ACT Research Report Series 91-8

Least-Squares Models to Correct for Rater Effects in Performance Assessment

Mark R. Raymond and Chockalingam Viswesvaran

December 1991

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

Least-Squares Models to Correct for Rater Effects in Performance Assessment

Mark R. Raymond and Chockalingam Viswesvaran

American College Testing, Iowa City, Iowa

Abstract

This study illustrates the use of three least-squares models to control for rater effects in performance evaluation: ordinary least squares (OLS); weighted least squares (WLS); and ordinary least squares subsequent to applying a logistic transformation to observed ratings (LOG-OLS). The three models were applied to ratings obtained from four administrations of an oral examination required for certification in a medical specialty. For any single administration, there were 40 raters and approximately 115 candidates, and each candidate was rated by four raters. R^2 values for the OLS and LOG-OLS models were comparable, while R^2s were substantially higher for the WLS model. The results indicated that raters exhibited significant amounts of leniency error, and that application of the least-squares models would change the pass-fail status of approximately 7% to 9% of the candidates. Ratings adjusted by the models demonstrated higher reliability and correlated slightly higher than observed ratings with the scores on a written examination.

Least Squares Models to Correct for Rater Effects in Performance Assessment

Oral examinations have long played an important role in evaluating an individual's readiness for professional practice. In many professions, the oral survives as the final exam among many assessment hurdles. Of the 23 specialty boards that are members of the American Board of Medical Specialties (ABMS), 15 include an oral examination as a requirement for certification (ABMS, 1990). Psychology boards in numerous states include the successful completion of an oral examination as part of their licensure requirements (Hill, 1984). Oral examinations have achieved prominence in other areas, as well. For example, many, if not most, doctoral programs require an oral examination as part of the criteria for graduation. Interviews conducted for the purpose of making admission decisions may, in many instances, represent nothing more than an oral exam, both in terms of the methods used to elicit responses, and in terms of the constructs that are evaluated. Given the recent interest in alternative methods of assessment (Bray & Byham, 1991; Linn, Baker, & Dunbar, 1991), one might expect an increase in the use of oral examinations.

Supporters of oral examinations suggest that orals provide a method for evaluating psychological constructs that often elude conventional written examinations. The interactive nature of an oral can permit an examiner to evaluate a candidate's depth of knowledge in a particular area, skill at interpreting and evaluating the utility of diagnostic tests (e.g., slides, radiographs, other images), and skill at evaluating and selecting alternative methods of managing a particular case (Hill, 1984; Levine & McGuire, 1970; Muzzin & Hart, 1985; Watson, 1984). Interpersonal behaviors and communication skills can also be assessed

during an oral examination. In fact, oral examinations in some professions consist of actual work samples that require examinees to demonstrate proficiency with surgical instruments or to interact with patients (Levine & McGuire, 1970; Small, 1982). The job-relatedness of such exams is direct and obvious.

Both supporters and critics of oral examinations acknowledge the limitations: they are expensive, subjective, unreliable, and may represent nothing more than a rite of passage (e.g., Muzzin & Hart, 1985; Watson, 1984). Perhaps the most serious criticism of oral examinations concerns the levels of reliability that are typically observed. Although interrater reliabilities have approached .80 in selected studies (O'Donohue & Wergin, 1978), it is also common to see interrater reliability coefficients in the .20s and .30s (Barnes & Pressey, 1929; Hubbard, 1971). Consequently, many oral examination programs make use of multiple raters in an effort to reduce the influence of measurement error and enhance reliability.

Ratings of performance are susceptible to two general classes of measurement error: random and systematic. The random error component is what is typically regarded as rater unreliability. If all candidates in a group are evaluated by the same raters, then the reliability coefficient can be estimated by:

$$\rho^2 = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_e^2/n_r} \qquad (1)$$

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; Shavelson, Webb, & Rowley, 1989).

