Estimated and True Scores for Two Levels of Missing Data (50% and 75%) and for the Four Methods—OLS, WLS, IMPUTE (IMP), and NOTHING (NOT), Each With 30 Replications Per Cell

Bias and N	50%				75%			
	OLS	WLS	IMP	NOT	OLS	WLS	IMP	NOT
Low								
50	.573	.579	.590	.611	.854	.867	.792	.918
100	.600	.598	.590	.648	.855	.856	.748	.933
High								
50	.589	.590	.609	.733	.870	.879	.809	1.136
100	.581	.579	.586	.744	.824	.822	.751	1.141

Table 2 reports summary statistics from the multivariate ANOVA for the main effects and interaction effects involving the method factor. The three- and four-way interactions were not statistically significant. There are three statistically significant two-way interaction effects [Amount of Missing Data (Missing) $\times$ Method, Method $\times$ Bias, and Method $\times$ N] reported in Table 2. Because of the

Table 2
ANOVA Results for ln(RMSE)
for Selected Hypotheses

Effect	F(3,114)
Method	595.5*
Method × Bias	147.0*
Low Bias	88.0*
High Bias	655.0*
Method × *N*	12.5*
N = 50	236.8*
N = 100	371.6*
Method × Bias × *N*	1.1
Missing × Method	47.9*
Missing = 50%[a]	593.9*
Missing = 75%	362.1*
Missing × Method × Bias	3.0
Missing × Method × *N*	.5
Missing × Method × Bias × *N*	.3

*Significant at $p < .01$.
[a]Hypothesis tested was mean of OLS, WLS, and IMPUTE equal to NOTHING. Degrees of freedom for this test were 1,116.

relative magnitude of these two-way interactions, subsequent discussion will focus on these effects and ignore the large main effect for Method. For purposes of testing the significance of each interaction effect, values of RMSE were collapsed across levels of factors not included in that particular effect. For example, the Missing × Method interaction effect was averaged across levels of bias and *N*. The Missing × Method, Method × Bias, and Method × *N* interactions are graphically depicted in Figure 1.

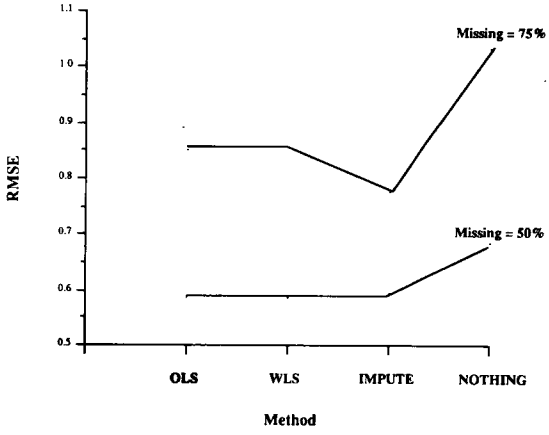
The Missing × Method interaction (Figure 1a) appears to be due primarily to the methods IMPUTE and NOTHING. At 50% missing data, OLS, WLS, and IMPUTE all appear to perform equally well and better than NOTHING. The follow-up test of the hypothesis that the average of the means of the OLS, WLS, and IMPUTE methods was equal to the mean of the method NOTHING, reported in Table 2, supports this conclusion. At 75% missing, NOTHING appears to perform worse, relative to the other methods, whereas IMPUTE appears to perform more accurately than OLS and WLS. The follow-up ANOVA test, reported in Table 2, rejected the hypothesis of the equality of the means across the four methods. Scheffé (1959) tests of all possible pairwise comparisons supported the conclusions that IMPUTE performed better than OLS and WLS and that all three methods performed better than NOTHING.

Figure 1b presents the Method × Bias interaction (averaged across amount of missing data and *N*). This interaction appears to be due primarily to the method NOTHING, which performs more poorly as the level of bias increases from low to high. Follow-up tests conducted at each level of the bias factor rejected the hypothesis of the equality of the means across the four methods. At each level of bias, Scheffé tests indicated that IMPUTE performed better than OLS and WLS, and that all three methods performed better than NOTHING.