Expression (1) applies to complete rating designs. However, most performance ratings utilize a design in which each candidate is evaluated by a subset of raters. If an incomplete rating design is used, then the reliability estimate must acknowledge the error due to the fact that raters may be differentially lenient or harsh in their ratings. This systematic error is typically referred to as leniency error. The reliability for many incomplete rating designs can be computed by:

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

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 equations (1) and (2) will provide equivalent results. This will seldom be the case, however. Given that performance ratings are frequently used to make important decisions about an individual's career, any effort to reduce the impact of measurement error may have social utility. Rater training represents one common strategy for minimizing rating errors. However, training programs are time consuming and costly, and their effectiveness is questionable: although some studies show positive effects, others show no effect, while still others have demonstrated a negative effect (e.g., Bernardin, 1978; Bernardin & Pence, 1980; Borman, 1979; Hedge & Kavanaugh, 1988; King, 1983; Lunz, Wright, & Linacre, 1990; Trier, 1983). A variety of different statistical models to correct for rating errors have also been proposed in the literature, including models based on item-response theory (de Gruijter, 1984; Lunz et al., 1990), multivariate analysis from incomplete data (Houston, Raymond, & Svec, in press), least-squares

regression (Braun, 1988; de Gruijter, 1984; Raymond, Webb, & Houston, 1991; Wilson, 1988), and other models (Cason & Cason, 1984). Prior research suggests that the use of statistical models can result in considerable reductions in measurement error (Braun, 1988; Houston et al., in press).

The purpose of this article is to describe and illustrate the use of a simple and flexible statistical model that can be applied to incomplete rating designs in order to identify and correct for leniency error. Actual rating data are used, obtained from four administrations of a certification examination in a medical specialty. The magnitude of the adjustments, their effect on pass-fail decisions, and their impact on correlations with the scores on a written examination are estimated. The next section presents three variations of a least-squares model: ordinary least squares (OLS), weighted least squares (WLS), and OLS applied to logit-transformed ratings (LOG-OLS). Subsequent sections describe the results of applying the models to ratings obtained from four independent administrations of an oral examination in a medical specialty.

Correction Methods

Ordinary Least Squares (OLS)

Regression-based procedures to identify and correct for rater effects have been proposed by 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. The model can be represented as follows:

$$y_{ij} = \alpha_i + \beta_j + e_{ij} \quad (3)$$

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

α_i is the true rating 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 or 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 leniency effects (i.e., is an unbiased estimator of α_i). 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 model in expression (3) can be estimated through least-squares regression. Let K be the total number of observed ratings assigned by p raters to n candidates. Then the matrix formulation for the OLS model is:

$$y = X \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + e \quad (4)$$

where y is a (K x 1) vector of observed ratings,

X is a (K x (n + p - 1)) design matrix,

α is an (n x 1) vector of true ratings for candidates,

β is a (p - 1) x 1 vector of rater bias indices, and

e is an (K x 1) vector of random errors.

The design matrix, X, consists of n + p - 1 columns; the column for the last rater is dropped to avoid a linear dependency in the columns of X. Because X is of full-column rank, the parameters can be estimated by any standard multiple-regression algorithm. The appendix presents an example of an incomplete rating matrix for a sample of five candidates and three raters, as well as 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 $\begin{bmatrix} a \\ b \end{bmatrix} = (X'X)^{-1} X'y$. As the vector in 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.

Weighted Least Squares (WLS)

The OLS procedure provides an unbiased estimate of the vector of true ratings. If, however, the consistency of scoring varies across raters (i.e., if correlations among raters vary considerably), then the usual regression assumption of equal error variances across all candidates and raters is violated. The statistical consequence is that the variances of the

parameter estimates will be inflated (Draper & Smith, 1981). 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. 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 squares analysis to obtain revised estimates of the candidates' true scores. The WLS parameter estimates are given by:

$$\begin{bmatrix} a \\ b \end{bmatrix} = (X'WX)^{-1} X'Wy \qquad (5)$$

where W is a K by K diagonal matrix of weights, with the elements of W corresponding to the value of $1/MSR_j$ for each rater.

OLS Applied to Logit-Transformed Ratings

The presence of floor or ceiling effects in the observed ratings will compress individual differences at the two ends of the rating scales. Such effects can be compensated, in part, through the use of a nonlinear transformation. The most commonly employed nonlinear transformations are the probit and logit transformations (Cohen & Cohen, 1983; Lord, 1980; Wright & Stone, 1979). Although the logit and probit transformations achieve similar

outcomes, the distribution is stretched more with the logit transformation (Cohen & Cohen, 1983).

For the present study, we assessed both the rater effect and any floor/ceiling effects by also applying the OLS model to logit-transformed ratings. The logit transformation was effected as .5*ln(P/(1-P)) where P was the ratio of the observed to maximum score possible. The LOG-OLS model assumes that observed ratings are compressed at the two ends of the scale in that the ability-observed rating relationship is not linear but takes the form of an ogive and corrects for such compression first, before correcting for rater effects. The mathematical tractability of the logit transformation (Cohen & Cohen, 1983; Lord, 1980) is one reason for its popularity over probit models in psychometrics. Also, since the logistic function approaches its asymptotes less rapidly than the probit model, aberrant ratings do lesser harm to the logistic model than to the probit model (Lord, 1980). That is, the logit transformation is less sensitive to random error than the probit transformation. The impact of the transformation can be evaluated by comparing the model fit of the LOG-OLS model to the fit of the OLS model with untransformed ratings.

Method

Rating Data

Data were comprised of observed ratings from four operational administrations of an oral certification examination administered by a medical specialty board. All candidates took a written multiple-choice exam in that specialty prior to participating in the oral examination. Only those candidates whose written scores fell within the middle range of the distribution were called to the oral exam. That is, for any given year, approximately 30%

of the candidates who took the written exam also took the oral; those with high written scores were exempted, while those with low written scores were ineligible.

The oral examination consists of 16 clinical cases, complete with the results of lab tests, x-rays, pathology slides, and other diagnostic information. The cases are organized into four subspecialties. Within each subspecialty, four separate clinical cases are presented to each candidate. Although all candidates receive the same cases, candidates are not evaluated by the same raters. Specifically, each candidate is evaluated by four raters, and each rater presents the same four cases to all candidates evaluated. Raters are nested within topics, which are crossed with candidates. Candidates are given ratings on three dimensions (factual recall; interpretation of data; clinical problem-solving) by each rater. The three ratings are on a Likert-type rating scale ranging from 1 to 12 and serve as indices of performance in each subspecialty. The sum of the three ratings is obtained as an index of overall performance, thus producing a rating scale with a possible range from 3 to 36 for each subspecialty (correlations in the .80s among the three dimensions support this simple combination of ratings). A candidate's observed final rating is the mean rating obtained over the four raters (in four subspecialties) who evaluated that candidate. Therefore, the final observed rating scale ranges from a low of 3 to a high of 36 with one-fourth point intervals.

A total of 456 candidates was examined over a four-year period (1987 to 1990). In any one year, approximately 115 candidates were examined by 40 raters; each candidate was examined by four raters, and each rater examined from 7 to 14 candidates with an average of 11.5 candidates. In any given year, the entire data matrix (40 raters by 115 candidates = 4,600 possible observations) was about 10% complete. Prior to the administration of the

actual examination in any given year, the raters participated in a practice session during which they rated videotapes of candidates and received feedback regarding the similarity of their ratings to those of their peers.

Procedures

Data from the four years were subjected to three variations of least-squares regression: OLS, WLS, and LOG-OLS. Data resulting from application of the least-squares models were subjected to four types of analyses. First, we evaluated the fit of the three models to determine if the more complex models (WLS and LOG-OLS) were more accurate than the simpler model (OLS). Model fit was evaluated by examining residual plots and computing the proportion of variance in observed ratings attributable to the candidate effect and rater effect (R^2). Second, the model-adjusted ratings (α) were compared to observed ratings and subjected to various descriptive analyses in order to gain an understanding of the magnitude of the adjustments. If the values of α are essentially identical to the observed ratings, then there would be little reason to use the models. Third, since the oral examination is used to certify physicians in specialty practice, we evaluated the impact of the model-adjusted scores on pass/fail decisions. The analyses were conducted to determine the decision consistency between model-adjusted ratings and observed ratings. Fourth, we conducted analyses to determine if the models altered the intercorrelations among ratings (i.e., interrater reliability) and correlations of oral ratings with the scores on a written examination. Measurement error attenuates correlations among variables; therefore, we expected that the model-adjusted ratings would correlate more highly than observed ratings with each other and with scores on a written examination. The data from the four years

were analyzed separately to assess the feasibility of applying the least-squares models in a real world setting.

Results

Model Fit

Table 1 provides the R^2 values for the three models for each of the four years analyzed. Comparisons of the R^2 values indicate a better fit for the WLS model for the four data sets. This result occurs because there are differences in the error variance of ratings among raters, and this source of variation is modelled by the WLS model. The R^2s in Table 1 also indicate that the logistic transformation did not result in improved model fit for the four years considered. In addition, an analysis of the residual plots indicated that a logistic transformation was unnecessary.

Table 1

R^2 for Three Least Squares Models

Year	OLS	WLS	LOG-OLS
1	.54	.57	.53
2	.59	.70	.59
3	.52	.60	.52
4	.54	.65	.56

The differences in fit among the three models can be regarded as evidence that differences in rating variability across raters introduce more error than floor or ceiling effects for the present data. The lack of floor or ceiling effects could be due to the fact that the scale is quite broad, ranging from 3 to 36. Since logit transformation of the observed, untransformed ratings resulted in no appreciable increase in model fit for any of the four

years, further discussions will be limited to the OLS and WLS models using the untransformed ratings only.