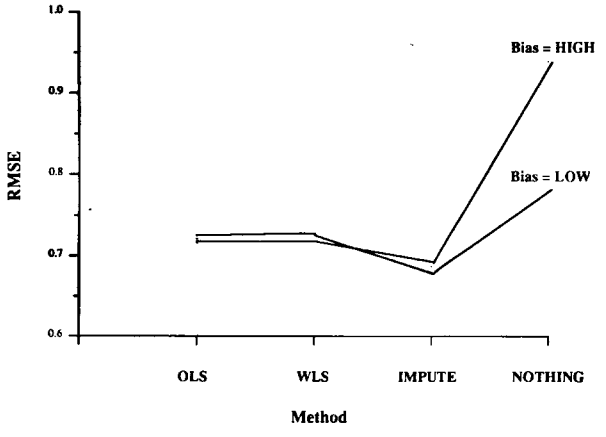
Figure 1c shows the Method × *N* interaction effect. Relative to OLS and WLS, IMPUTE performed better and the NOTHING method performed worse as the number of examinees increased from 50 to 100. The follow-up tests at each level of *N*, reported in Table 2, rejected the equality of the means across the four methods. Once again, Scheffé tests indicated that IMPUTE performed better than

Figure 1
Mean RMSE for Significant Interactions Involving the Correction Methods

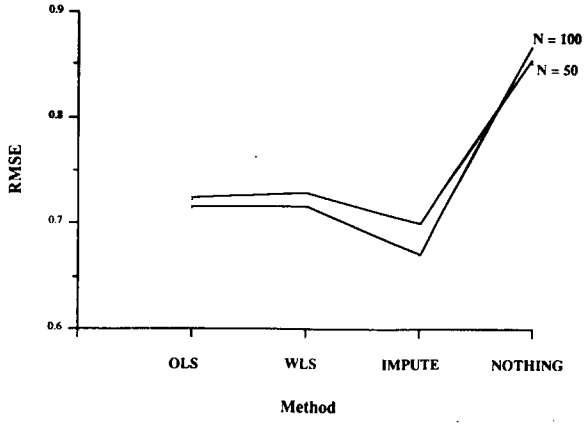
a. Missing Data × Method Interaction



b. Bias × Method Interaction



c. *N* × Method Interaction



OLS and WLS, and that all three methods were better than NOTHING with both $N = 50$ and $N = 100$.

Discussion

The three methods of correcting for rater bias effects were all superior to simply averaging the observed ratings. For example, when averaged across all conditions, the value of RMSE for the method NOTHING was approximately 24% higher than the RMSE for the method IMPUTE. The superiority of the correction methods relative to the method NOTHING increased as the amount of missing data increased and as the amount of rater bias increased. The improvements in accuracy that can be realized by employing a correction method can be quite substantial. Under the least favorable conditions (high bias, 75% missing data), RMSE for NOTHING was 40% (for $N = 50$) to 52% (for $N = 100$) higher than the RMSE for IMPUTE.

IMPUTE yielded better results than OLS and WLS with 75% missing data. With 50% missing data averaged across levels of rater bias and N, all three correction methods were approximately equally effective. The two least-squares regression procedures were found to be equally effective across all conditions investigated. However, WLS can be expected to become more effective than OLS as differences among rater error variances increase. Wilson (1988) demonstrated the WLS method to be far more accurate than the OLS method in an artificial situation in which two of eight raters had severely elevated error variances.

The results of simulation studies such as this are limited by the assumptions of the models used to generate and analyze the data. The simulated rating data were generated from a linear model. Similarly, the correction procedures are based on linear models. Consequently, there was a degree of predetermined model fit engineered into the study. If, in practical settings, the rating data did not conform to a linear model, then the correction procedures investigated would not be as effective. This should not be a problem, however, because all procedures could still be effectively used by subjecting the rating data to the necessary transformation (Cason & Cason, 1984; de Gruijter, 1984). Clearly, in practical applications of the correction procedures, it is essential to evaluate the degree to which the data could be modeled by a linear function.

It is also important to acknowledge that the data were generated from a model that assumed that the error terms and the true scores across examinees were normally distributed. However, these assumptions are not strictly required for any of the methods investigated. IMPUTE uses the EM algorithm to obtain maximum likelihood estimates of the mean and covariance matrix of the assumed multivariate normal distribution in the presence of missing data. This algorithm is noted for being slow to converge. In the present application with 75% missing data, 40 to 60 iterations were required. The relative superiority of IMPUTE may not hold for data that follow distributions markedly different from multivariate normal. Although this study used random elimination to create the missing data, none of the three correction methods requires strict randomization. It should be noted that random assignment of raters to examinees is extremely important if true scores are to be estimated by averaging the observed ratings (i.e., using the method NOTHING).