The rater effect was found to be statistically significant ($p < .01$) for all four years, indicating the presence of leniency error. The variance components and reliability coefficients are presented in Table 2. As Table 2 indicates, the rater effect is appreciable—about one-half the magnitude of the candidate effect for any year. The reliability of four ratings (ρ^2) ranges from .48 to .56 with an average of .52.

Table 2

Variance Components and Reliability Coefficients

Year	σ_i^2	σ_r^2	σ_ε^2	ρ^2
1	8.36	3.23	25.01	.54
2	8.70	5.36	22.15	.56
3	7.05	4.13	26.36	.48
4	8.08	4.45	29.92	.49

For the OLS model, estimated b_js for individual raters ranged from -6.15 to 7.24 for the four years; the average (over four years) minimum and maximum values were -5.22 and 6.03. The mean of the absolute values of b_j ranged from 1.87 to 2.41, with an overall mean across all four years of about 2.11. That is, any rater drawn at random would be expected to be biased by a little more than two points.

The levels of bias for the WLS model were very similar to those for the OLS model. However, the WLS model weights each examiner's ratings by the reciprocal of their mean squared residual ($1/MSR_j$). The values of MSR_j for the OLS model did exhibit considerable variability, as suggested by the increased model fit for the WLS model. Specifically, the average (over four years) minimum value of MSR_j was about 4.26, and the maximum value

was 43.12. The mean value of MSR_j over the four years was 17.22. It is important to note that these values are squared residuals.

Properties of Adjusted Ratings

Descriptive statistics were computed for each of the four years in order to compare the distributional properties of the observed ratings and model-adjusted ratings. The distributional properties of the observed and model-adjusted ratings were very similar for all years and are provided in Table 3.

Table 3

Descriptive Statistics for Observed and Model-Adjusted Ratings

Year	Rating Type	N	Mean	S.D.	Min.	Max.
1	Obs.	129	22.65	3.78	13.25	31.25
	OLS		22.66	4.03	13.19	32.57
	WLS		22.65	3.93	12.90	32.20
2	Obs.	114	22.73	3.89	13.25	30.00
	OLS		22.81	3.85	13.19	30.47
	WLS		22.81	3.82	11.92	30.03
3	Obs.	121	22.29	3.81	11.00	31.50
	OLS		22.29	3.73	11.85	29.89
	WLS		22.29	3.73	11.90	30.39
4	Obs.	92	21.10	4.12	9.50	31.50
	OLS		21.13	4.07	10.58	31.52
	WLS		21.14	4.10	10.82	30.77

The correlations among the model-adjusted ratings and the observed ratings ranged from .90 to .98. These high correlations indicate that the rank ordering of candidates is not affected significantly by the model adjustments. As expected, the correlation between the WLS-adjusted ratings and observed ratings was lower than the correlation between OLS-

adjusted ratings and the observed ratings. This is consistent with the fact that the WLS model results in a greater adjustment of ratings by differential weighting of raters.

For norm-referenced selection decisions, the correlations among model-adjusted and observed ratings would be of primary interest; that is, one would be concerned primarily with rank order information. Within the context of a domain-referenced examination that utilizes an absolute cut-off score, rank order information is insufficient for evaluating the potential impact of the OLS and WLS models. In such applications, the magnitude of adjustments is of critical importance. To examine the magnitude of the adjustments imposed by the OLS and WLS models, a difference score was computed for each candidate by subtracting the model-adjusted ratings (based on both OLS and WLS) from the observed ratings.

As indicated in Table 4, the magnitude of the adjustments exceeded ± 3.0 points and approached ± 5.0 points for year 4. As expected, the magnitude of the adjustments was greater for the WLS model than for the OLS model. The average adjustment over the four years for the OLS-adjusted ratings was 1.01, while the average adjustment for the WLS models was 1.26. One way to gauge the relative magnitude of these adjustments is to compare them to the standard deviations of the ratings (about 3.90). The magnitude of the adjustments is about .26 SD units for the OLS model and .32 SD units for the WLS model. A more conservative index of adjustment can be obtained by using the range of the scale as the basis for comparison. Using this index, the magnitude of the typical OLS adjustment is about 5.2% of the rating scale, and the typical WLS adjustment is about 6.5% of the rating scale. Adjustments of this magnitude may affect the pass/fail decisions of borderline candidates.

Table 4

Descriptive Statistics for Differences Between Model-Adjusted Ratings and Observed Ratings

Year	Model	Difference Scores		
		Min	Max	Mean (Absolute)
1	OLS	-3.06	3.04	0.89
	WLS	-3.27	4.16	1.11
2	OLS	-2.94	4.14	1.11
	WLS	-4.47	4.53	1.16
3	OLS	-3.91	3.40	0.97
	WLS	-3.53	4.89	1.32
4	OLS	-3.25	4.09	1.08
	WLS	-4.62	4.91	1.47

Impact on Pass/Fail Decisions

The consistency of pass/fail decisions for observed ratings and model-adjusted ratings was evaluated. For each year, several artificial but realistic cut-off points were imposed. The pass points were selected so that the pass rates would range from 70% to 90%—pass rates typical of a certification exam in the health professions. Multiple cut-off points were used so that the impact of model adjustments on the pass/fail decisions could be examined over a wider range of the distribution of ratings. This will help ensure that conclusions about the stability of the impact of adjustments on pass/fail decisions are not due to local abnormalities in the distribution. The effects of the adjustments on pass/fail decisions are summarized in Table 5.

The pass rates for each cut-off score resulting from the use of adjusted ratings are presented, along with the pass rates resulting from the use of observed ratings. The pass

rates were similar for all three types of ratings. This result is consistent with the fact that the least-squares models did not significantly alter the overall distribution of ratings. Also presented in Table 5 are the rates of disagreement for the pass/fail decisions produced by the OLS and WLS models, as compared with pass/fail decisions for the observed ratings. The disagreement rates range from 2.5% to 14.1%, and are consistently higher for the WLS model. The average rate of disagreement is 6.8% for the OLS model and 8.5% for the WLS model. Further analysis also revealed that the percentage of decisions changing from pass to fail was about the same as the percentage of decisions changing from fail to pass.

It can be seen that the disagreement rates increase as the passing point approaches the mean. Deviations from this trend reflect local characteristics of the rating distribution. The trend for disagreement rates to increase as the passing point approaches the mean is due to the fact that the opportunity for chance agreement decreases toward the midpoint of the distribution. Although the Kappa coefficient (Cohen, 1960) corrects for chance agreement, and would eliminate this trend, its use in the present context seemed to be more of a hinderance than an asset.

Improvements in Validity

As the purpose of the least-squares adjustments is to reduce measurement error, it seemed important to empirically determine the extent to which rater reliability is actually improved. This was done by obtaining the variance components of the model-adjusted ratings and computing the reliability coefficients. Stanley (1961) has noted that rater agreement can be interpreted as evidence supporting the construct validity of ratings. The variance component analyses produced two noteworthy findings. First, the variance

Table 5

Consistency of Pass/Fail Decisions Based on Observed and Adjusted Ratings for Selected Pass Points

		Percent Pass			Percent Disagreement With Observed	
Year	Pass Point	Observed	OLS	WLS	OLS	WLS
1	17.0	92.3	89.9	90.7	5.4	4.7
	17.5	89.9	89.2	89.2	3.9	5.4
	18.0	87.6	86.1	88.4	4.7	7.0
	18.5	87.6	84.5	86.1	3.1	6.2
	19.0	83.7	83.7	83.0	4.7	8.5
	19.5	82.2	79.8	77.5	5.4	9.3
	20.0	75.2	76.7	76.0	6.2	8.5
	20.5	72.1	70.5	72.1	6.2	4.7
	21.0	68.2	65.9	66.7	7.0	7.8
2	17.0	91.2	91.2	93.0	3.5	3.5
	17.5	89.5	91.2	91.2	3.5	5.3
	18.0	86.0	88.6	88.6	4.4	6.1
	18.5	85.1	85.1	87.7	7.0	7.9
	19.0	83.3	83.3	85.1	8.8	7.0
	19.5	79.8	78.1	80.7	10.5	7.9
	20.0	76.3	77.2	77.2	9.7	9.7
	20.5	72.8	73.7	75.4	7.9	11.4
	21.0	69.3	71.1	70.2	8.8	11.4
3	17.0	91.7	90.9	91.7	5.8	6.6
	17.5	88.4	88.4	89.3	5.0	9.1
	18.0	82.6	84.3	84.3	3.3	6.6
	18.5	80.2	81.0	82.6	2.5	9.1
	19.0	78.5	78.5	81.0	5.0	9.1
	19.5	76.9	76.0	76.9	5.8	6.6
	20.0	76.9	73.6	71.9	6.6	9.9
	20.5	74.4	69.4	69.4	8.3	11.6
	21.0	69.4	67.8	68.6	8.3	10.8
4	17.0	84.8	83.7	81.5	7.6	9.8
	17.5	83.7	81.5	80.4	6.5	12.0
	18.0	78.3	76.1	79.4	6.5	7.6
	18.5	78.3	71.7	76.1	6.5	4.4
	19.0	75.0	67.4	69.6	9.8	7.6
	19.5	69.6	65.2	67.4	10.9	13.0
	20.0	64.1	59.8	60.9	13.1	14.1
	20.5	57.6	57.6	55.4	10.9	13.0
	21.0	53.3	52.8	51.1	12.0	13.0