Finally, it is important to recognize that the three correction methods investigated all attempt to correct for relative rater bias effects in the estimation of examinee true scores. If all raters overestimate or underestimate the trait in question, then the estimated examinee true scores will still reflect this absolute bias. OLS and WLS concentrate on correcting for relative rater bias and are restricted to τ-equivalent or parallel measurement models. IMPUTE can be extended to congeneric measurement models.

The results indicate that statistically correcting for rater effects might result in substantial improvements in the accuracy of performance ratings. Because a powerful experimental design was

employed in this simulation study, the p values and/or F ratios should not be interpreted as indicators of the magnitude of treatment effects. Whether differences among the approaches to correct for rater bias effects are important and meaningful depends on the particular application.

Although the present study used RMSE as the criterion variable, it may be useful to investigate other indices of effectiveness. For example, correlations between estimated and actual true scores may be a viable criterion, especially if the results are intended to generalize to personnel selection problems for which only an accurate rank ordering of candidates is required. In evaluation settings involving dichotomous decisions, it would be important to evaluate the accuracy of various methods in terms of false-positive and false-negative error rates and the consistency of the decisions across correction methods (Raymond et al., 1991).

Although the aforementioned limitations urge a cautious interpretation, the results of this study are certainly encouraging: They demonstrate that errors of leniency/stringency can be modeled and partially controlled, thereby increasing the accuracy of performance ratings. Because performance ratings are both pervasive and generally unreliable, any method that shows some promise of improving the quality of rating data is worthy of careful consideration. However, further research is needed prior to implementing any of the correction procedures on an operational basis.

The results of this study could be extended to at least three classes of research. First, the correction methods could be expanded to correspond to more complex rating situations. In some settings, all candidates are rated on multiple behavioral dimensions. In other settings, the rating dimensions or testing materials may vary across candidates and/or raters. For example, Braun (1988) used a linear model to adjust for multiple effects (e.g., rater, topic, day) in a study of essay ratings.

Second, the experimental conditions could be expanded to include more factors or levels within factors. One possible extension to the present study would be to vary the degree of model fit. This could be done by increasing or decreasing the interrater correlations. As the interrater correlations increase (i.e., as error variances decrease), the correction methods become more effective, relative to averaging the observed ratings. It can be shown that as error variances approach 0, estimated true scores obtained from the three correction methods approach the actual true scores. In addition, it would be informative to investigate the utility of the correction procedures under varying levels of rater bias or under conditions of higher levels of missing data. For example, in many assessment settings levels of incomplete data can range from 80% to 90% incomplete (e.g., Cason & Cason, 1984; Raymond et al., 1991). In such instances, the need to correct for rater effects is paramount; however, parameter estimation becomes more tenuous due to the sparseness of the data.

A third way to extend the present study would be to extend the measurement model in Equation 1 to allow raters, or examinees, to be a random effect (i.e., assume that the δ_j are sampled from some underlying distribution). Much additional work is needed in order to determine the differential effectiveness of various correction methods and the conditions that moderate their effectiveness.

References

Beale, E. M. L., & Little, R. J. A. (1975). Missing data in multivariate analysis. *Journal of the Royal Statistical Society, 37,* (Series B), 129–145.

Berk, R. A. (Ed.) (1986). *Performance assessment.* Baltimore: Johns Hopkins University Press.

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

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, 22,* (Series B), 302–307.

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

de Gruijter, D. N. M. (1984). Two simple models for rater effects. *Applied Psychological Measurement, 8,* 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, 39,* (Series B), 1–38.

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

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

Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data. *Biometrika, 27,* 783–832.

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.

Little, R. (1976). Inferences about means from incomplete multivariate data. *Biometrika, 63,* 593–604.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data.* New York: Wiley.

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

Neufeld, V. R., & Norman, G. R. (Eds.). (1985). *Assessing clinical competence.* New York: Springer.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1–18.

Orchard, T., & Woodbury, M. (1972). A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Problems, Vol 1,* 697–715.

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,* 1326.

Raymond, M. R., Webb, L. C., & Houston, W. M. (1991). Correcting performance rating errors in oral examinations. *Evaluation and the Health Professions, 14,* 100–122.

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

Rubin, D. (1976). Inferences and missing data. *Biometrika, 63,* 581–592.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

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

Acknowledgments

The authors thank Ron Cope, Brad Hanson, Deb Harris, Mike Kane, Richard Sawyer, the Editor, and two anonymous reviewers for their insightful comments.

Author's Address

Send requests for reprints or further information to Walter M. Houston, American College Testing, P.O. Box 168, Iowa City IA 52243, U.S.A.