component due to raters for the model-adjusted ratings went to zero. This result was expected since the least-squares models statistically remove the systematic variation among raters. Second, the reliability increased from an average of .52 (Table 2) to an average of .63 for the four years.

The validity of the model-adjusted ratings was further examined by computing the correlations between the ratings (observed and model-adjusted) and scores obtained on a written certification examination. As noted earlier, only 30% of all candidates who took the written exam also qualified for the oral examination: high scoring candidates are exempted, and low scoring candidates are disqualified. As this rather severe range restriction substantially depresses the observed correlations, a correction for range restriction was applied.

Table 6

Correlations of OLS, WLS, and Observed Ratings With Written Test Scores[1]

Year	Observed	OLS	WLS
1	.34 (.70)	.33 (.69)	.33 (.69)
2	.36 (.76)	.41 (.81)	.39 (.79)
3	.17 (.45)	.19 (.49)	.20 (.51)
4	.14 (.38)	.21 (.53)	.23 (.56)

Table 6 presents the uncorrected and corrected correlations between written scores and the three sets of ratings. Overall, the model-adjusted ratings correlate more highly with the

[1]The values within parentheses are correlations corrected for range restriction.

scores in the written examination than the observed ratings although the magnitude of the improvement is modest. The average increase in correlations over the four years is .06 for the OLS model and .07 for the WLS model. The results indicate that the model-adjusted ratings are more reliable than observed ratings, and that the increased reliability enhances their relationship with an external criterion.

Discussion

This investigation illustrates the feasibility of applying three variations of least-squares regression to correct for rater errors in performance evaluations. Clearly, the efficacy of the regression methods are a function of the degree to which the models fit the rating data, as well as the magnitude of the rater effect (i.e., leniency error). The rater effect was statistically significant for each of the four years. Although the values of b_j reached nearly ± 5 points, the average level of bias was a little over ± 2 points. R^2 values for the OLS model ranged from .52 to .59 over the four years, while the R^2 values for the WLS ranged from .57 to .70 for the four years. The use of the logit transformation was not necessary for the present data, as suggested by the R^2s for the LOG-OLS models. An analysis of the residual plots confirmed the suitability of the linear model. It is likely that the original rating scales—a series of three 12-point Likert scales—were sufficiently broad to deter floor and ceiling effects. It is certainly possible that the use of narrower scales would have produced detectable floor and ceiling effects.

For the present sets of data, correcting for rater effects made a modest difference in the overall rating received by the typical candidate. The OLS model resulted in about a ± 1-point adjustment on average, although the magnitude of the adjustments for several

candidates fell in the ±3.0-point range. Adjustments based on the WLS model were about 25% larger, with the average adjustment a little more than 1.25 points and several adjustments falling in the 3.0- to 4.0-point range. For both models, a few adjustments exceeded 4.0 points. In the present rating situation, the assignment of four examiners to each candidate helped minimize the negative impact due to the bias of any single rater. That is, lenient and harsh raters tended to cancel one another to some extent. Since the index of rater bias, b_j, was more than two points on average, it is clear that the magnitude of the adjustments imposed by the OLS and WLS models would certainly have been larger had only two or three raters been assigned to each candidate.

As indicated in Table 5, the modest levels of leniency error altered the pass/fail decisions for selected candidates. The results indicated that using OLS-adjusted ratings would result in changed decisions for approximately 7% of the candidates in any single year. Use of the WLS model would alter the decisions for 8% to 9% of the candidates. Is this an important effect? In the context of the present examination, it is. The candidates were physicians seeking board certification in a specialty area. Nearly all of the candidates had completed four years of college, three to four years of medical school, and at least another three years of residency training in this particular medical specialty. Even a few erroneous pass/fail decisions can have significant personal, social, and economic consequences.

Since the data were not perfectly modelled, it is difficult to know whether all of the adjustments to the ratings were appropriate. That is, the least-squares models could have resulted in a candidate's status being incorrectly changed from pass to fail or vice versa. In addition, there are likely to be some candidates whose pass/fail status should have changed but did not. However, the improvements in reliability for all years, as well as the increases

in the correlations with written scores for three of the four years suggest that the adjustments to the ratings imposed by the least-squares models were, in fact, in the appropriate direction for most candidates. That is, the model-adjusted ratings appear to improve the construct validity of the ratings.

The utility of the least-squares models is a direct function of model fit. In particular, the adjustments to the ratings will be larger and more beneficial as the variance component due to error decreases and the variance component due to the rater effect increases. Although the R^2 values in the present study are less than what have been observed for essay ratings (Braun, 1988), it is generally the case that essays can be rated with far more reliability than performance in an oral examination. The R^2 values in the present data are also less than those reported by Cason and Cason (1984), who applied a rather complex, iterative, least-squares model to probit-transformed ratings. In short, the degree of model fit for the present data should be regarded with cautious optimism.

Empirical research on model fit is lacking. Consequently, it is difficult to unambivalently advocate the use of the least-squares models on an operational basis. In the presence of borderline model fit, one is faced with a difficult decision. On one hand, less-than-desirable model fit may discourage the use of model-adjusted ratings. On the other hand, poor model fit may suggest that even the observed ratings are not reliable enough for making important decisions. The choice is not an easy one; more research is needed to establish guidelines. The results of one empirical investigation using numerous sets of simulated rating data clearly supported the use of three models—OLS, WLS, and imputing missing ratings via the E-M algorithm (Houston et al., in press). However, the simulated data contained less error variance than the present data, and the rating designs were more

complete; therefore, the results of that study do not necessarily generalize to the data investigated in the present study. Within the context of essay ratings, Braun (1988) demonstrated that, under certain circumstances, the use of an OLS model could actually result in greater increments in reliability than could be obtained by doubling the number of raters. The least-squares models appear to offer a viable method of reducing the negative consequences of rating errors. Future simulation studies will need to investigate the effectiveness of the least-squares models using data that are consistent with the levels of reliability typically observed in oral examinations (Muzzin & Hart, 1985) and in work settings (Rothstein, 1990).

In addition to detecting and correcting for rater effects, the least-squares models can generate statistical information useful for describing the psychometric properties of the rating data. For example, in the present study the index MSR_j was used for computing weights for the WLS model. Since MSR_j is inversely related to the correlation of rater j with all other raters, it can be interpreted as an index of rater reliability, rater discrimination, or rater fit. It is parenthetically noted that the OLS and WLS models can be loosely interpreted within an item-response theory framework, whereby the difficulty parameter of a rater is described by b_j, and the slope is a function of MSR_j. The presence or absence of the logit transformation will determine whether the model is linear or conforms to the logistic function.

The least-squares regression models provide the many useful inferential statistics (e.g., tests that $b_j = 0$ for each j) and diagnostic procedures (e.g., residual plots) that are described in selected texts on linear models (e.g., Belsley, Kuh, & Welsch, 1980; Draper & Smith, 1981). Residuals can also be cumulated across candidates, and an index MSR_i can

be used to describe the degree to which each candidate fits the underlying rating model. A large value of MSR_i would indicate that candidate i has not, for some reason, been measured as well (or on the same constructs) as other examinees. In item-response theory, conceptually similar indices are interpreted as measures of "appropriateness measurement" (Hulin, Drasgow, & Parsons, 1983; Levine & Rubin, 1979) or "person fit" (Wright & Stone, 1979). Values of MSR_i, when computed within selected demographic groups (e.g., males, females), can also be used to study issues related to fairness and bias. The least-squares models also provide standard errors for each candidate and each rater. For example, the standard errors associated with each candidate's adjusted rating for the OLS model exhibit minimal variability, whereas the standard errors based on the WLS model exhibit considerable variability, as they are sensitive to the differential levels of error variance associated with the individual raters who assigned the ratings.

Although the present study addressed rating errors within the context of oral examinations, the least-squares models can be applied in many other rating circumstances. The rating design must be structured so that each candidate is evaluated by more than a single rater and each rater evaluates more than a single candidate. Also, candidates and raters should be crossed (incompletely); cohorts of students cannot be nested within fixed teams of raters. That is, a certain degree of overlap must exist in the rating data in such a way that each rater can be directly or indirectly linked to each of the other raters (de Gruijter, 1984). On the surface, the incidence of rating designs that meet these requirements may seem quite low. To the contrary, such designs often naturally occur, or could often occur in many practical contexts such as: interviews of job applicants, ratings given at assessment centers, reviews of manuscripts or proposals, evaluation of faculty by

review committees, and institutional accreditation (hospitals, universities), to name a few. Grades in college courses also represent an instance of an incomplete rating design. Previous investigations have used either a simple additive model (Elliott & Strenta, 1988) or an IRT graded-response model (Young, 1990; 1991) to adjust college grades for leniency error. The OLS models presented in this paper appear to offer another alternative for adjusting college grades.

Controlling rater error (leniency, discrimination) has heretofore been left to the good intentions, but limited effects, of rater training programs (e.g., Bernardin & Pence, 1980; Trier, 1983). It is possible that statistical methods to control for this type of rater error may prove to be a useful and inexpensive adjunct to rater training. We hope that future research will address this possibility.

References

American Board of Medical Specialties. (1990). *Annual report & reference handbook-1990.* Evanston, IL: Author.

Barnes, E. J., & Pressey, S. L. (1929). The reliability and validity of oral examinations. *School and Society, 30,* 719-722.

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

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63,* 301-308.

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.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410-421.

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

Bray, D. W., & Byham, W. C. (1991). Assessment centers and their derivatives. *Journal of Continuing Higher Education, 39*(1), 8-11.

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

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

Cohen, J. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 137-146.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Erlbaum.

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

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

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement, 25*, 333-347.

Hedge, J. W., & Kavanaugh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73*, 68-73.

Hill, D. S. (1984). Oral examinations: Standards and strategies. *Professional Practice of Psychology, 5*, 69-77.

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

Hubbard, J. P. (1971). *Measuring medical education.* Philadelphia: Lea & Febiger.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

King, R. B. (1983). A preliminary statistical analysis of an oral examination. In J. S. Lloyd (Ed.), *Oral examinations in medical specialty board certification.* Chicago, IL: American Board of Medical Specialties.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72—107.

Levine, H. G., & McGuire, C. H. (1970). The validity and reliability of oral examinations in assessing cognitive skills in medicine. *Journal of Educational Measurement, 7*(2), 63-73.

Levine, M. W., & Rubin, D. F. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

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

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.

O'Donohue, W. J., & Wergin, J. F. (1978). Evaluation of medical students during a clinical clerkship in internal medicine. *Journal of Medical Education, 53*, 55-58.

Raymond, M. R., & Houston, W. M. (1990). *Detecting and correcting for rater effects in performance assessment* (Research Report No. 90-14). Iowa City: The American College Testing Program.

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.

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

Small, S. M. (1982). Evaluation methodology for the oral examination of the American Board of Psychiatry and Neurology. In J. S. Lloyd (Ed.), *Evaluation of noncognitive skills and clinical performance*. Chicago: American Board of Medical Specialties.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika, 26*, 205–219.

Trier, W. C. (1983). Oral examiner training by the American Board of Plastic Surgery. In J. S. Lloyd (Ed.), *Oral examinations in medical specialty board certification*. Chicago: American Board of Medical Specialties.

Watson, M. A. (1984). The Colorado model: A structured oral examination. *Professional Practice of Psychology, 5*, 79–85.

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

Wright, B.D., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement, 27*, 175–186.

Young, J. W. (1991). Improving the prediction of college performance of ethnic minorities using the IRT-based GPA. *Applied Measurement in Education, 4*, 229–239.

Appendix

The following tables present a sample rating matrix suitable for least-squares regression and the corresponding design matrix. Ratings on a seven-point scale are provided in the cells.

Incomplete Rating Design Suitable for Methods to Correct for Rater Effects

	Rater		
Candidate	A	B	C
1	3		2
2		3	3
3	5	4	
4		5	4
5	7	5	

Design Matrix for OLS Method Based on Data in Table Above

Observed	Candidates					Raters	
Rating	C_1	C_2	C_3	C_4	C_5	R_A	R_B
3	1	0	0	0	0	1	0
2	1	0	0	0	0	-1	-1
3	0	1	0	0	0	0	1
3	0	1	0	0	0	-1	-1
5	0	0	1	0	0	1	0
4	0	0	1	0	0	0	1
5	0	0	0	1	0	0	1
4	0	0	0	1	0	-1	-1
7	0	0	0	0	1	1	0
5	0	0	0	0	1	0	1